Harnessing Opportunities for Quality Improvement from Primary Care Electronic Health Records

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy (PhD) in the Faculty of Biology, Medicine and Health

2017

Benjamin Brown
School of Health Sciences, Centre for Health Informatics
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1</td>
<td>Introduction</td>
<td>p10</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Methods</td>
<td>p26</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Results (1) <strong>The Case for Conceptual and Computable Cross-Fertilisation Between Audit and Feedback and Clinical Decision Support</strong></td>
<td>p32</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Results (2) <strong>Clinical Performance Feedback Intervention Theory (CP-FIT): A Meta-synthesis of Qualitative Studies</strong></td>
<td>p47</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Results (3) <strong>Interface Design Recommendations for Electronic Audit and Feedback: Hybrid Usability Evidence from a Research-led System</strong></td>
<td>p104</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>Results (4) <strong>Multi-method Laboratory User Evaluation of an Actionable Clinical Performance Information System: Implications for Usability and Patient Safety</strong></td>
<td>p143</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>Results (5) <strong>Implementing Actionable Clinical Feedback in UK Primary Care: a Longitudinal Optimisation Study Using Clinical Performance Feedback Intervention Theory</strong></td>
<td>p196</td>
</tr>
<tr>
<td>Chapter 8</td>
<td>Discussion</td>
<td>p241</td>
</tr>
<tr>
<td>Chapter 9</td>
<td>Conclusion</td>
<td>p261</td>
</tr>
<tr>
<td>Appendix 1</td>
<td>Search terms (Chapter 4)</td>
<td>p262</td>
</tr>
<tr>
<td>Appendix 2</td>
<td>Example data extraction form (Chapter 4)</td>
<td>p268</td>
</tr>
<tr>
<td>Appendix</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>----------</td>
<td>-------------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Appendix 3</td>
<td>Study details (Chapter 4)</td>
<td>p273</td>
</tr>
<tr>
<td>Appendix 4</td>
<td>Theories, models, and frameworks considered during the synthesis process (Chapter 4)</td>
<td>p294</td>
</tr>
<tr>
<td>Appendix 5</td>
<td>CP-FIT Constructs (Chapter 4)</td>
<td>p301</td>
</tr>
<tr>
<td>Appendix 6</td>
<td>User tasks and Goal-Action structure provided to system evaluators (Chapter 5)</td>
<td>p335</td>
</tr>
<tr>
<td>Appendix 7</td>
<td>Example usability issue data collection form (Chapter 5)</td>
<td>p338</td>
</tr>
<tr>
<td>Appendix 8</td>
<td>Sample of the usability issue severity evaluation form (Chapter 5)</td>
<td>p340</td>
</tr>
<tr>
<td>Appendix 9</td>
<td>Screening questionnaire (Chapter 6)</td>
<td>p342</td>
</tr>
<tr>
<td>Appendix 10</td>
<td>Tasks (Chapter 6)</td>
<td>p344</td>
</tr>
<tr>
<td>Appendix 11</td>
<td>Post-test questionnaires (Chapter 6)</td>
<td>p346</td>
</tr>
<tr>
<td>Appendix 12</td>
<td>Interview schedule (Chapter 7)</td>
<td>p348</td>
</tr>
</tbody>
</table>

Word count: 62413
## Figures

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Figure 1: A typical example of the result of an audit for patients with hypertension in primary care</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>Figure 2: Effectiveness of common quality improvement strategies</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>Figure 3: Screenshot from example electronic health record with clinical decision support in primary care</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>Figure 2: Screenshot from a typical electronic Audit and Feedback system in primary care</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>Figure 1: Synthesis process</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>Figure 2: Flowchart of screening process of studies</td>
<td>57</td>
</tr>
<tr>
<td>4</td>
<td>Figure 3: Clinical Performance Feedback Intervention Theory</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>Table 4: How CP-FIT explains findings from the Cochrane review</td>
<td>81</td>
</tr>
<tr>
<td>4</td>
<td>Table 5: Tentative best practices in literature for A&amp;F compared to CP-FIT</td>
<td>82</td>
</tr>
<tr>
<td>5</td>
<td>Figure 1: PINGR’s Overview page</td>
<td>110</td>
</tr>
<tr>
<td>5</td>
<td>Figure 2: PINGR’s Preview page</td>
<td>111</td>
</tr>
<tr>
<td>5</td>
<td>Figure 3: Overview of hybrid usability inspection methodology</td>
<td>114</td>
</tr>
<tr>
<td>5</td>
<td>Figure 4: Demonstration of process for deriving Goal-Action structure from user tasks</td>
<td>115</td>
</tr>
<tr>
<td>5</td>
<td>Figure 5: Flowchart of usability issue discovery and finalisation</td>
<td>118</td>
</tr>
<tr>
<td>5</td>
<td>Figure 6: Stacked bar plot of usability issues per heuristic category versus interface components</td>
<td>119</td>
</tr>
<tr>
<td>6</td>
<td>Figure 1: The Performance Improvement plaN GeneratoR</td>
<td>148</td>
</tr>
<tr>
<td>6</td>
<td>Figure 2: Summary of the data collection and analysis process</td>
<td>156</td>
</tr>
<tr>
<td>6</td>
<td>Figure 3: Time taken to complete each task by participants</td>
<td>157</td>
</tr>
<tr>
<td>6</td>
<td>Figure 4: Number of errors made during each task by participants</td>
<td>158</td>
</tr>
<tr>
<td>6</td>
<td>Figure 5: Number of errors according to usability heuristics and interface component</td>
<td>159</td>
</tr>
<tr>
<td>6</td>
<td>Figure 6: Visual attention heatmap for Task 2</td>
<td>163</td>
</tr>
<tr>
<td>6</td>
<td>Figure 7: Transition matrix of eye movement sequences between areas of interest during Task 2</td>
<td>164</td>
</tr>
<tr>
<td>6</td>
<td>Figure 8: Participant responses to the Object-Action Interface questionnaire</td>
<td>171</td>
</tr>
<tr>
<td>7</td>
<td>Figure 1: Clinical Performance Feedback Intervention Theory</td>
<td>197</td>
</tr>
<tr>
<td>7</td>
<td>Figure 2: PINGR</td>
<td>202</td>
</tr>
<tr>
<td>7</td>
<td>Figure 3: GP practices use of PINGR per month</td>
<td>212</td>
</tr>
<tr>
<td>7</td>
<td>Figure 4: Quality indicators accessed per month</td>
<td>213</td>
</tr>
<tr>
<td>7</td>
<td>Figure 5: Process map of PINGR use by individual users</td>
<td>218</td>
</tr>
<tr>
<td>Chapter</td>
<td>Table</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>Table 1: Example Read codes</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>Table 2: Thesis ‘results’ chapters and their relation to research objectives and publication status</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>Table 1: Comparison of electronic CDS and A&amp;F</td>
<td>39</td>
</tr>
<tr>
<td>4</td>
<td>Table 1: Inclusion criteria</td>
<td>52</td>
</tr>
<tr>
<td>4</td>
<td>Table 2: Frequency of main paper characteristics</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>Table 3: Quality appraisal results</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>Table 1: Overview of tasks performed by evaluators during usability inspection</td>
<td>116</td>
</tr>
<tr>
<td>5</td>
<td>Table 2: Usability issue severity rating scale</td>
<td>116</td>
</tr>
<tr>
<td>6</td>
<td>Table 1: Overview of tasks performed by participants</td>
<td>152</td>
</tr>
<tr>
<td>7</td>
<td>Table 1: How Clinical Performance Feedback Intervention Theory (CP-FIT) informed the design and implementation of PINGR</td>
<td>201</td>
</tr>
<tr>
<td>7</td>
<td>Table 2: Quality indicators included in PINGR</td>
<td>206</td>
</tr>
<tr>
<td>7</td>
<td>Table 3: GP practice (pseudonyms) and study participant characteristics (ordered by recruitment date)</td>
<td>209</td>
</tr>
<tr>
<td>7</td>
<td>Table 4: Processes detailed in Clinical Performance Feedback Intervention Theory in practices using PINGR</td>
<td>214</td>
</tr>
<tr>
<td>7</td>
<td>Table 5: Patient-level processes detailed in Clinical Performance Feedback Intervention Theory according to quality indicator (ordered by dates introduced)</td>
<td>216</td>
</tr>
<tr>
<td>8</td>
<td>Table 1: How PINGR has evolved during the course of this thesis</td>
<td>237</td>
</tr>
</tbody>
</table>
Abbreviations

A&F: Audit and feedback  
AF: Atrial fibrillation  
CCG: Clinical Commissioning Group  
CDS: Clinical Decision Support  
CKD: Chronic Kidney Disease  
COPD: Chronic Obstructive Pulmonary Disease  
CP-FIT: Clinical Performance Feedback Intervention Theory  
CW: Cognitive Walkthrough  
DM: Diabetes mellitus  
e-A&F: Electronic Audit and Feedback  
EHR: Electronic Health Record  
ENTREQ: Enhancing transparency in reporting the synthesis of qualitative research  
GP: General Practitioner  
HE: Heuristic Evaluation  
NHS: National Health Service  
NICE: National Institute for Health and Care Excellence  
PINGR: Performance Improvement plaN GeneratoR  
PPI: Patient and public involvement  
PR: Pulmonary rehabilitation  
PRIMER: Primary Care Research in Manchester Engagement Resource  
PRISMA: Preferred reporting items for systematic reviews and meta-analyses  
PROSPERO: International prospective register of systematic reviews  
QI: Quality improvement  
QOF: Quality and outcomes framework  
RAG: Red Amber Green  
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms  
US: United States
Abstract

**Background:** UK primary care accounts for 90% of patient contacts in the NHS, and over 300 million consultations every year. Consequently, when primary is suboptimal it has important impacts on population health. At the same time, virtually all general practices use electronic health records (EHR) to capture patient data. Clinical Decision Support (CDS) systems use it to highlight when individual patients do not receive care consistent with clinical guidelines, though ignore the wider population. Electronic Audit and Feedback (e-A&F) systems address the wider population, but their results are difficult to interpret. EHR data has the richness to suggest ways in which care quality could be improved, though this is currently not exploited. The aim of this thesis was to make progress towards better use of primary care EHR data for the purposes of quality improvement (QI) by focusing on e-A&F as a vehicle. Research Objectives were: 1) Develop a model and recommendations to guide EHR data analysis and its communication to health professionals; 2) Use these models and recommendations to develop a system for UK primary care; 3) Implement and evaluate the system to test the models and recommendations, and derive generalisable knowledge.

**Methods:** The overall approach of this thesis was informed by guidance from the Medical Research Council on the development of complex interventions, and Boyriceti al.‘s evidence-based framework for the development of health information technologies (Chapter 2). Theory was first identified through a critical examination of the empirical and theoretical literature regarding CDS and e-A&F systems (Chapter 3), then built upon in a systematic literature search and metasynthesis of qualitative studies of A&F (and e-A&F) interventions (Chapter 4). This resulted in the development a new theory of A&F (Clinical Performance Feedback Intervention Theory; CP-FIT), which was used to inform the development of an e-A&F system for UK primary care – the Performance Improvement plaN GeneratoR (PINGR; version 1). PINGR was then iteratively optimised through a series of three empirical studies. First, its usability was evaluated by software experts using Heuristic Evaluation and Cognitive Walkthrough methodologies (Chapter 5). GPs then performed structured tasks using the system in a laboratory whilst their on-screen interactions and eye movements were recorded (Chapter 6). Finally, PINGR was implemented in 15 GP practices, and CP-FIT used to guide the mixed methods evaluation including examinations of usage records, and interviews with 38 health professionals.

**Results:** There are both empirical and theoretical arguments for combining features from CDS and e-A&F systems to increase their effectiveness; a key recommendation is that e-A&F systems should suggest clinical actions to health professionals (Chapter 3). This is supported by CP-FIT, which has three core propositions: 1) A&F interventions exert their effects through health professionals taking action; 2) Health care organisations have limited capacity to engage with A&F; and 3) Health care professionals and organisations have a strong set of beliefs and behaviours regarding how they provide patient care (Chapter 4). Based on these findings, the unique feature of PINGR is that it suggests improvement actions to users based on EHR data analysis (‘decision-supported feedback’). Key findings from PINGR’s usability evaluation with software experts translated into a set of design guidelines for e-A&F interfaces regarding: summarising clinical performance, patient lists, patient-level information, and suggested actions (Chapter 5). When tested with GPs, these guidelines were found to impact: user engagement; actionability; and information prioritisation (Chapter 6). Following its implementation in practice, PINGR was used on 227 occasions to facilitate the care of 725 patients. These patients were 1.8 (95% CI 1.6-1.9) times more likely to receive improved care according to at least one clinical guideline. Barriers and facilitators to its success included: the resources available to use it; its perceived relative advantages; how compatible it was with pre-existing beliefs and ways of working; the credibility of its data; the complexity of the clinical problems it highlighted; and the ability to act on its recommendations (Chapter 7).

**Conclusion:** It is both feasible and acceptable to health professionals to make better use of EHR data for QI by enabling e-A&F systems to suggest actions for them to take. When designing e-A&F interfaces, attention should be paid to how they summarise clinical performance, and present patient lists and detailed patient-level information. Implementation of e-A&F interventions is influenced by availability of resources, compatibility with existing workflows, and ability to take action based on their feedback results. Unresolved tensions exist regarding how they may deal with patient complexity. Policymakers should consider the relevance of these findings for National Clinical Audits and pay-for-performance initiatives.

7
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.
Acknowledgments

I am grateful to: Wellcome Trust for providing me with a Clinical Research Training Fellowship in order to undertake this work; my PhD supervisors Professor Iain Buchan, Dr Matthew Sperrin, Dr Thomas Blakeman, and Professor Stephen Campbell; my co-authors; the health professionals who participated in these studies; and my wife Lulu for putting up with all the lost evenings and weekends over the last three years.

About the author

I graduated from Manchester Medical School in 2007 and completed Foundation Training at Salford Royal Hospital. From 2009 to 2010 I undertook a Fellowship in health services management at the North West Strategic Health Authority where I provided clinical input to regional Urgent Care policy. From 2010 to 2014 I trained to be a GP as a National Institute for Health Research (NIHR) Academic Clinical Fellow, and completed an MSc at The University of Birmingham’s Health Services Management Centre in 2012. I qualified as a GP in August 2014, and started work on this PhD the following month. I have undertaken clinical work one day per week for its duration.
Chapter 1

Introduction

1.1 UK primary care and electronic health care records

Primary care can be defined as a health care service that provides: first-contact access for each new need; long-term person (not disease) focused care; comprehensive care for most health needs; and coordinated care when it must be sought elsewhere [1]. Since its inception in 1948, primary care has been provided by general practitioners (GPs), and more recently nurses and other allied health professionals, as a crucial part of the UK National Health Service (NHS) [2]. It acts as ‘gatekeeper’ to all other NHS services (other than emergencies), and in recent years has taken on considerable responsibility from secondary care with regard to chronic disease management, in addition to public health responsibilities such as screening and health promotion [3]. Currently, primary care accounts for 90% of all patient contacts in the NHS [4], translating to over 300 million consultations every year [5].

GPs used to document their interactions with patients on paper notes (Lloyd Georges), which were often illegible, brief, and difficult to search or find [6]. Since the early 1990s UK primary care has led the way in terms of computerisation to help with research and the organisation of care. The first electronic health record (EHR) was introduced in 1970 by Dr John Preece at Whipton near Exeter, and by 1996, 96% of general practices were computerised [7]. Relatively recent drivers in the computerisation of primary care has been the introduction of the Quality and Outcomes Framework (QOF), a pay-for-performance programme in which EHR data are extracted from most GP practices to determine how they are paid [8]. This system hinges on the use of coded data inputs (currently Read codes invented by Dr James Read, but soon to be replaced by [Systematized Nomenclature of Medicine – Clinical Terms]), which clinicians use to describe the care provided to patients [9]. These codes can relate to most aspects of care including diagnoses, physiological measurements, and medications, and can be supplemented by descriptive free text that the clinician will type into the EHR (Table 1). Consequently, given the volume, history, and nature of work conducted by UK primary care, its EHR data are among the most detailed longitudinal records of coded data in the world [10].
### Table 1: Example Read codes [9]

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases</td>
<td>G713 Extradural haemorrhage</td>
</tr>
<tr>
<td>Procedures</td>
<td>4662 Urine glucose test negative</td>
</tr>
<tr>
<td>History/symptoms</td>
<td>1377 Ex-smoker</td>
</tr>
<tr>
<td>Examination/signs</td>
<td>2BB6 O/E – retinal exudates 2BB7</td>
</tr>
<tr>
<td>Administration</td>
<td>9313 FP1001 claim up to date</td>
</tr>
</tbody>
</table>

This digitisation of primary care records has led to a number of initiatives to expand the use of EHR data, with mixed results. In general these initiatives extract coded data only from EHRs; recent progress has been made with regards to extracting free text, but currently this is too expensive to do on a large scale. Success stories include the use of EHR data for research purposes with the emergence of databases such as Clinical Practice Research Data Link and Research One, which have provided important epidemiological insights and have the potential to transform the way clinical trials are performed [11]. Health service planning has also successfully used EHR data to understand the needs of local populations, which has been particularly relevant for primary care since it acquired responsibility for commissioning services with the formation of Primary Care Trusts and now Clinical Commissioning Groups (CCGs) – primary care-led organisations responsible for commissioning services in a particular geographical area. Less successful initiatives have involved national efforts to share this data for the espoused purposes of improving clinical care in Summary Care Records [12], and for more opaque purposes in the “care.data” programme. Particularly in the latter, the approach to not clearly communicating the benefits of sharing primary care EHR data with organisations outside the GP practice (e.g. improved drug safety monitoring, and public health surveillance), and problems with obtaining consent, led to public distrust [13]. As a result a number of local initiatives led by CCGs and other regional organisations have formed to share local primary care EHR data for specific purposes, such as DataWell in Greater Manchester [14].

### 1.2 Care quality and its improvement

It has been suggested that good quality care not only saves lives but can also save money [15]. However, quality in health care is a multi-dimensional concept that is difficult to define. The US Institute of Medicine defines quality as ‘the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge’ [16]. They identify that high quality care has six dimensions relating to:

- Safety – avoids injuries to patients;
- Effectiveness – provides care based on scientific knowledge to those who could benefit and refrain from those not likely to benefit;
• Patient-centredness – is respectful of and responsive to individual patient preferences, needs, and values;
• Timeliness – reducing waits and harmful delays;
• Efficiency – avoiding waste;
• Equity – does not vary in quality because of patients’ personal characteristics.

In contrast, Campbell et al. [17] reviewed the literature on care quality and developed their more simplified definition, which has two major components: Accessibility and Effectiveness. Accessibility means that patients can access health care when they need it, and had subcomponents of Availability and Affordability [17]. Effectiveness is the extent to which care delivers its intended outcome or results, and has subcomponents of Clinical care and Inter-personal care [17]. In the NHS, the definition of care quality, which is now part of health care law [18] dates back to the Next Stage Review in 2008 [19], and has three components:

• Patient safety – The environment is safe and clean, and reduces avoidable harm such as medication errors or healthcare associated infections;
• Patient experience – Patients are treated with compassion, dignity and respect;
• Effectiveness – Treatment is successful according to measures such as mortality or patient-reported outcomes.

Quality Improvement (QI) on the other hand refers to ‘better patient experience and outcomes achieved through changing provider behaviour and organisation through using a systematic change method and strategies’ [20]. QI has a history in quality control that emerged from industry and factory production in the early 1920s, and the influence of United States (US) experts Deming, Juran and Feigenbaum, and Japanese expert Ishikawa [21]. Similar strategies of thinking of health care in terms of systems to be improved were pioneered by Donabedian in the 1960s [22], who introduced the structure-process-outcome model, and later by Berwick in the 1980s and 1990s [23].

In the NHS, a key development in relation to QI was the introduction of Clinical Governance in 1989, which was a policy defined as the ‘a system through which NHS organisations are accountable for continuously improving the quality of their services and safeguarding high standards of care by creating an environment in which excellence in clinical care will flourish’ [24]. After high profile failures in paediatric cardiac surgery in Bristol, organisational shortcomings were blamed, and statutory obligations consequently placed on boards of NHS organisations for managing the quality of care they provided [25]. Clinical governance subsequently developed into specific activities, sometimes called ‘pillars’, generally encompassing: clinical risk management; clinical audit; patient involvement; education and training; clinical effectiveness; research and development;
staff focus; and use of information [26]. Of particular note is audit, which has been a mandatory requirement for NHS doctors since the late 1980s [27]. The first audits were thought to have been conducted by Florence Nightingale in the Crimean War of 1853-1855, and later by Ernest Codman, a US surgeon in the early 20th century [28]. Although what is considered audit has developed over time, a widely accepted modern description is: ‘a quality improvement cycle that involves measurement of the effectiveness of healthcare against agreed and proven standards for high quality, and taking action to bring practice in line with these standards so as to improve the quality of care and health outcomes.’ [29]

Based on this definition, the Healthcare Quality Improvement Partnership (HQIP; a UK organisation that promotes and supports the use of audit in the UK), suggests that audit has four stages [30]: preparation and planning; measuring performance; implementing change; and sustaining improvement. The process recommends that the audit should measure performance against standards for clinical processes and patient outcomes that are based on the ‘best available evidence’, and that data to measure performance should be collected from either paper or electronic health records on a large sample of patients. This generally means that the results of audits are presented as quantitative results stating the number or proportion of patients meeting a particular standard or outcome (often called Performance Measures [PMs]; Figure 1). Interestingly, there is little guidance in policy documents with regards to how these results should be communicated to health professionals. This is in contrast to the academic literature, where audit is generally referred to as Audit and Feedback (A&F), and there is recognition that there is variation in how audit results can be reported (e.g. verbally vs written, graphs vs text, bar charts vs line charts, target performance vs no target) [31], which in turn can have significant impacts on its effectiveness [32]. Given the importance of how audit results are presented to health professionals, I will use the term ‘A&F’ in preference to ‘audit’ for the rest of this thesis in order to be more accurate.

Figure 1: A typical example of the result of an audit for patients with hypertension in primary care

![Graph showing % Hypertensive Patients with Controlled Blood Pressure](image)
Both policy and the academic literature appear to agree that the ‘best available evidence’ on which to base A&F should be from clinical guidelines, such as those published by the National Institute for Health and Care Excellence (NICE) [31,33]. Clinical guidelines arose from the evidence-based practice movement, with a focus on synthesising existing research through systematic review methods into ‘statements to assist practitioner and patient decisions about appropriate healthcare for specific clinical circumstances’ [34]. For example, the relevant standard for Figure 1 above is: ‘Aim for a target clinic blood pressure below 140/90 mmHg in people aged under 80 years with treated hypertension’ [35]. A key criticism of clinical guidelines, particularly in primary care, are that they are based on evidence from randomised controlled trials (RCTs), where the patients involved are unrepresentative of those seen in clinical practice [36,37]. For example, they may be younger and with fewer comorbidities, and therefore their recommendations may be inappropriate or irrelevant [38,39]. Furthermore, in order to be able to measure quality as defined in clinical guidelines from medical record data, the definition of quality necessarily has to be relatively narrow. For example, taking the NHS definition of quality – Safety, Experience, Effectiveness – patient experience and holistic aspects of care are often ignored when measuring quality according to clinical guidelines [36]. There have been some developments to challenge this in terms of collecting patient experience measures (PREMs), though there are concerns about their utility, and they are not routinely collected in medical records [40].

From a practical point of view, in primary care, A&F may be conducted in a variety of ways: by individual doctors or nurses who decide to examine their own practice, or conducted/facilitated by external agencies [29]. In particular, these external agencies may be local or national. Local initiatives may include those run by CCGs, who may have teams to help with conducting the A&F. National initiatives include National Clinical Audits (NCAs) where the A&F is conducted with or without the co-operation of the local clinician. NCAs are run by HQIP, though currently the only one relevant to primary care is the diabetes audit [41].

A&F is not the only QI technique by any means, though may be considered fundamental to QI because in order to know whether improvement has occurred, some sort of measurement needs to occur [42]. Indeed, all commonly used QI techniques in health care such as Plan-Do-Study-Act (PDSA), Statistical Process Control, Six Sigma, and Lean, use measurement and monitoring of data in a very similar way to A&F [43]; the PDSA cycle in particular bears resemblance to the four-stage audit model proposed by HQIP [30]. In the academic literature, a different approach to defining QI interventions is taken. The Cochrane organisations’ Effective Practice and Organisation of Care (EPOC)
group have conducted systematic reviews of QI interventions (other than A&F) including [44]:

- Educational meetings – conferences, lectures, workshops or traineeships.
- Education outreach – Use of a trained person who met with providers in their practice settings to give information with the intent of changing the provider’s practice.
- Printed educational materials – Distribution of published or printed recommendations for clinical care, including clinical practice guidelines, audio-visual materials and electronic publications
- Reminders – Patient or encounter specific information, provided verbally, on paper or on a computer screen, which is designed or intended to prompt a health professional to recall information.

Similar to A&F, these interventions also tend to use standards described in clinical guidelines as the aspect of care they are trying to improve [45]. Results of EPOC’s systematic reviews show similar findings in that all these interventions are moderately effective at improving measured care quality, with wide variation in their results (Figure 2), including negative effects results in some cases. For example, the latest A&F review estimates that A&F improves compliance with desired practice by a median of 4.3% (range -9.0-70.0%). In general, it is unclear why there is such variation in these interventions, and what features are associated with their success [46].
1.3 Using EHR data for quality improvement: Audit and Feedback, and Clinical Decision Support

In terms of the QI interventions described above, EHR data are commonly used to highlight suboptimal care for either individual patients (i.e. reminders) or populations (i.e. A&F). In general, these interventions are delivered as electronic software: electronic reminders are usually referred to as Clinical Decision Support (CDS) systems, whereas I will call their A&F counterpart ‘electronic A&F’ (e-A&F). CDS systems commonly used in primary care are generally in-built to the EHR system and provide a ‘pop up’ message when patients visit the GP practice and their care is inconsistent with clinical guidelines (Figure 3). For example, for a patient with hypertension it may show that their blood pressure measurement last month was high. This type of system is problematic because it: ignores hard-to-reach patients that do not see doctors often (and who may need care the most [48]), is often over-ridden during a consultation [49]; hijacks the consultation from the patient’s agenda [50]; and does not address organisational barriers to improving care quality [51]. e-A&F systems on the other hand highlight all patients that may be receiving suboptimal care. Continuing the example, they may inform a doctor the number of their patients with hypertension whose latest blood pressure measurement was too high. However, finding the time to access this type of system, and translate its results into improvement action is difficult [52]. An example of an e-A&F system is provided in Figure 4.
In UK primary care, QOF has dominated both CDS and e-A&F systems. Although QOF is a mechanism for paying GP practices and is not a QI intervention, its original rationale was to improve care quality, and is often mistakenly viewed as a proxy for quality [53]. QOF works by paying GP practices on the proportion of patients who have achieved particular standards of care, which is analogous to A&F. However, these standards, although informed by NICE guidance, are subsequently negotiated with the British Medical Association about what is achievable in practice [53]. Therefore there are often disparities between what NICE guidelines recommend and what GP practices are incentivised to achieve: for example, QOF encourages target blood pressure to be controlled to 150/90 mmHg, which is easier to achieve than the 140/90 target set by NICE [35]. Furthermore, QOF as technically does not provide feedback to GP practices. Performance is measured from 1st April each year, and payment is made to a practice based on their performance on 31st March the following year. Most EHR systems used by GP practices however, have integrated e-A&F functions that can be used to monitor their performance against QOF measures. These systems can also be used by the practice to create their own measures to conduct A&F on a clinical topic of their choice, though the practice has to be skilled in developing the relevant EHR search queries. Other e-A&F systems separate from the EHR are also used in primary care. Examples include those developed by PRIMIS at the University of Nottingham that are relative complex to use [54], or from Public Health England, which are problematic for the reasons above because they are generally based on QOF data [55].

Figure 3: Screenshot from example electronic health record with clinical decision support in primary care

Note: Clinical decision support alert is in pink box on bottom right.
It is arguable that both CDS and e-A&F systems do not make the most of the richness of primary care EHR data. Simply highlighting whether a specific patient (in the case of CDS) or group of patients (in e-A&F) have not received/achieved a particular standard of care does not necessarily provide guidance on what action could be taken. For example, using the NICE hypertension standard from Figure 1 above [35], a patient may have not met the standard because they have: not had their blood pressure measured; had their blood pressure measured but are currently prescribed suboptimal medication; been prescribed optimal medication but are not taking it; a condition in which it is inappropriate to apply this standard (e.g. terminal cancer) [56]. Each of these reasons is associated with a different specific clinical action, and most, if not all, can be determined from Read codes in EHR data (Table 1): Read codes capture physiological measurements (e.g. blood pressure recordings), medications and when they are prescribed, and diagnoses. Providing clinical actions alongside both CDS and e-A&F interventions may be one answer to reduce the variability of, and improve the relatively moderate levels of, their effectiveness (Figure 2) [32,57].

1.4 Hypotheses of this thesis

The underlying hypotheses tested by this PhD has been that better use of primary care EHR data is feasible, acceptable, and useful in improving the quality of primary care. In particular, this is in regard to the provision of actions that could be taken by health professionals. Taking into account the problems outlined above in section 1.3 with regards to CDS, the main approach of this thesis will be with regards to using EHR data by e-A&F
systems. That being said, the theoretical relevance of CDS to this approach will be addressed in Chapter 3.

1.5 Aims and objectives

The overall aim of this thesis is to make progress towards better use of primary care EHR data for the purposes of QI, using e-A&F as a vehicle. It intends to develop both generic frameworks and models for use by researchers and practitioners, in addition to using these outputs to build specific tools and implement them into clinical practice, whilst using rigorous evaluation to both optimise their design and derive generalisable knowledge. Consequently, there are three Research Objectives:

1. Develop a model and set of recommendations that can be used to guide EHR data analysis and its communication to health professionals for QI purposes by e-A&F systems;
2. Use these models and recommendations to develop an e-A&F system for UK primary care;
3. Implement and evaluate the system to test the models and recommendations, and derive generalisable knowledge about using EHR data for QI.

1.6 Importance of this work

Given the amount of care delivered by GP practices in the UK, when this is suboptimal it has important impacts on population health. For example, in England alone it is estimated that 7000 quality adjusted life years could be saved through a 15% increase in patients with better-managed hypertension [58] – a condition predominantly treated in primary care. Furthermore, given the ubiquity of EHRs and the extraction and use of their data both in the UK and beyond, small improvements in the way they could be used to improve primary care could result in large benefits in population health.

1.7 Structure of this thesis

This thesis is written in a ‘journal’ format, in which the results are presented as self-contained chapters suitable for publication in a peer-reviewed scientific journal. Some of the papers included have been already published, whereas some are under consideration or being prepared for submission. The remainder of this thesis is split into seven further chapters. The next chapter describes the overall methodology taken I have taken to achieve the above objectives, and is followed by five ‘results’ chapters (Table 2):

- Chapter 3 (*The Case for Conceptual and Computable Cross-Fertilization Between Audit and Feedback and Clinical Decision Support*) is a conceptual paper that represents my early thinking towards how primary care EHR data could be put to better use in both CDS and e-A&F systems. It critically reviews the literature to make an argument for the conceptual union of their design, and support for the
provision of action plans with care quality measurement. A version of this chapter was presented at the leading health informatics conference (Medinfo) in Sao Paulo, 2015, and published in Studies in Health Technology and Informatics [59].

• Chapter 4 (Clinical Performance Feedback Intervention Theory (CP-FIT): A Meta-synthesis of Qualitative Studies) reports a meta-synthesis of findings from qualitative studies of A&F interventions to develop a theoretical model of causal effects for these interventions. Work on this chapter started at the beginning of my doctoral studies though was only finished towards the end. Consequently, its emerging (rather than finalised) findings were used to inform Chapters 5 and 6. However, given its conceptual relevance I believe it is important to be presented before them. This chapter is currently being prepared for submission to PLOS Medicine.

• Chapter 5 (Interface design recommendations for computerised clinical audit and feedback: Hybrid usability evidence from a research-led system) uses recommendations from Chapter 3 and emerging findings from Chapter 4 to develop a prototype e-A&F system for UK primary care – the Performance Improvement plaN GeneratoR (PINGR; version 1). It describes an evaluation of the system’s usability by software experts, which derive guidelines for the design of e-A&F systems in general. A version of this chapter was published in the International Journal of Medical Informatics in 2016 [60].

• Chapter 6 (Multi-method laboratory user evaluation of an actionable clinical performance information system: implications for usability and patient safety) reports the evaluation of an optimised version of PINGR (version 2) in a lab-based setting with primary care clinicians using screen recording and eye-tracking technology. Its findings are used to refine the set of design recommendations derived in Chapter 5. A version of this chapter will be published in the Journal of Biomedical Informatics in 2018 [61].

• Chapter 7 (Implementing actionable clinical feedback in UK primary care: a longitudinal optimisation study using Clinical Performance Feedback Intervention Theory) uses finalised results from Chapter 4 to inform a pilot-test of PINGR (improved after Chapter 6; version 3) in GP practices. This chapter is currently being prepared for submission to the journal Implementation Science.

These results chapters are then followed by a chapter discussing the implications of their findings, limitations, comparison to wider literature, and planned future work. The thesis finishes with a brief concluding chapter examining whether the research objectives of the thesis have been met.
# Table 2: Thesis ‘results’ chapters and their relation to research objectives and publication status

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Research objectives addressed</th>
<th>Publication status at time of writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>The Case for Conceptual and Computable Cross-Fertilization Between Audit and Feedback and Clinical Decision Support</td>
<td>RO1</td>
<td>Published [59]</td>
</tr>
<tr>
<td>4</td>
<td>Clinical Performance Feedback Intervention Theory (CP-FIT): A Meta-synthesis of Qualitative Studies</td>
<td>RO1</td>
<td>Under preparation</td>
</tr>
<tr>
<td>5</td>
<td>Interface Design Recommendations for Electronic Audit and Feedback: Hybrid Usability Evidence from a Research-led System</td>
<td>RO1-3</td>
<td>Published [60]</td>
</tr>
<tr>
<td>6</td>
<td>Multi-method Laboratory User Evaluation of an Actionable Clinical Performance Information System: Implications for Usability and Patient Safety</td>
<td>RO1-3</td>
<td>Published [61]</td>
</tr>
<tr>
<td>7</td>
<td>Implementing Actionable Clinical Feedback in UK Primary Care: a Longitudinal Optimisation Study Using Clinical Performance Feedback Intervention Theory</td>
<td>RO2-3</td>
<td>Under preparation</td>
</tr>
</tbody>
</table>

## 1.8 Chapter references


Health Foundation, Quality improvement made simple: What everyone should know about health care quality improvement, London, 2016.


[34] Institute of Medicine, Guidelines for Clinical Practice: from development to use, National Academies Press, Washington D.C., 1992.


N. Ivers, J. Barnsley, R. Upshur, K. Tu, B. Shah, J. Grimshaw, et al., My approach to this job is... one person at a time, Can Fam Physician. 60 (2014) 258–266.

M. Roland, B. Guthrie, Quality and Outcomes Framework: what have we learnt?, BMJ. 354 (2016) i4060.


K. Kawamoto, C. a Houlihan, E.A. Balas, D.F. Lobach, Improving clinical practice using clinical decision support systems: a systematic review of trials to identify
doi:10.1136/bmj.38398.500764.8F.


Chapter 2

Methods

2.1 Introduction
This chapter provides an overview and critique of the overall approach to this thesis. It is not intended to address details on the methods for each of the studies (chapters 3-7), as these are discussed in the relevant chapters. It does however address potential risks and mitigating steps taken during the research process, and a discussion of what is beyond the scope of the thesis.

2.2 Theoretical approach
Health informaticians have been criticised for not considering the impacts and relevance of the technologies and findings their research produces on the real world [1]. In contrast, as a clinician and researcher, my personal and academic experience has taught me that technologies often have unintended consequences and may be ineffective. Consequently, in undertaking this thesis I have considered the use of population-level EHR data to drive improvements in care quality a complex intervention (i.e. electronic Audit and Feedback; e-A&F) [2]. An intervention may be considered ‘complex’ based on the number or variability of: intervention components, behaviours required by those receiving the intervention, groups targeted, possible outcomes, or intervention tailoring [2]. When applied to using EHR data for quality improvement, it is clear to see that it fulfills these criteria because it could be used in different ways, by different health professionals, to undertake different activities, all of which may or may not be successful. A number of general and informatics-based guidance exists on how to develop and evaluate complex interventions that aim to improve care quality (e.g. [3–6]). For the purposes of this thesis, I have used two widely used and accepted frameworks from health research in general by the Medical Research Council guidance (MRC) [2], and from health informatics by Borycki et al [7]. These were chosen because of their extensive use and range of supporting evidence.

MRC guidance recommends a four-stage process of intervention development, feasibility/piloting, evaluation, and implementation [2]. This thesis represents the first two stages of this process. A core part of the intervention development stage involves the identification of relevant evidence and theory. A series of Cochrane reviews of audit and feedback (A&F) [8–10] provides a robust evidence base for using population-level measures of quality to produce improvements in patient care, though a suitable theory is more elusive [11]. Therefore, Chapter 3 directly addresses this evidence gap by
developing a theory by examining the qualitative literature and building on existing models. The second stage of feasibility/piloting is addressed by Chapters 5-7, and follows Boyrcki et al.’s evidence-based framework for the development of health information technologies prior to full deployment [7]. This framework advocates a using a sequence of evaluation studies, with correction of defects after each stage:

1. Heuristic Evaluation [12] and Cognitive Walkthrough [13], where software experts undertake standardised tasks with the system and evaluate the software according to specific design guidelines;
2. Usability Testing, where representative users (e.g. clinicians) undertake representative tasks using the system [14];
3. Clinical Simulation, which is similar to Usability Testing, though is conducted in a representative context [15] e.g. a clinical setting;
4. Naturalistic Testing, where the system is deployed into routine clinical practice to a limited extent, and its routine use studied [7].

This thesis has employed both Heuristic Evaluation and Cognitive Walkthrough in Chapter 5, Usability Testing in Chapter 6, and Naturalistic Testing in Chapter 7, whereas Clinical Simulation has not been used. The main value of Clinical Simulations are to identify issues that arise when technologies are introduced into patient-facing situations (i.e. clinical consultations) where actors pose as patients to prevent any potential unintended consequences arising from the technology that may impact patient safety [15]. From both my personal and research experience [16], using population-level EHR data for quality improvement purposes, is generally conducted outside patient-facing situations. Consequently, little value was perceived to be gained over using a Clinical Simulation study design beyond Usability or Naturalistic Testing.

Mixed methods are employed throughout the thesis, which are recommended in more modern guidance regarding complex general [17] and health informatics-specific interventions [1,5].

2.3 Risks and mitigating steps taken
I identified three main risks with regards to any systems/interventions developed during this thesis: firstly, that they would not be used by health professionals; secondly, that if implemented, they may have unintended adverse effects on patient care as has happened with other health informatics interventions, such as distractions from patient-centredness [18]; and thirdly, that given I led both their development and evaluation, these findings could be prone to a positive bias [19]. Consequently, a number of steps were taken to mitigate these risks. Firstly, as recommended in the evaluation of complex interventions [17,20], mixed methods were used throughout the thesis. This served to
address the first risk because if the interventions were not used quantitatively, qualitative investigations would be able to find out why this happened, and what steps could be taken to improve [17]. In this sense, the approach is similar to Black Box thinking [21] (named after the analysis of flight recordings from aviation accidents), where reasons for failure are studied in detail using a range of data sources in order to avoid it happening again. It also served to address the second risk by enabling triangulation of data sources to identify potential unintended consequences that may otherwise be missed [17]. Secondly, a multidisciplinary research team was involved in the evaluation process, and where appropriate, data collection and analysis focused on working practices and processes of using population-level EHR data for quality improvement, rather than participant’s subjective views. This addressed risk two and three, by moving the research process (particularly for the qualitative aspects) to being more detached and critical [22].

2.4 Scope and delimitations
Given the discussions above and in the previous Chapter, there are number of aspects that will explicitly not be addressed by this thesis. Firstly, given it encompasses the first two stages of the MRC framework (intervention development and feasibility/piloting), it does not attempt to evaluate the effectiveness of using EHR data for quality improvement [20]. However, potential effects on patient-level outcomes are explored in Chapter 7. Secondly, the definition of ‘quality’ care is, as outlined in Chapter 1, the degree to which health care services for individuals and populations increase the likelihood of desired outcomes and are consistent with current professional knowledge’ [23], where ‘professional knowledge’ generally refers to that espoused in clinical guidelines. Further, as primary care EHR data are being used to measure care quality, the thesis is limited to what is recorded. Subsequently, important aspects of care quality not routinely recorded in EHRs are not addressed, such as patient satisfaction [24,25]. Thirdly, given it has the strongest evidence base, the methods of how to improve care quality using population-level EHR data in primary care will focus on A&F, rather than other techniques such as CDS or education outreach (described in Chapter 1; however, the relevance of CDS is discussed in Chapter 4). In addition, it will not address how primary care EHR data can be used for research purposes to generate new epidemiological insights or to evaluate clinical interventions that may serve to generate new evidence for inclusion in clinical guidelines. Fourthly, given the limited resources available to conduct a PhD, and the focus of quality improvement on changing health professional behaviour, exploring the experiences of patients when EHR data are used for quality improvement was not addressed. Finally, free text data is currently not available from primary care EHRs in a timely and efficient manner at the population-level, so only coded data were considered for analysis. This may be problematic because useful information not stored as clinical codes may be missed, however, there is evidence that using quality improvement systems
can increase the quality of recorded data [26], and therefore may partly address this issue.

2.3 Chapter conclusion

This chapter has described the overall theoretical approach taken during this PhD to address the aims and objectives set out in Chapter 1. In doing so, it has also addressed potential risks of the research plan and mitigating steps taken, in addition to considering what is beyond the scope of its enquiry. The next five chapters describe the ‘results’ of the PhD, starting with a critical examination of the empirical and theoretical literature in Chapter 3.

2.4 Chapter references


Chapter 3

The Case for Conceptual and Computable Cross-Fertilisation Between Audit & Feedback and Clinical Decision Support

Benjamin Brown,1,2 Niels Peek,1,2 Iain Buchan1,2

1 Health eResearch Centre, Farr Institute for Health Informatics Research, University of Manchester, UK
2 NIHR Patient Safety Translational Research Centre Greater Manchester, University of Manchester, Manchester, UK

Contributions
Author BB intellectually conceived the idea for the paper, searched the literature, and wrote the paper. Authors NP and IB reviewed versions of the manuscript, and provided critical feedback on its development.

Publication status

Chapter introductory note
This chapter presents a conceptual paper that represents my early thinking on how electronic health record (EHR) data could be put to better use. It addresses Research Objective 1 by critically reviewing the literature regarding Clinical Decision Support (CDS) and electronic Audit and Feedback (e-A&F) systems to make an argument for the conceptual union of their design. Many of these ideas were later used to inform the initial design of the Performance Improvement plaN GeneratoR (PINGR).
3.1 Abstract

Many patients do not receive care consistent with best practice. Health Informatics interventions often attempt to address this problem by comparing care provided to patients (e.g. from electronic health record data) to quality standards (e.g. described in clinical guidelines) and feeding this back to clinicians. Traditionally these interventions are delivered at the patient-level as clinical decision support (CDS) or at the population level as audit and feedback (A&F). Both CDS and A&F can improve care for patients but are variably effective; the challenge is to understand how their efficacy can be maximised. Although they are traditionally considered separate approaches, we argue that CDS and A&F share common mechanisms, and that their efficacy may be improved by cross-fertilizing relevant features and concepts. We draw on the Health Informatics and Implementation Science literature to argue this includes functions typically associated with the other, in addition to other features that may prove fruitful for further research.
3.2 Introduction

The prominence of evidence-based medicine has led to widespread acceptance of what constitutes good care. Implementing this evidence in clinical practice, however, is challenging – often referred to as the ‘second translational gap’ [1]. Such problems lead to adverse outcomes for patients: for example, in the UK alone there are thought to be over 3000 unnecessary strokes per year because patients with atrial fibrillation do not receive anticoagulant medication [2]. Barriers to implementing evidence-based care may occur at different levels: individual patient-practitioner level; provider team; provider organisation; and health system policy [3]. Health Informatics interventions often attempt to address these barriers by highlighting to clinicians when patients may not receive care consistent with best practice evidence (e.g. in clinical guidelines) through analysis of patient data (e.g. from electronic health records [EHR]). When these interventions are delivered via computers during clinical encounters with patients they are called “clinical decision support” (CDS). When delivered outside clinical consultations and at the population-level, they are typically described as “audit and feedback” (A&F). Systematic reviews of both types of intervention suggest they are moderately effective at ensuring patients receive improved care [4–6]. However, they also suggest they are highly variable: sometimes they work very well, and sometimes they do not [4–6]. The current challenge is therefore to understand how to maximise their efficacy. Although they are traditionally considered separate approaches, in this paper we suggest that electronic CDS and A&F share common mechanisms, and that their efficacy may consequently be improved by cross-fertilising relevant features and concepts between each other.

To build our argument, we draw on literature from Health Informatics and from Implementation Science. Despite our focus on computer-based tools, we also draw on relevant evidence from non-electronic interventions. First we examine electronic CDS and A&F separately: their functions, mechanisms and features associated with success. Next we consider their common aspects and provide a rationale for cross-fertilisation. Finally, we provide examples of how this could be achieved both with functions typically associated with the other, in addition to other features that may improve their success. We end with a discussion on implications for future research and other tools that facilitate human interpretation of patient data.

3.3 Clinical Decision Support

CDS (both electronic and non-electronic) refers to a heterogeneous set of tools that can be broadly defined as ‘active knowledge systems which use two or more items of patient data to generate case-specific advice’ [7]. This is contrasted with passive knowledge
systems, in which the user themselves must search the system [7]. Musen et al. classify CDS in three basic varieties [8]:

1. Patient-specific, situation-specific alerts, reminders, physician order sets, or other recommendations for direct action;
2. Information about the current clinical context to retrieve highly relevant online documents (e.g. ‘infobuttons’);
3. Organisation and presentation of information in a way that facilitates problem solving and decision making (e.g. graphical displays, documentation templates, structured reports).

We equate electronic CDS with the first variety. These are considered classic electronic CDS systems [8], and are arguably the most common [9,10]. Such systems provide custom-tailored assessments or advice based on patient-specific data (e.g. from EHRs or order entry systems) in consultation with a knowledge-base (usually best-practice evidence e.g. clinical guidelines), delivered via a computer to professionals at the point of care. Examples of these systems include [8,9]:

- Alerting clinicians if they are about to perform an action that may have adverse consequences (e.g. prescribing a macrolide antibiotic in a patient taking a statin);
- Reminding clinicians to perform a task (e.g. that a patient requires a cholesterol blood test);
- Suggesting management options for a particular patient based on their specific circumstances (e.g. suggesting changes in cholesterol-lowering treatment for a patient with high cholesterol).

Electronic CDS typically employs event-condition-action rules such as those in Arden syntax [8]. If a patient’s data (the “event”; e.g. cholesterol level), meets criteria in accordance with the knowledge-base (the “condition”; e.g. >5mmol/L), then the CDS is triggered (the “action”; e.g. suggestion of options for intensification of statin treatment). We exclude probabilistic CDS tools (e.g. Bayesian diagnostic systems) from our argument because they are usually not based on predefined clinical standards such as clinical practice guidelines.

We can surmise that electronic CDS attempts to improve compliance with best practice evidence by addressing barriers at the individual (patient-practitioner) level [3]. These include the health professional’s lack of awareness or familiarity with the evidence, or their inertia of previous practice [11]. CDS works by making information available and visible to the health professional during the clinical encounter when action can be taken. However,
CDS only works for patients that are encountered, and it is often ignored during time-pressured clinical encounters, or when a patient has an over-riding competing clinical priority. As a result, the efficacy of CDS is modest and highly variable. A recent Cochrane review of ‘on-screen, point of care computer reminders’ demonstrated they improved processes of care by a median of 4.2% (interquartile range [IQR], 0.8% to 18.8%) [12]. Another review found only 58% of trials demonstrated an improvement in either processes of care or patient outcomes [4]. This review also demonstrated that CDS is more likely to be effective if it is delivered outside the EHR or order entry system, provides advice to patients as well as health professionals, and requires the user to articulate why they ignored a recommendation [4]. It has also been suggested that CDS may be variably effective because it does not target organisation-level barriers [13].

3.4 Audit & Feedback

A&F (both electronic and non-electronic) can be defined as ‘any summary of clinical performance of health care over a specified period of time’ [14]. Other names for A&F include ‘clinical performance feedback, ‘performance measurement’ and ‘quality measurement.

The audit part of A&F involves analysing data to produce a summary measure of clinical performance (interchangeably called a ‘quality indicator’, ‘performance measure’, or some combination of the two). Data may be obtained from medical records, computerised databases, or observations from patients [14]. Clinical performance is judged for a specified population according to accepted best practice (e.g. clinical guidelines). Quality indicators usually quantify clinical performance in a Donebedian [15] classification as:

- Structural measures e.g. number of nurses on a ward;
- Process measures e.g. proportion of eligible atrial fibrillation patients on anticoagulation;
- Outcome measures e.g. proportion of diabetic patients with good glycaemic control (intermediate outcome) or number of myocardial infarctions per year per unit of population.

These are generally calculated as proportions by comparing individual patients’ data to the performance standard: the number of patients meeting the criteria form the numerator (e.g. number of children given a vaccination), and the total number eligible to meet the criteria form the denominator (e.g. number of children in the population eligible to receive the vaccine).
Feedback of audit results takes place after the clinical encounter, generally outside the clinical environment, and may target an individual, team or organisation [16]. It may be delivered in a written, verbal or electronic format, and may include supporting materials, such as suggestions for improvement [6,14]. Feedback may also include ‘benchmarking’ – comparison of recipients’ performance with colleagues [17].

Traditionally, A&F was laboriously undertaken by the health professionals using paper medical records [18]. However, widespread use of EHRs and web-based technologies means it is now much easier to undertake across multiple providers by external agencies such as governments or health service managers. As a result there is now an abundance of electronic A&F tools in healthcare systems around the world, variably termed ‘dashboards’, ‘benchmarking tools’, or ‘report cards’ [19]. Some are crude implementations of generic business intelligence software, others are more carefully developed for healthcare. These tools present information to health professionals (and often other audiences such as patients) via websites, computer applications or e-mail. Unlike non-electronic A&F, electronic A&F tools rarely make suggestions for improvement action to be taken by recipients.

Like CDS, A&F addresses barriers on the individual (patient-practitioner) level by making health professionals cognisant of their performance. As feedback is delivered outside the clinical encounter, it provides space and time for reflection and self-awareness, which ideally leads to behaviour change [20]. However, A&F also has the potential to address team and organisation-level barriers too, such as lack of resources and structural constraints [11] through the following ways:

- Feedback may be delivered to teams of clinicians and health care managers in addition to individual practitioners;
- Feedback provides recipients with a systematic and comprehensive view of entire patient populations served by a team or organisation, rather than only focusing on individual clinical encounters.

In addition to the space and time for reflection afforded by A&F, these factors encourage the formulation of service re-design plans for quality improvement. A limiting factor of A&F is that these plans must be formulated and undertaken, for which there must be sufficient time and resources. Consequently, like CDS, the efficacy of A&F is also modest and highly variable. The most recent Cochrane review of A&F demonstrated a median improvement in processes of care of 4.3% (IQR 0.5% to 16.0%) [6]. This review also demonstrated that A&F is more likely to be successful if the recipient is not performing
well at baseline, and if feedback is provided by a supervisor or senior colleague, regularly, in multiple formats with clear targets and an action plan [6]. It has also been suggested that A&F may be more effective if it includes individual patient-level data (in addition to population summaries) [21–23], individual clinician-level data (in addition to team- or organisation-level) [23,24], and if it is provided in a timely manner [24,25].

3.5 Rationale for Cross-Fertilisation

Although electronic CDS and A&F have traditionally been considered separate approaches to quality improvement, we argue the above evidence (summarised in Table 1) suggests they are in fact highly related for the following reasons:

1. **They use the same ‘substrates’**: Both interventions use EHR data and compare the observed clinical workflow against a clinical standard (e.g. guidelines).
2. **They use analogous analytic methods**: The number of event-condition-action rules triggered in electronic CDS systems for a specific patient population are equivalent to the numerator value of quality indicators in an electronic A&F system. The total number of patients for which the event-condition-action rules could be applied is equivalent to the denominator.
3. **They use similar methods to effect behaviour change**: Both feed back to recipients assessments of observed care versus a clinical standard.

As described above, it is already established that CDS and A&F are moderately effective at improving patient care. The current research challenge is to therefore understand how their efficacies can be maximised [4,20]. We argue that given their similarities, there is a need to explore potential, systematic cross-fertilisation and learning between these separate systems. As we demonstrate in the next section, there is evidence that cross-fertilisation of associated functionality is effective, which supports our assertion that computerised CDS and A&F are related. We argue that this relationship should be exploited in systematic, computable ways.
Table 1: Comparison of electronic CDS and A&F

<table>
<thead>
<tr>
<th>Feature</th>
<th>CDS</th>
<th>A&amp;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data source</td>
<td>EHR</td>
<td>EHR</td>
</tr>
<tr>
<td>Analytic method</td>
<td>Event-condition-action rules</td>
<td>Quality indicators</td>
</tr>
<tr>
<td>Unit of analysis</td>
<td>Individual patient</td>
<td>Population</td>
</tr>
<tr>
<td>Delivery</td>
<td>During clinical encounter</td>
<td>Outside/after clinical encounter</td>
</tr>
<tr>
<td>Users</td>
<td>Individual clinicians</td>
<td>Individual clinicians, teams, organisations</td>
</tr>
<tr>
<td>Recommends improvement actions</td>
<td>Sometimes</td>
<td>Rarely</td>
</tr>
<tr>
<td>Features associated with success</td>
<td>• Delivered outside EHR</td>
<td>• Low baseline performance of recipient</td>
</tr>
<tr>
<td></td>
<td>• Providing advice to patients</td>
<td>• Feedback provided by supervisor/senior</td>
</tr>
<tr>
<td></td>
<td>• Requiring over-ride reasons</td>
<td>• Regular feedback</td>
</tr>
<tr>
<td></td>
<td>• Targeting organisation-level barriers</td>
<td>• Multiple formats of feedback</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Clear targets and action plans fed back</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Individual level information (patient and clinician)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Timely information</td>
</tr>
</tbody>
</table>

3.6 Suggestions for Cross-Fertilization

Evidence for Synergies of Typical Functions

Electronic CDS is more effective if delivered separately from the EHR or ordering system [4], and may also be improved if it targets team and organisation-level barriers [13]. Both these are features typically associated with A&F (Table 1). For example, an electronic CDS system may only remind a user to order an annual thyroid function blood test for someone on long-term thyroxine if their EHR is opened during a clinical encounter. However, this only works for patients that are seen, and the reminder may be ignored in a time-pressured clinical environment [26]. If it was delivered outside the EHR, it would be possible to see all the patients who needed the blood test (like A&F), which may provide the time and space to formulate a plan to ensure they all had the blood test. Furthermore, providing this more population-based information may also help address some of the team and organisation-level barriers (like A&F), such as understanding that additional services may be needed to facilitate all the patients receiving the blood test e.g. providing additional phlebotomy clinics or new phlebotomy staff.

Electronic A&F is more effective if provided in a timely manner [24,25], with suggested action plans [6], individual patient-level data [21–23] and individual clinician-level feedback [23,24]. These are features typically associated with CDS (Table 1). For example, an electronic A&F system normally only highlights the proportion of hypertensive patients who have uncontrolled blood pressure. This requires searching for the patients and formulating action plans to improve performance, for which there may not be the
resources or skills [21]. Electronic A&F may therefore be improved if the summary also provides individual-level data on which patients have uncontrolled blood pressure, in addition to suggestions for improvement action (like some kinds of CDS). This may include individual patient actions such as choices for medication optimisation, but also team and organisation-level changes such as introducing a home blood pressure monitoring service or installation of a blood pressure machine in the clinic waiting room. Some electronic CDS systems already have similar functions to this termed “population registries”, however they are uncommon in practice [9] and only provide suggestions for individual patient actions. Furthermore, to address barriers at the individual patient-practitioner level, such as health professionals’ lack of awareness or familiarity with clinical guidelines, requires knowledge of which clinicians need targeting. This is facilitated if feedback reports specify individual clinician performance in addition to their team or organisation (like CDS) [24]. This may also be more effective if provided close to the time of the clinical encounter (like CDS) [24], when the experience is fresh in the clinician’s mind and the patient’s care is amenable to action, for example they have not left the hospital or moved address.

Conceptual Extension
The evidence above suggests that functions typically associated with A&F are likely to improve the efficacy of electronic CDS, and vice versa, which re-enforces our argument that the two interventions are related. It also suggests that cross-fertilisation of other features, for which there may not currently be supporting evidence, may be worth investigating. For example, computerised CDS may be more successful if it also provided population-level summaries like A&F during clinical encounters. One may hypothesise that when an electronic CDS system is triggered (e.g. for raised cholesterol), knowing the proportion of the eligible population for whom the alert would also fire (e.g. proportion of patients with high cholesterol) will put the information in a broader context of clinical performance, emphasising the alert’s importance. This may more effectively motivate the recipient to take action, reducing alert fatigue [27].

This principle may also extend to other features that are not necessarily considered typical of either CDS or A&F, but that have been shown to improve their efficacy (Table 1). For example, electronic CDS is more effective if it provides advice for patients and if it requires a reason for over-riding its advice [4]. To our knowledge these features have not been investigated extensively in computerised A&F, though may improve its efficacy: providing advice to patients may improve adherence to medication and engagement with healthcare in the same way as CDS; and requiring users to justify why feedback is ignored may improve cognitive engagement with the quality improvement process. We suggest these and other areas in Table 1 may prove fruitful areas for further research.
We acknowledge there may be other features of electronic CDS and A&F interventions that are associated with improved efficacy that have not been mentioned above. This may include characteristics for which there are currently conflicting opinions or evidence, such as the use of benchmarking in A&F [20,22,28,29], or other features not yet discovered. The application of behavioural change theory and use of qualitative evaluations in future research may help identify these additional features. This idea is beginning to gain traction in the A&F literature [20,29], but to our knowledge has not yet been mirrored with regard to CDS. We suggest that when new theories or features are discovered for either electronic CDS or A&F, they should be inclusively applied to both and their influence on efficacy assessed.

3.7 Discussion

We have studied the features of electronic CDS and A&F and argued that they should be considered as related rather than separate approaches to healthcare quality improvement. In doing so, we have suggested their efficacy may be improved through the cross-fertilisation of features typically associated with the other, and that future research should explore linking the two in a computable synergy.

Previous attempts to increase the understanding of (the effectiveness of) CDS and A&F have largely been limited to bottom-up aggregation of empirical evidence concerning a heterogeneous set of intervention studies, with little success to date [30]. Our approach advocates consideration of the mechanisms that underpin how they work and what they are trying to achieve, which we think will be more successful.

Our argument is not that electronic CDS and A&F should be used as ‘multifaceted’ or ‘co-’ interventions; these terms suggest separate tools glued together. Our vision is that given their similarities these interventions should seamlessly incorporate successful features and other learning from each other. Interestingly, in support of this assertion, merely adding reminders to A&F has not been shown to impact efficacy [6], nor adding summary feedback to CDS [4], and there is little support from systematic reviews for multifaceted over single-component interventions in general [31]. Furthermore, we do not advocate that the empirical evidence relating to electronic CDS and A&F be simply combined to increase the power of meta-regressions in systematic reviews, as this belies the clear differences between them. Our argument is rather that a shared conceptualisation can strengthen the generation and testing of specific hypotheses that draw on their shared mechanisms. We believe our understanding of both interventions can be advanced by considering the empirical evidence across them, borrowing evidential strength from adjacent areas where appropriate.
In addition to improving the efficacy of both electronic CDS and A&F, there are corollary benefits to their cross-fertilisation. For example, it encourages the development of common technical standards, which promises to save implementation time and effort [32]. Linking quality indicators to possible improvement actions also provides a more accurate measure of care quality, which is important if used for accountability purposes (e.g. performance-based payment). For example, it is more accurate to know the proportion of uncontrolled hypertensive patients prescribed suboptimal medication, rather than simply the proportion of uncontrolled hypertensives.

Although we have limited our discussion to CDS and A&F, our argument may well extend to a broader set of electronic interventions that facilitate clinician interpretation of patient data against pre-defined standards of clinical quality. Examples include range checks for laboratory test results, electronic checklists, care pathways, and risk prediction tools. And although we excluded probabilistic CDS systems from our discussion, they may also be relevant. For example there are arguments for the application of risk prediction tools in A&F [21] and a need for actionable suggestions (like some CDS) in risk prediction [33]. Furthermore, there may be implications for non-electronic interventions too, as there are suggestions that A&F is most effective when there is external facilitation [34], which may be considered an ‘educational outreach’ feature [14].

A limitation of our argument is that it has, to enable wider conceptual arguments, drawn on evidence from non-electronic CDS and A&F. Future research should therefore empirically test whether our suggestions for cross-fertilisation hold in electronic settings. It should also test extended cross-fertilisation hypotheses as outlined above, which should be informed by relevant theoretical models. Our group has already started to develop these ideas into deliverable interventions to test a range of hypotheses [35,36].

3.8 Conclusion

We argue that electronic CDS and A&F systems are not separate but highly related approaches to quality improvement. We suggest that cross-fertilisation of features and learning between them may improve their efficacy. We have provided examples of how this may be achieved in computable ways, along with suggestions for future research.

3.9 References


Ross Scrivener, C. Morrell, R. Baker, S. Redsell, E. Shaw, K. Stevenson, et al.,


[21] N. Ivers, J. Barnsley, R. Upshur, K. Tu, B. Shah, J. Grimshaw, et al., My approach to this job is ... one person at a time, Can Fam Physician. 60 (2014) 258–266.


Chapter 3 concluding note

This chapter addressed Research Objective 1 by presenting a critical review of the literature regarding CDS and e-A&F systems to make an argument for combining their features in order to increase their effectiveness. One key recommendation was that e-A&F systems should provide recommendations for clinical action to health professionals, as is the defining feature of PINGR. The next chapter extends this conceptual thinking further, with a particular focus on qualitative evaluations of A&F interventions in the published literature.
Chapter 4

Clinical Performance Feedback Intervention Theory (CP-FIT): A Meta-synthesis of Qualitative Studies

Benjamin Brown,1,2 Wouter T. Gude,3 Thomas Blakeman,2 Gavin Daker-White2
1 Health eResearch Centre, Farr Institute for Health Informatics Research, University of Manchester, UK
2 NIHR Patient Safety Translational Research Centre Greater Manchester, University of Manchester, Manchester, UK
3 Department of Medical Informatics, Academic Medical Center, University of Amsterdam

Contributions
BB intellectually conceived the idea for the study, designed the study, collected and analysed the data, and wrote the manuscript. WG contributed to data collection and analysis, and reviewed versions of the manuscript. TB and GDW contributed to study design, data analysis, and reviewed versions of the manuscript.

Publication status
A version of this paper is currently being prepared for submission to PLOS Medicine.

Chapter introductory note
This chapter addresses Research Objective (RO) 1 by developing a new theoretical model to explain causal pathways of effectiveness in Audit and Feedback (A&F) interventions. It does this by conducting a systematic search and meta-synthesis of qualitative studies. This is a key chapter because the both its emerging and finalised results were used to guide the design and evaluation of the Performance Improvement plan Generator (PINGR) as described in chapters 5-7.
4.1 Abstract

**Background:** Audit and Feedback (A&F) is a widely used quality improvement technique in health care, though multiple Cochrane reviews have demonstrated its effects as highly variable. The application of theory may help explain this variation and understand how to maximise A&F effectiveness, however, existing theories lack detail and specificity to health care.

**Aim:** To develop Clinical Performance Feedback Intervention Theory (CP-FIT), a theoretical model to understand the causal pathways in A&F effectiveness.

**Design:** Systematic review and meta-synthesis of findings from qualitative studies of A&F interventions.

**Method:** Qualitative studies were identified in MEDLINE, EMBASE, CINAHL, Web of Science, and Google Scholar published from inception until 2016 inclusive. Data were extracted and analysed through line-by-line coding of individual studies. They were then synthesised using Realistic Evaluation and Framework Analysis to build causal explanations for A&F intervention effectiveness. Existing theoretical models related to A&F were used to explain findings and generalise across studies.

**Results:** 15413 potential papers were identified, of which 65 were synthesised, reporting studies of 73 different A&F interventions in 24 different countries. Effective A&F is a cyclical process of Goal setting, Audit, Feedback, recipient Interaction, Perception, and Acceptance of the feedback, followed by Intention and Behaviour. Progress round this cycle is influenced by variables relating to characteristics of the Goal, Audit methods, Feedback message, Implementation process, Organisational context, Co-interventions, Health professional, and Patient population, that exert their effects via mediators relating to Actionability, Resource match, Complexity, Relative advantage, Compatibility, Credibility, and Social influence. Unintended outcomes of A&F include Gaming and Tunnel vision.

**Discussion:** Based on our findings, CP-FIT has three propositions: 1) A&F interventions exert their effects by inducing patient-level behaviours in health professionals; 2) Health care organisations have limited capacity to engage with and respond to the demands of A&F interventions; and 3) Health care professionals and organisations have a strong set of beliefs and behaviours regarding how they provide patient care that influence their interactions with A&F. CP-FIT also suggests that A&F interventions have two key
mechanisms by which they exert effects: facilitating Direct Action and raising Knowledge/Awareness. A&F interventions that maximise their effects via the Direct Action are most effective.

**Conclusion**: CP-FIT is a new theoretical model of causal pathways in A&F interventions that can be used to design, implement, and evaluate interventions. It provides tentative explanations for the observed variation in Cochrane reviews, though future research should test and refine its hypotheses.
4.2 Introduction

Audit and Feedback (A&F) is a widely used quality improvement technique that has been used for decades as a strategy to implement evidence-based care into clinical practice [1]. In A&F, a summary of health professional's clinical performance is provided to them with the intention of changing their practice, however the mechanisms by which this works are poorly understood [2]. Three Cochrane reviews of randomised controlled trials (RCTs) of A&F interventions have found they produce “small but potentially important improvements in professional practice” [3]. However, this effect is highly variable: the most recent review found A&F confers a median increase in desired processes of care of 4.3%, ranging from -9% to 70% [4].

A lack of theory and inadequate exploration of underlying mechanisms in A&F research has been cited as a key reason for sub-optimal effectiveness [2]. Analysis of the 140 RCTs included in the most recent Cochrane review found only 20 (14%) reported using theory, and concluded that “no consensus exists among empirical researchers on how to approach A&F from a theoretical perspective” [5]. Consequently, policymakers, practitioners and researchers designing A&F have only a tentative set of best practices regarding how it can be conducted most effectively [2,6].

Three main theories of the effects of feedback on human behaviour have gained popularity in A&F research recently [2]: Control Theory [7], Feedback Intervention Theory [8] and Goal Setting Theory [9]. However, Control Theory does not describe potential factors that may influence the ‘success’ of feedback, and all three ignore potentially important determinants of quality improvement interventions in health care settings, such as organisational context [10]. Both Control Theory and Feedback Intervention Theory informed the analyses performed in the most recent Cochrane review [4], in addition to other systematic reviews of A&F interventions [11,12], with mixed success. Some observed findings could not be explained by either theory [4,11], whereas others were inconsistent with their predictions [12]. Furthermore, when these theories have been used to prospectively design A&F interventions, they have been found to be ineffective (e.g. [13]). Consequently, one conclusion may be that these existing theories are not sufficiently detailed, nor comprehensive enough to adequately model A&F mechanisms.

Qualitative research tends to generate theory and contribute detail on what, how and why certain factors may influence the effectiveness of complex interventions [14]. Meta-syntheses of such studies can be used to develop theoretical models to provide useful insights into their mechanisms and how they may be optimised [15]. Examples include interventions for tuberculosis therapy [16], smoking cessation [17], the prevention of skin cancer [18], lay health worker programmes [19], and telephone counselling [20]. A similar
approach may therefore be both useful and feasible for A&F, and to the best of our knowledge has not yet been attempted.

**Aim and objectives**

We sought to develop a theoretical model (Clinical Performance Feedback Intervention Theory; CP-FIT) of causal pathways in A&F effectiveness through a systematic review and meta-synthesis of qualitative research studies. Our primary aim was to provide empirical guidance on how to optimally design and implement effective A&F interventions. A secondary aim was to use the model to ‘extend’ [21] the most recent Cochrane review of A&F [4] by providing potential explanations for the observed variability in effectiveness.

In accordance with guidance on developing and evaluating complex interventions in health care [22,23], the specific objectives of this review were to understand how:

1. A&F interventions are implemented into practice to produce improvements in patient care;
2. Health professionals respond to, and interact with, A&F interventions;
3. Unintended consequences may arise from undertaking A&F;
4. Features of A&F interventions and the context into which they are implemented affect their outcomes.

**Why it is important to do this review**

A&F is used in health care settings around the world, and technological advances such as electronic health records (EHRs) and web-based systems have made it easier and less expensive to undertake [24]. Therefore, understanding how to optimise the effectiveness of A&F has the potential to improve care for large numbers of patients, in addition to reducing the opportunity cost from unsuccessful A&F interventions.

**4.3 Methods**

We followed guidance from the Cochrane Qualitative and Implementation Methods Group [21,25], and published our protocol on the International Prospective Register of Systematic Reviews (PROSPERO; registration number CRD42015017541 [26]). In reporting our findings, we followed both PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) [27] and ENTREQ (enhancing transparency in reporting the synthesis of qualitative research) statements [28]. Our international multidisciplinary research team has extensive experience in qualitative research and its metasynthesis, and A&F (as researchers, practitioners, and recipients) [29].
Search strategy

We intended to find all published qualitative evaluations of A&F interventions by replicating the latest Cochrane review’s search strategy [4], substituting terms for RCTs with qualitative research filters optimised for sensitivity [30–32]. We searched MEDLINE (Ovid), EMBASE (Ovid) and CINAHL (Ebsco) without time limits on 25th March 2015. An information specialist evaluated and provided advice regarding our search strings (Appendix 1). ‘Supplementary’ [25] citation, related article, and reference list searches were undertaken up to 31st December 2016 for all included studies, RCTs in the latest A&F Cochrane review [4], and relevant reviews and essays (e.g. [2,5,6,11,12,33–39]). Citation and related article searches were performed using Web of Science (core collection), limited to the first 100 results to avoid papers of limited relevance [40]. Reference list searching involved manually examining bibliographies. Further studies were found through contact with international experts in A&F [2] and Google Scholar alerts activated during the review process.

Inclusion criteria

Table 1 describes our inclusion criteria, which were informed by the latest Cochrane review [4], guidance from Cochrane’s Effective Practice and Organisation of Care (EPOC) Group [41], and specialist A&F organisations [1,42,43]. As our objectives included understanding the influence of context and the implementation process on A&F feature effectiveness, we only included papers that studied specific A&F interventions in routine clinical practice, as opposed to those in training or simulated environments, or those that discussed the general concept of A&F.

Table 1: Inclusion criteria

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Typical exclusion examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td></td>
</tr>
<tr>
<td>The intervention primarily targeted health care professionals (including clinicians and non-clinicians e.g. managers).</td>
<td>Feedback intended to help patients choose health care provider or treatment (e.g. [44]).</td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td></td>
</tr>
<tr>
<td>The intervention provided feedback to participants.</td>
<td>Audit reports (e.g. [45]); pay-for-performance programmes where feedback was not explicitly provided (e.g. [46]).</td>
</tr>
<tr>
<td>Feedback primarily concerned health professionals’ clinical performance in work-based clinical settings, defined as compliance with pre-defined clinical standards (e.g. clinical guidelines) and/or clinical patient outcomes. This may have referred to the performance of an individual, their team, or organisation.</td>
<td>Interventions that provided only: fictitious feedback (e.g. [47]); feedback used in training or simulated settings (e.g. [48]); feedback on non-clinical aspects of performance or data not directly related to clinical performance, such as costs of care (e.g. [49]), patient experience (e.g. [50]), or epidemiological surveillance (e.g. [51]).</td>
</tr>
<tr>
<td>Clinical performance data were primarily obtained from medical records, computerised databases, or observations from patients.</td>
<td>Clinical performance feedback based only on peer or supervisor observation (e.g. [52]).</td>
</tr>
<tr>
<td>Feedback primarily relates to the care of multiple patients.</td>
<td>Feedback focused on the care of individual patients, such as: reminder or alert systems (e.g. [53]); patient-level summaries (e.g. [54]); significant event analyses or case reviews (e.g. [55]).</td>
</tr>
</tbody>
</table>
Feedback could inform quality improvement actions for teams or organisations, not solely for individual patients. [55].

Improvement collaboratives that primarily consisted of mentoring visits, improvement advisors, and educational sessions, with ‘benchmarking’ as an additional component (e.g. [57]).

| Feedback is a core and essential component of the intervention i.e. in multifaceted interventions, it is unlikely other components would have been offered in the absence of feedback. | Dashboards that summarised patients’ current clinical status to primarily inform point-of care decisions (e.g. [56]). |

<table>
<thead>
<tr>
<th>Comparator</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not applicable</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

**Comparator**
- Not applicable

**Outcome**
- The intervention primarily aimed to improve clinical performance as defined above.
- Interventions that primarily intended to reduce costs (e.g. [49]).

**Study**

<table>
<thead>
<tr>
<th>Studies of interventions described in enough detail to determine whether they met the above criteria.</th>
<th>Studies of groups of interventions, the characteristics of which are not clearly described. For example, studies of ‘audit’ in general (e.g. [58]).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluations of interventions that reported both qualitative data collection (e.g. semi-structured interviews, focus groups, unstructured observations) and analysis methods (e.g. ethnography, grounded theory, thematic analysis, framework analysis). To judge adequacy, studies must have provided either a full methodological description, or reference to a specific relevant approach.</td>
<td>Studies reporting interviews or focus groups but no description of analytic methods (e.g. [59]); intervention descriptions or protocol papers (e.g. [60]); editorials or opinion papers (e.g. [61]); quantitative surveys with or without open ended questions (e.g. [62]); manuscripts with insufficient detail to judge adequacy, such as abstracts or letters (e.g. [63]).</td>
</tr>
</tbody>
</table>

**Comparator**
- Not applicable

**Outcome**
- The intervention primarily aimed to improve clinical performance as defined above.
- Interventions that primarily intended to reduce costs (e.g. [49]).

**Study**

<table>
<thead>
<tr>
<th>Studies of interventions described in enough detail to determine whether they met the above criteria.</th>
<th>Studies of groups of interventions, the characteristics of which are not clearly described. For example, studies of ‘audit’ in general (e.g. [58]).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluations of interventions that reported both qualitative data collection (e.g. semi-structured interviews, focus groups, unstructured observations) and analysis methods (e.g. ethnography, grounded theory, thematic analysis, framework analysis). To judge adequacy, studies must have provided either a full methodological description, or reference to a specific relevant approach.</td>
<td>Studies reporting interviews or focus groups but no description of analytic methods (e.g. [59]); intervention descriptions or protocol papers (e.g. [60]); editorials or opinion papers (e.g. [61]); quantitative surveys with or without open ended questions (e.g. [62]); manuscripts with insufficient detail to judge adequacy, such as abstracts or letters (e.g. [63]).</td>
</tr>
</tbody>
</table>

**Study selection**

Titles and abstracts were independently screened by lead author (BB) and one other reviewer. Full manuscripts of citations marked as potentially relevant by at least one reviewer were obtained and the inclusion criteria independently re-applied. Disagreements were resolved through discussion, and consultation with the wider review team as necessary. Papers found during citation and related article searches were screened by BB, and any potentially relevant full manuscripts reviewed by another reviewer.

**Data extraction**

Data from included studies were extracted independently by BB and WG, regarding study and A&F intervention details, and study findings. Study details included: aims, setting, main findings, use of theory, strengths, limitations, and research parameters listed in the consolidated criteria for reporting qualitative research (COREQ) [65]. Intervention details included: aims, whether it had been judged a success (and how this was reached), a description of its design and implementation according to behavior change techniques [66] and 17 known modifiable elements of A&F interventions [67]. Sister studies (e.g. RCT results or protocols) found during supplementary searches were used to add detail. Both study and intervention details were entered into individual Microsoft Word documents (Appendix 2). Study findings included both first order (direct quotations from participants) and second order (author interpretations) constructs [68]. These were primarily found in
the abstract, results and discussion sections. Any disagreements between reviewers regarding extracted data were resolved through discussion, with the wider review team consulted as necessary.

Quality appraisal
Studies were appraised by BB and WG on both technical and theoretical aspects [25] using Walsh and Downe’s criteria [69], which occurred concurrently with data extraction (Appendix 2). Judgments were made regarding whether 12 essential quality criteria were met, aided by 53 specific prompting questions. Papers meeting all 12 criteria had no methodological limitations; those meeting nine to 11 were deemed to have ‘minor’ limitations; five to eight had ‘moderate’ limitations; and four or less had ‘major’ limitations [70]. Appraisal results were used to weight study findings during synthesis, rather than exclude low quality studies, to maintain theoretical richness [25,71].

Data synthesis
We used Framework Analysis [72] informed by Realistic Evaluation [73] to synthesise study findings. Realistic Evaluation conceptualises that interventions introduced into certain contexts, trigger particular mechanisms to generate specific outcomes [73]. Framework Analysis enables concurrent coding of these constructs, and encourages exploration of relationships between them to identify causal pathways [72]. The synthesis process was iterative, and took place in three main stages: coding of individual studies, generalisation across them, then model consolidation and confidence assessment (Figure 1).

Figure 1: Synthesis process
Stage 1: Coding of individual papers

Study findings were independently coded in batches of four by BB and WG. Each manuscript was read line-by-line to identify phrases representing first and second order constructs relevant to our research aim and objectives. Passages were entered as separate rows in a Microsoft Excel spreadsheet and their associated codes in adjacent columns. As per Realistic Evaluation [73], the initial coding framework consisted of columns for features of the A&F intervention, study context, resultant outcomes, and potential explanatory mechanisms. Consequently, potential causal mechanisms from original papers were preserved [74]. Intervention codes described aspects of intervention design or implementation extracted as described in section 2.4. Context codes described anything that existed prior to the introduction of the A&F intervention, such as setting characteristics or participants’ personalities [75]. Outcome codes described a consequence of introducing the intervention in a particular context, which could be intended or unintended, and could differ depending on the A&F intervention aim. Mechanism codes explained how aspects of the A&F intervention or its implementation resulted in a particular outcome in a specific context [73]. An additional column captured memos where reviewers could add notes. For the first batch of manuscripts, codes were developed inductively from the data, whereas in subsequent batches, a codebook from the emerging model was used (see section 2.5.2 below). After coding each batch of four manuscripts, reviewers compared findings and resolved discrepancies through discussion.

Stage 2: Generalisation across studies

The consolidated coding results and papers from each batch were discussed with two further reviewers who had independently read the papers (GDW and TB), with the aim of generalising findings from individual studies into a ‘middle range’ theory [73]. This involved a modified version of Analytic Induction where tentative explanations for observed outcomes were compared across studies [76]: positive cases supported an explanation, whereas deviant cases led to their adjustment or abandonment. A supporting codebook and narrative description of the emerging model were written, with diagrams and supporting data to develop and illustrate ideas [77]. Relevant constructs from existing theories, models and frameworks were incorporated to help describe codes and explain observed findings [36]. Candidate theories were found from included papers, supplementary searches (section 2.1), contact with experts, relevant papers (e.g. [78]) and a separate literature search conducted by lead reviewer BB following the methodology detailed by Booth and Carroll [79]. Each was assessed, initially by BB, against three criteria distilled from Cochrane guidance [80]:

1. Explanatory power: Does the theory explain phenomena of interest?
2. Clarity: Does the theory contain unambiguous concepts?

3. Testability: Are the theoretical propositions empirically testable?

All three criteria had to be met, and agreed with the rest of the review team, in order to be included in the synthesis. The prototype codebook was then used to code the next batch of studies in a new Stage 1 with amendments made or new concepts created as necessary. This cycle between Stage 1 and Stage 2 served to iteratively test and improve the emerging model, early versions of which were presented at national and international A&F research meetings, further refining and clarifying findings.

Stage 3: Model consolidation and confidence assessment

When all papers had been synthesised, the agreed codes and supporting data were transferred into Nvivo (version 10; QSR International) for further management. Lead author BB re-read all supporting data, re-coding earlier findings as necessary according to the final agreed model. Data were then interrogated using Matrix Coding queries to assess the confidence in each of the model’s findings according to the Confidence in the Evidence from Reviews of Qualitative Research (CERQual) method [70]. Each group of studies supporting a finding were assessed for: methodological limitations, relevance, coherence, and adequacy [70]. Based on these criteria, an overall judgment of the confidence in each finding was made: high, moderate, low, or very low [70]. The final codebook, narrative description, and confidence in the findings of the model were discussed and agreed with the entire authorship team.

4.4 Results

Retrieved studies

In total, we identified potential 15413 potential papers, of which 65 were ultimately synthesised. They reported the results of 61 different studies of 73 different A&F interventions, involving 1699 unique participants. Figure 2 demonstrates the flow of papers through the screening process.

Study characteristics

Table 2 summarises the main characteristics of included papers, full details of which are provided in Appendix 3. The median date of publication was 2013 (range 1996-2016). Papers most frequently reported studies set in hospital (n=30, 46%) and primary care (n=28, 43%), based in England (n=18, 28%), Canada (n=12, 18%), and the US (n=11, 17%), though studies were also reported from 21 additional other countries. A&F interventions were most commonly delivered as paper documents or face-to-face, to physicians and nurses, focusing on chronic care, patient experience and medication prescribing. Overall, papers reported a median of 14 out of 17 (82%) essential design
elements of A&F interventions studied (range 4-17) [67]. The most common design elements not reported were whether or graphical elements were used in the feedback (n=30, 46%), and the rationale for using A&F to improve care quality (n=31, 48%).
Figure 2: Flowchart of screening process of studies

- Records identified through database searching (n = 8576)
- Records after duplicates removed (n = 8413)
- Records screened (n = 8413)
- Full-text articles assessed for eligibility (n = 347)
  - Full-text articles excluded (n = 318)
    - Not a specific implemented A&F intervention (n = 188)
    - Not a qualitative study (n = 63)
    - A&F not core and essential (n = 46)
    - Not a full peer-reviewed publication (n = 12)
    - Not English (n = 9)
- Studies meeting inclusion criteria (n = 29)
- Records excluded (n = 8049)
- Articles screened from citation, reference list and related article searching, Google Scholar alerts, and contact with experts (n = 6837)
- Studies excluded (n = 6629)
- Additional full-text articles assessed for eligibility (n = 208)
  - Full-text articles excluded (n = 172)
    - Not a specific implemented A&F intervention (n = 84)
    - Not a qualitative study (n = 60)
    - A&F not core and essential (n = 28)
- Additional studies meeting inclusion criteria (n = 36)
<table>
<thead>
<tr>
<th>Continent</th>
<th>Count (%)</th>
<th>A&amp;F topic</th>
<th>Count (%)</th>
<th>Target A&amp;F recipient</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>37 (57)</td>
<td>Chronic care (general)</td>
<td>15 (23)</td>
<td>Physicians</td>
<td>45 (69)</td>
</tr>
<tr>
<td>North America</td>
<td>22 (34)</td>
<td>Patient experience</td>
<td>14 (22)</td>
<td>Nurses</td>
<td>40 (62)</td>
</tr>
<tr>
<td>Africa</td>
<td>2 (3)</td>
<td>Prescribing</td>
<td>11 (17)</td>
<td>Non-clinicians</td>
<td>24 (37)</td>
</tr>
<tr>
<td>Australia</td>
<td>2 (3)</td>
<td>Health care structures</td>
<td>10 (15)</td>
<td>Surgeons</td>
<td>6 (9)</td>
</tr>
<tr>
<td>South America</td>
<td>2 (3)</td>
<td>General nursing</td>
<td>8 (12)</td>
<td>Allied clinicians</td>
<td>6 (9)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting</th>
<th>Count (%)</th>
<th>A&amp;F topic</th>
<th>Count (%)</th>
<th>Target A&amp;F recipient</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital inpatient</td>
<td>30 (46)</td>
<td>Surgery</td>
<td>7 (11)</td>
<td>Junior physicians</td>
<td>3 (5)</td>
</tr>
<tr>
<td>Primary care</td>
<td>28 (43)</td>
<td>Cancer</td>
<td>5 (8)</td>
<td>Midwives</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Hospital outpatient</td>
<td>3 (5)</td>
<td>Diabetes</td>
<td>5 (8)</td>
<td>Pharmacists</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Nursing home</td>
<td>3 (5)</td>
<td>Stroke</td>
<td>5 (8)</td>
<td>Pathologists</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Mental health</td>
<td>1 (2)</td>
<td>Obstetrics</td>
<td>5 (8)</td>
<td>Radiologists</td>
<td>1 (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Preventive care</td>
<td>4 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infectious disease</td>
<td>3 (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Patient demographics</td>
<td>2 (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Staff experience</td>
<td>2 (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intensive care</td>
<td>2 (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mental health</td>
<td>1 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>General surgery</td>
<td>1 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Heart failure</td>
<td>1 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Orthopaedics</td>
<td>1 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Paediatrics</td>
<td>1 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physiotherapy</td>
<td>1 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rheumatology</td>
<td>1 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cost</td>
<td>1 (2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Each category may add up to more than 100%, as papers may have multiple features within each one.*
Quality appraisal

Table 3 summarises the quality appraisal of included papers. All papers had methodological limitations, with the majority (n=47, 72%) classified as ‘moderate’. Nine (14%) papers each were deemed to have ‘minor’ and ‘major’ limitations. Aspects of quality most poorly demonstrated included consistency between method design and research intent, researcher reflexivity, and sensitivity to ethical concerns.

Table 3: Quality appraisal results (using criteria from [69])

<table>
<thead>
<tr>
<th>Quality criterion</th>
<th>Number of papers meeting criterion (%)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear statement of, and rationale for, research question/aims/purposes</td>
<td>63 (75)</td>
</tr>
<tr>
<td>Study thoroughly contextualised by existing literature</td>
<td>58 (89)</td>
</tr>
<tr>
<td>Method/design apparent, and consistent with research intent</td>
<td>8 (12)</td>
</tr>
<tr>
<td>Data collection strategy apparent and appropriate</td>
<td>51 (78)</td>
</tr>
<tr>
<td>Sample and sampling method appropriate</td>
<td>25 (38)</td>
</tr>
<tr>
<td>Analytic approach appropriate</td>
<td>34 (52)</td>
</tr>
<tr>
<td>Context described and taken account of in interpretation</td>
<td>17 (26)</td>
</tr>
<tr>
<td>Clear audit trail given</td>
<td>48 (74)</td>
</tr>
<tr>
<td>Data used to support interpretation</td>
<td>56 (86)</td>
</tr>
<tr>
<td>Researcher reflexivity demonstrated</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Demonstration of sensitivity to ethical concerns</td>
<td>8 (12)</td>
</tr>
<tr>
<td>Relevance and transferability evident</td>
<td>59 (91)</td>
</tr>
</tbody>
</table>

*Papers did not meet criterion either when it was evident or unclear

Theories considered and used

We considered using constructs from 48 different theories to help develop codes and explain observed findings in included papers (Appendix 4). Ultimately we used constructs from 18 theories (Table 4), relating to: feedback mechanisms [7,8,81,82], goal-setting and motivation [9], psychology [83–85], organisational context and implementation [86–88], sociology [89,90], guideline adherence [91,92] and general behaviour change [93].
### Table 4: Theories included in the synthesis

<table>
<thead>
<tr>
<th>Feedback theories</th>
<th>Ilgen's model of feedback [81]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feedback intervention theory (FIT) [98–100]</td>
</tr>
<tr>
<td></td>
<td>Control theory (CT) [7]</td>
</tr>
<tr>
<td>Goal setting and action planning theories</td>
<td>Goal setting theory [9]</td>
</tr>
<tr>
<td>Guideline adherence theories</td>
<td>Cabana guideline model [91]</td>
</tr>
<tr>
<td></td>
<td>Guidelines interdependence model [92]</td>
</tr>
<tr>
<td>General behaviour change frameworks</td>
<td>Theoretical Domains Framework (TDF) [78,93]</td>
</tr>
<tr>
<td>Psychological theories</td>
<td>Cognitive dissonance [85]</td>
</tr>
<tr>
<td></td>
<td>Self Affirmation Theory [83]</td>
</tr>
<tr>
<td></td>
<td>Persuasion theory [84]</td>
</tr>
<tr>
<td>Technology theories</td>
<td>Value chain of information [82]</td>
</tr>
<tr>
<td>Context and implementation theories</td>
<td>Diffusion of innovations [87,137,138]</td>
</tr>
<tr>
<td></td>
<td>Consolidated framework for implementation research (CFIR) [86]</td>
</tr>
<tr>
<td></td>
<td>Multilevel quality improvement approach [88]</td>
</tr>
<tr>
<td>Sociological theories</td>
<td>Social comparison [89]</td>
</tr>
<tr>
<td></td>
<td>Reference group behaviour [90]</td>
</tr>
</tbody>
</table>
Synthesised results
We coded 1369 phrases from included papers. Our final codebook with references to supporting papers and CERQual ratings are provided in Appendix 5. To maintain readability, in this section we only discuss findings with a CERQual ‘high’ confidence rating unless stated otherwise, and provide key references to included papers only. All other lower confidence findings are listed in Appendix 5. Box 1 provides two case studies of A&F interventions, with descriptions regarding how our codebook explains their evaluations. Box 2 provides example quotes from included studies illustrating key concepts discussed below.

A&F processes and outcomes
Overall, we found that effective A&F consists of a series of sequential processes (Figure 3) consisting of: choosing desirable aspects of clinical performance against which health professionals will be measured (Goal setting); collection and analysis of clinical performance data on a defined population of patients (Audit); communication of measured clinical performance to health professionals (Feedback); who receive, comprehend, and accept the feedback message (Interaction, Perception, and Acceptance respectively); then undertake a planned behavioural response (Intention and Behaviour); ultimately leading to improved clinical performance (Clinical performance). The cycle then repeats generally starting again with further Audit. In certain situations, Goal setting may be revisited after Audit has taken place if (moderate confidence): there are problems collecting the clinical performance data (e.g. [155]), the goals are set too high to be achievable (e.g. [156]), or current clinical performance is so high that further improvement is unlikely (e.g. [157]). A further transgression occasionally occurs between Perception and Acceptance, where an A&F recipient interrogates the data underlying their feedback message in order to understand whether it presents a true picture of their clinical performance (Verification; e.g. [103]) – consequently this is most likely to happen when feedback reports a recipients’ performance as suboptimal (Performance level), or if they believe the underlying data to be inaccurate (Accuracy; moderate confidence).

Although the main intended outcome of A&F interventions is to improve clinical performance, two main negative unintended outcomes may also arise (Box 1): Gaming and Tunnel vision (both moderate confidence). In Gaming, health professionals may unethically manipulate clinical data or change their patient population in order to artificially improve their measured clinical performance. Whereas in Tunnel vision health professionals may excessively focus on the topic against which clinical performance is measured, to the detriment of other clinical areas. This may manifest during the care of individual patients or during quality improvement activities. A number of factors may influence the likelihood of unintended consequences (all low confidence) including the use
of payments to recipients (Financial rewards), reporting the feedback message to external bodies (Reporting), or perceived punishment related to suboptimal clinical performance (Function). Collectively they may be explained by a variety of mechanisms including the desire to gain financial income (Resource match) or to maintain a self and group perception as a high performing clinician (Compatibility [85] and Social influence).

Facilitating and inhibiting factors
A&F may become less effective if any process (described above) fails, for example if an audit is not conducted (e.g. [141]), or a recipient does accept the feedback message (e.g. [154]). Whether a process fails or is successful is determined by 68 different ‘moderating’ variables (Appendix 5) [158], relating to eight high-level factors: the clinical topic and expected standard of performance chosen on which to focus (Goal); how clinical performance data are collected and analysed (Audit methods); the design and delivery of the feedback (Feedback message); how the A&F is implemented into practice (Implementation process); the presence and nature of any additional quality improvement interventions (Co-interventions e.g. financial incentives); and characteristics of the recipient of the feedback (Health professional), the health care organisation or team (Organisational context), and the patients targeted by the A&F (Patient population). These factors influence A&F processes via seven key ‘mediating’ variables [158], largely adapted from existing theories and models: the feedback message’s ability to induce behaviour in A&F recipients (Actionability e.g. [9]); whether there is enough resource for the A&F intervention (Resource match e.g. [86]); the difficulty of implementing the A&F intervention (Complexity e.g. [86]); the perceived benefits of the A&F intervention over alternative ways of working (Relative advantage e.g. [137]); the degree of ‘fit’ between the A&F intervention and the recipient and their organisation (Compatibility e.g. [137]); the perceived trustworthiness and reliability of the A&F intervention (Credibility e.g. [81]); and social influences between A&F recipients, such as competition (Social influence e.g. [89]).

In general, moderating variables have direct effects on A&F processes and outcomes, however some may also affect other variables via ‘ripple effects’ [159] (Appendix 5; Figure 3) – for example Feedback messages that are delivered to a team rather than solely to individual recipients (Delivery to a group), may improve the ability of the organisation to work as a team (Teamwork) via Social influence [84].

Below, we describe key moderating variables organised under the main mediating variables through which they exert their effects, and grouped according to their moderator categories. Moderators may effect change through multiple mediators, though only one or two are presented for simplicity. Ripple effects are not described below, but are detailed in Appendix 5.
**Actionability:** The more a feedback message can directly facilitate behaviours in A&F recipients, the more effective it is. With regard to the **Goal** of the A&F intervention, a clinical topic perceived to be within the control of the recipient is more actionable because it is more likely their behaviour will influence its outcome (**Controllability**). Furthermore, a **Feedback message** that suggests performance is suboptimal is more actionable than if performance is high, because there is room for improvement (**Performance level**). Similarly, a **Feedback message** that provide lists of patients who have received suboptimal clinical care (**Patient lists**), represents the clinical performance of an individual clinician rather than their wider team (**Specificity**), and is based on recent rather than historical data (**Timeliness**) is more actionable because it communicates to recipients specific ways in which their current behaviour could improve patient care. In terms of the wider **Organisational context**, health care teams that work together effectively towards a common goal (**Teamwork**), or actively communicate with other health care organisations (**Extra-organisational networks**), make A&F interventions more actionable by providing practical support to recipients on communicating, interpreting and responding to feedback messages.

**Resource match:** If a health care organisation’s resources match those required to implement an A&F intervention in terms of time, human or financial costs, they are more likely to successfully engage and respond. If the **Organisational context** is associated with greater numbers of responsibilities (**Competing priorities**), or fewer staff, equipment, or finances, they understandably have less resource to dedicate to A&F and organisational-level change (**Resource**). Similarly, an **Audit** conducted by the recipients themselves (**Conducted by recipients**), or performed manually (**Manual vs automatic**), requires more resource, whereas **Feedback messages** delivered solely face-to-face require significant time commitments from recipients (**Active delivery**). In contrast, **Health professionals** with a greater knowledge of clinical and quality improvement theory increase **Resource match** because they have the requisite skills to interpret and respond effectively to feedback messages (**Knowledge and skills – clinical and quality improvement**). Certain **Co-interventions** can address situations where this knowledge is lacking, such as the provision of support to help recipients understand reasons for suboptimal performance and develop corrective action plans (**Problem solving and Action planning**), facilitated interactions with peers (**Social support**), or additional staff with necessary skills (**External change agents**).

**Complexity:** The simpler an A&F intervention is to engage with, the less resource it requires. **Feedback messages** can reduce their **Complexity** by: communicating the relative importance of its contents, for example by highlighting areas that require urgent attention (**Prioritisation**); providing information on recipients’ historical performance to help them...
interpret their current performance (Trend); and by ensuring they are user-friendly (Usability). Furthermore, Feedback messages that are ‘pushed’ to recipients, such as formal dissemination programmes in organisations, makes them simpler to receive (Active delivery).

**Relative advantage:** The more an A&F intervention has a perceived advantage over current ways of working, the more likely it is to be adopted with the organisations’ available resources. Consequently, most of what constitutes Relative advantage is context-specific. However, in general Relative advantage can be increased if the Implementation of A&F demonstrates their benefits to recipients, for example by highlighting improvements recipients have made since using A&F (Observability), or when Health professionals have positive beliefs towards the potential benefit of A&F in the first place (A&F attitude).

**Compatibility:** The more an A&F intervention aligns with the beliefs, systems, and processes of an organisation and its staff, the less disruptive and more successful it will be. An A&F intervention that focuses on a Goal perceived to represent comprehensive ‘good’ clinical care (Importance), that is relevant to the recipients’ job (Relevance), will align with their views on what is clinically significant, and their working life. However, during Audit, if a health professional cannot exclude patients they deem inappropriate to be included in their clinical performance measurement, it will clash with their autonomy and motivation to provide high quality patient-centred care (Exclusions). Patient population reasons why this may happen include those who have made explicit choices to refuse care (Choice misalignment), or those who have clinical characteristics that preclude them from receiving care processes (Clinically inappropriate). Similarly, this also happens where recipients believe the purpose of a Feedback message is to punish suboptimal performance rather than support improvements in care quality (Function). During Implementation, Compatibility can be increased if an A&F intervention is tailored to the needs of the health care organisation and its staff (Adaptability), and a sense of ownership fostered rather than imposing the intervention (Ownership). Finally, A&F can harness the strengths of existing systems and procedures within its Organisational context where teams have effective internal communication channels (Intra-organisational networks) and by aligning with their current ways of working (Workflow fit).

**Credibility:** The more trustworthy and reliable an A&F intervention, the more health professionals will believe it will help them improve patient care, and the more likely they will engage with it. Audits perceived by recipients to be accurate measures of clinical performance increase their credibility (Accuracy), whereas Feedback messages that detail the patients used to calculate their clinical performance increase trustworthiness (Patient
lists), especially when the recipient wants to verify the feedback message (Verification). Furthermore, Feedback messages delivered by a person or organisation with an appropriate perceived level of knowledge and skill (either clinical or technical) increases their reliability (Source – knowledge and skill).

Social influence: The more an A&F intervention can harness the social dynamics between health professionals, the more likely they are to be implemented. Feedback messages that compare recipients’ performance to other health professionals may promote competition between recipients [89], and influence them to change their behaviour if others behave differently [84] (Benchmarking). This may be augmented if the identities of the health professionals are made visible, by recipients striving to maintain their status as a member of a group of high performing clinicians [90]. In the wider Organisational context, when senior managers advocate the A&F intervention, this facilitates A&F processes and organisational-level behaviours by instructing staff to engage and respond appropriately [84] (Leadership support).
Figure 3: Clinical Performance Feedback Intervention Theory (high confidence findings only; Appendix 5)

Notes: Boxes – Blue boxes = moderating variables, orange boxes = mediating variables, Green boxes = processes and outcomes. Moderating variables – Green text = facilitators, Red text = inhibitors, Orange text = mixed, Black text = ripple effects only.
Box 1: Case studies

Case study 1: The pharmacist-led information technology-enabled (PINCER) intervention [139]

Setting and topic
Primary care (England); Medication safety

Effectiveness
Effective (determined in randomised controlled trial [160])

Description of A&F intervention
Pharmacists as additional members of staff were allocated to GP practices for a period of three days per week for up to 12 weeks. They conducted an educational session at the beginning of the intervention period with practice staff on the importance of medication safety. Then periodically throughout the intervention period provided practice staff with population summaries, and lists, of patients at-risk of medication safety errors in their practice (e.g. Patients with asthma also prescribed beta-blockers). Pharmacists used educational outreach and root cause analysis techniques to identify potential causes of clinically important errors in medicines management, and to assist practices in making changes to patients’ medication.

Key findings from qualitative paper (references to processes, moderators and mediators from the meta-synthesis in italics)
GP practice staff found the medication safety topic had both Relevance and Importance (Credibility), and therefore engaged with the intervention. The GP practice staff often struggled to find the time or staff (Resource match) to undertake action (Behaviour) based on the feedback. The pharmacists (External change agents) provided additional Resource through which most of the improvement action took place. The pharmacists were also perceived as having sufficient Credibility to influence change. The use of Patient lists was key to identifying patients in whom action was required.

Case study 2: Clinical Resource and Audit Group (CRAG) indicators [161]

Setting and topic
Secondary care (Scotland); Patient outcomes

Effectiveness
Ineffective

Description of A&F intervention
CRAG indicators are compiled and disseminated by the Scottish National Health service. They included 38 clinical indicators detailing patient outcomes for each hospital based on a variety of data sources. To minimise random variation each indicator spans a period of at least three years, and are case-mix adjusted. Typical indicators include: five-year survival in women with breast cancer, and 30-day survival after emergency admission for stroke.

Key findings from qualitative paper (references to processes, moderators and mediators from the meta-synthesis in italics)
The main reasons for lack of impact of CRAG indicators related to:
- Concerns that the data were not reliable (Accuracy; Credibility);
- Significant time-lags between collection and presentation of data, which was often a few years (Timeliness; Actionability);
- Lack of awareness of the indicators amongst health professionals to whom the indicators measured their clinical performance, because they were not provided the feedback (Interaction; Active delivery; Resource match);
- Lack of training, facilitation and knowledge amongst health professionals regarding how they could interpret and act upon the indicators (Knowledge and skills – quality improvement; Training and support; Resource match; Actionability);
- Lack of data regarding individual clinician performance (Specificity; Actionability);
- Focus on outcome rather than process indicators. Process indicators are felt to be more accurate, easier to act upon, and quicker to see an improvement in (Process vs Outcome; Actionability; Credibility).
Box 2: Example quotes from included studies

Controllability
Clinical managers also felt this powerlessness. At one facility, the primary care clinical lead was responsible for meeting “...350 something” (Facility 3) measures though the ability to improve clinical processes was outside their sphere of control: “…when I walked in here, it was basically primary [care]… everybody’d leave the room and ...I’d say, ‘Wait a minute! Everybody owns these measures… a guy goes to Audiology to get his hearing aids, he don’t [sic] ever come to Primary Care, but I’m going to get dinged because he didn’t get his flu shot?’ (Author interpretation and participant quote [123])

Performance level
For the most part, my numbers are usually in the middle, around the average... near the bar... I always feel like if I’m within one standard deviation of the bar, I can live with that. (Participant quote [103])

Patient lists
The informants suggested that the identities of the inappropriately treated patients should be revealed in prescriber feedback.... Prescribing data on individual patients were considered more relevant than aggregated data even if patients’ identities were not revealed. However, having received prescriber feedback with in-formation on inappropriate treatment on non-identifiable patients, one GP said: “It was frustrating that I had a quality problem without being able to do something about it ... (but) ...I am not sure whether I actually have a quality problem” (Author interpretation and participant quote [162])

Specificity
Participants were asked what they thought of the idea of individualized AF [audit and feedback], in particular, receiving data on individual compliance. Here responses were largely positive. They believed the personalized sheets, with each patient identified and the consequent infection rate, were very useful with regard to directly seeing patient outcomes based on their practices. (Author interpretation [163])

Timeliness
The elapsed time between collection and publication of data was a major drawback to the indicators being used in a meaningful way for continuous quality improvement. In many cases the CRAG indicators are at least a year out of date and considerably more for some indicators such as breast cancer. “It’s pretty basic information and it comes out several years after it is taken. Things have changed over that period of time. So, in relation to say treatment of cervical cancer, the whole way of cancer management has changed. The change had already occurred by the time the data were issued” (medical director)” (Author interpretation and participant quote [161])

Teamwork
We also observed that when there were prior habits of collaborating or when nurses and physicians were already sharing responsibilities for the management of diabetic patients, it facilitated the reflective and action-planning processes (Author interpretation [111])

Extra-organisational networks
Moreover, the collaboration of local stakeholders, particularly the RNs registering data in Riks-Stroke, along with the national Riks-Stroke staff, facilitated learning about the registry and about stroke care. “Having access to the experience and knowledge of [an authority on the Riks-Stroke team] is valuable. Really.” (Author interpretation and participant quote [147])

Competing priorities
Two practices struggled to embed the work within practice routines and expressed the concern that DQIP work could be sidelined by competing work pressures. (Author interpretation [164])

Resource
Improvement against a number of audit standards in this study relied on the availability of specific equipment and supplies. These were often in insufficient supply at all study sites. As a TB nurse in Peru put it:“Lack of resources, inadequate infrastructure, staff with too much workload, which prevents quality care. The patient care is by quantity not quality, when it should be the other way around.” (Author interpretation [157])
Conducted by recipients
Another barrier reported in the focus group was that participating in the InFoQI program was very time-consuming, especially the collection of valid and reliable indicator data. (Author interpretation [165])

Manual vs automatic
Most data were available for participating hospitals. However, extracting data from existing sources for some PIs proved burdensome; e.g. collecting staff orientation indicators was challenging: We discovered that the level of computerization of our human resources departments was very bad so some hospital [staff] had to take pencils to calculate them manually. (Author interpretation and participant quote [166])

Active delivery – inhibiting factor
Feedback challenges Scorecard data were not always presented to staff. In both hospitals, meetings were sometimes cancelled due to competing priorities or practical issues. The most pronounced example occurred in a Hospital A intervention ward: …there was a weekly juniors’ meeting specifically for the [Hospital Award] team. It seems to have gone into abeyance […] largely because of lack of space (Author interpretation and participant quote [167])

Active delivery – facilitating factor
However, we found that although consultants and chief executives were aware of the data, most nurse managers and junior doctors reported that they had little or no knowledge of the indicators. Only one trust disseminated these data to nurse managers and junior doctors. “There should be more widespread dissemination of this information [the CRAG reports]. It would certainly be useful to push it down to my level of service manager” (Author interpretation and participant quote [161])

Knowledge and skills – quality improvement
None of the trusts ran specific training or education programmes on the appropriate use and interpretation of clinical indicators…I don’t think there is sufficient knowledge about CRAG data. It is not taught in medical schools” (Author interpretation and participant quote [161])

Knowledge and skills – clinical
For ‘antithrombotics in AF’, clinical reasons reported were mainly ‘unfitness for warfarin’ (examples reported in interview included fragility and dementia, heavy alcohol use and previous gastrointestinal bleeding), but also included ‘paroxysmal AF’ (reflecting a misconception that stroke risk is lower than for patients with chronic AF) (Author interpretation and participant quote [164])

Problem solving
An active and interactive approach was observed in teams A and B, reflected in the planning of regular team meetings for discussions of scores, possible problems and solutions, and appointing a responsible person to take action. This approach was lacking in teams C and D, as confirmed by the surgeon from team D: “We should have looked at the data more often and also discussed the results to discover weaknesses. Rather, we acted reactively.” (Author interpretation and participant quote [102])

Action planning
The interviewer explored whether participants felt it was appropriate for the SPCERH team to make suggestions for changes in their service, or if there were any sense of resentment about these external suggestions. All participants who had received and discussed the letter were comfortable with SPCERH having made suggestions and with the suggestions they made. Appropriate for SPCERH to suggest areas for change “I thought it was quite appropriate. Initially, I thought they had just plucked three items. … So, obviously thought had gone into it and I felt that it was quite handy to have.” (Author interpretation and participant quote [129])

Social support
Participants mentioned that the workshop allowed them to target what they wished to improve in their collaborative management of diabetes in order to de-clutter and reorganize services to improve the healthcare system, deliver more homogenous and standardized care in the region, increase treatment adherence, and develop new tools to improve follow-up and interprofessional collaboration (Author interpretation [111])

External change agents
PINCER pharmacists may be viewed as ‘change agents’… It is the change agency’s aim to implement the innovation with a focus on the collective goals of the social system (here to improve prescribing safety) (Author interpretation [139])
Prioritisation
Surgeons reported that the vast number of metrics was a contributing factor to the complexity of the reports; a few indicated a preference for a more summary-based approach. "I think it has just way too (much) information, especially NSQIP, and it’s kind of hard to hone in on what it’s really saying or what it’s telling you." (Author interpretation and participant quote [168])

Trend
Participants (both nurses and unit managers) reported the import ance of being able to see and trend the data over time and know that the units are keeping on track and whether collective goals are achieved for the selected BPGs (Author interpretation [169])

Usability
Moreover, the EFS did not communicate with the in-house IT system which resulted in time consuming clicking around on the screen. “It is things like that which demands mental surplus, you know . . . To go in and look at it [the sortable lists] collectively like that, I really don’t think we have enough time for it.” (Author interpretation and participant quote [170])

Observability
The appreciation of a visible relationship between audit activity and real improvements was a key motivational factor for clinicians to change practice as seen in both Cuba and Bolivia. As the director of a health centre in Bolivia commented: [...] the audits really helped us to improve our quality of service. Because thanks to these, we’ve been able to correct mistakes, improve the quality of patient care and, above all, treat patients more humanely [...]…A visible relationship between audit activity and patient care can be a key motivational factor for clinicians to change their practice. (Author interpretation and participant quote [157])

A&F attitude
Fewer doctors (n=7) were enthusiastic. They believed that our overall approach could provide data complementary to the individual patient focus of primary care. “General practice is a multi-pronged job as you’re looking after individuals as well as numbers and quantitative indicators. Half our job is to make sure the patient as a person is fine but it’s important to see how you’re doing on standard issues such as hypertension and ischaemic heart disease which can be categorised and quantified, so it’s (the intervention) a useful part of the audit circle” (Author interpretation and participant quote [171])

Importance
All GPs interviewed highly valued the process of reviewing patients identified as receiving high-risk NSAID or antiplatelet prescriptions. The topic is, I would go so far as to say, essential. I don’t even think you can say it’s urgent. It’s essential that practices are doing this. They could be killing patients totally unnecessarily and it’s not as if it’s difficult, because in a lot of circumstances, the vast majority of them are nonsteroidals in elderly people (Author interpretation and participant quote [172])

Relevance
All the GPs and most of the nurses reported having a special interest in, or responsibility for, diabetes care. Those who had a special interest in diabetes had often taken the lead organizing diabetes care for the team. “I am studying diabetes management currently so this fitted in with my interest." (GP) (Author interpretation and participant quote [147])

Exclusions & Choice misalignment
Many CCRs at participating facilities are locally programmed to remain active even if a patient refuses a PM [performance measurement]-related intervention. Although this is not mandated, CCR programmers are often instructed to match the CCR programming as closely as possible to national PM definitions, which do not exclude patient refusals from the score. As a consequence, many nurses (14) say they waste time asking patients the same questions at multiple visits. “There are some [patients] that are saying I’m just flat-out not doing it, don’t ask me anymore. And one of the nurses actually told me that somebody threw FOBT cards at her” (Author interpretation and participant quote [173])

Exclusions & Clinically inappropriate
In other instances there are limitations due to systemic con straints which do not allow for flexibility in prescribing practices in specific clinical contexts where deviation from recommended doses may be clinically indicated. “Military patients have a set pain protocol which involves (…) prescribing a number of opioids. So every time that I put somebody on this pain protocol, I get a red alert saying
‘multiple opioid drugs prescribed, are sure you want to proceed?’, so I tick yes but obviously then on the dashboard I will get a negative mark if you like.” (Author interpretation and participant quote [174])

Function
Data also revealed a general concern with the ‘Big Brother Watching’ mentality that many GPs seemed to fear as part of their future work situation.”(...) These lists compare my prescriptions to those of my colleagues. They leave me completely cold. It’s . . . I think I have done well enough, right? And you have to protect yourself against all the things you can be measured and weighed by” (Male #1D, solo practice). “(Author interpretation and participant quote [170])

Adaptability
In the audit discussions the purpose and process of the groups’ practice development were negotiated within the group and together with the facilitators. A critical attitude to the clinical guidelines was considered necessary in order to adapt the guidelines to local contexts and thereby translate research-based evidence into local clinical practice (Author interpretation [156])

Ownership
Many perceived challenges to the regional audit process were concerned with problems with ownership, both at organising level, and within local departments. It tends to be the same people in terms of the consultants who organise it . . . it could be seen as being exclusive (Participant 2). Being a passive recipient [in regional audit] . . . you don’t feel as much ownership . . . you don’t feel as dedicated to the process.” (Author interpretation and participant quote [175])

Intra-organisational networks
The practice accreditation program requires the involvement of the whole team. Therefore it is advised to organize structural team meetings to evaluate the progress of improvement plans. Participants experienced implementation of the program as more effective when indeed all members of the team were involved and processes were structurally evaluated in team meetings. (Author interpretation [141])

Workflow fit
Our observations of EMR use also revealed that variable user acceptance of eHealth led to constraints on the ability to use EMR data for performance feedback. ART providers appeared to avoid using some EMR functionality when the EMR workflow did not support established clinical processes, often related to optimizing provision of care under a heavy workload. (Author interpretation [97])

Accuracy
Interviewees expressed a range of opinions about the validity of PROMs. The factors identified were related to possible biases, confounding, and chance. Participants were aware that incorrect administration and completion of the measures would affect the data quality. In particular, they were concerned about the potential to manipulate scores by failing to recruit patients who may be more likely to have a poor outcome, thus creating a selection bias. (Author interpretation [176])

Source – knowledge and skill
The audits were universally viewed as important and valuable. According to interviewees, this is due to the credibility of the established professional societies and authoritative bodies associated with the reports, rather than the content of the reports themselves: ‘I don’t think DAHNO has ever told us anything we didn’t already know, it’s just given us data that’s authoritative […]. It’s providing data that makes people realise that we’re telling the truth’. (Author interpretation and participant quote [101])

Benchmarking – non-identifiable
Several (11) mentioned that because providers can see and track their own individual performance a spirit of competition can occur that is motivating: I like to see where I fall in that list and I definitely don’t like to see myself in the middle or lower down, so I try to keep up and be in the top few percent. . . Yeah, the competitive nature, for me that’s how I motivate myself. (Author interpretation and participant quote [177])

Benchmarking – identifiable
It [the CRAG data] is used to help strengthen the case for change. In terms of sanctions it is a peer one—not letting your peers down (Participant quote [161])

Leadership support
A lack of support from upper level management was cited as a barrier in 9% of services, two of these were Older Person’s services that were co-located in nursing homes. The problems they faced are illustrated in the following quotation: But even when you go one step with higher management... because this unit is one of ‘them’ units [you’re] left to your own devices for decision making in a lot of things because they haven’t got a clue ... it’s just not their main patient group. (Author interpretation and participant quote [178])

**Gaming – data manipulation**

A third unwanted consequence was that technovigilance produced incentives for staff to “game” the system… A small number of staff also suggested that sometimes nurses clicked that a dose had been given when in fact it had not, though we did not witness any instances of this occurring." (Author interpretation and participant quote [179])

**Gaming – patient population manipulation**

Most surgeons thought the impact of taking care of high-risk patients would worsen their SSR metrics. One surgeon referenced the cardiac surgery experience with individual outcomes. He/she described how physicians would “cherry-pick” their patients, likely in an attempt to keep “excellent outcomes.” (Author interpretation [168])

**Tunnel vision – during point-of-care**

When time is limited, patient health concerns may be given lower priority than PM-related areas of care (23 mentioned). Occasionally a patient will come in with a complaint and the providers will make sure all the alerts are answered rather than addressing the complaint per se. Because you have this and this to do and you don’t address the fact that they have low back pain because that’s not a performance measure or their ankle hurts or something. (Author interpretation and participant quote [152])

**Tunnel vision – during quality improvement activities**

Many participants argued that much quality assurance work is being done within the field of diabetes care. As a counterweight, many felt that conditions like hypertension and chronic obstructive pulmonary disease (COPD) were in more need of attention. (Author interpretation [170])
4.5 Discussion

Clinical Performance Feedback Intervention Theory

We synthesised 65 qualitative evaluations of 73 different A&F interventions, using 18 different theoretical models to identify causal pathways of effectiveness in A&F interventions. We found that A&F is a cyclical process of Goal setting, Audit, Feedback, recipient Interaction, Perception, and Acceptance of the feedback, followed by Intention, Behaviour and Clinical performance improvement. Progress round this cycle is influenced by 68 moderating variables relating to characteristics of the Goal, Audit methods, Feedback message, Implementation process, Organisational context, Co-interventions, Health professional, and Patient population, that exert their effects via mediators relating to Actionability, Resource match, Complexity, Relative advantage, Compatibility, Credibility, and Social influence. These moderators generally have direct effects on A&F processes, though can also exert their influence indirectly via other moderators in ‘ripple effects’ [159]. Although the main aim of A&F is to improve clinical performance, unintended outcomes of Gaming and Tunnel vision may also occur.

Based on these findings, we propose a model of effective A&F called Clinical Performance Feedback Intervention Theory (CP-FIT; Figure 3). CP-FIT has three propositions that govern causal pathways in A&F: Behavioural induction; Capacity limitations; and Identity and culture. In general, Proposition 1 (Behavioural induction) is supported by the Actionability mediating variable; Proposition 2 (Capacity limitations) is supported by Resource match, Complexity, and Relative advantage; and Proposition 3 (Identity and culture) is supported by Compatibility, Credibility, and Social influence. Furthermore, CP-FIT suggests that A&F exerts its effects on patient care via two different mechanisms: Direct Action; and Knowledge/Awareness. Both mechanisms relate to Proposition 1 (Behavioural induction). Below, we discuss the propositions and mechanisms in detail.

Proposition 1 – Behavioural induction: A&F interventions exert their effects by inducing behaviours in health professionals related to individual patient care.

As demonstrated in Figure 3, A&F can only produce improvements in clinical performance if it results in health professionals undertaking Behaviour. As clinical performance in A&F is measured in care provided to individual patients in a specified population, this behaviour must relate to individual patient care. However, behaviours may also systematically change the way care is delivered by a health care organisation, such as starting, stopping, or modifying: services and protocols (e.g. [180]), or staff roles and training (e.g. [170]). These ‘organisational-level’ behaviours further boost clinical performance because they directly facilitate multiple ‘patient-level’ behaviours by
augmenting the clinical environment in which they take place, therefore benefitting populations of patients (e.g. [179]). Conversely, the absence of ‘organisational-level’ behaviours can lead to limited improvements in clinical performance (e.g. [104]). Furthermore, Figure 3 also demonstrates a health professional must undertake a number of additional processes prior to Behaviour i.e. Interaction, Perception, (sometimes Verification,) Acceptance, and Intention. Consequently, A&F that induces Behaviour that most effectively influences patient care, or facilitates these necessary precursor processes, are most successful.

Proposition 2 – Capacity limitations: Health care organisations have limited capacity to engage with and respond to the demands of A&F interventions.

Health care organisations in general have limited resources in terms of staff, time, finances and equipment (e.g. [104]). This is exacerbated by multiple demands in addition to A&F has to contend with, such as their primary responsibility to provide patient care, or wider organisational priorities such as existing quality improvement initiatives and re-structures (e.g. [142]). Concurrently, each A&F process (Figure 3) necessitates a non-trivial resource commitment: interacting with a feedback message requires time (e.g. [149]), and executing an action plan often requires material resource (e.g. [103]). Furthermore, health professionals often do not possess the necessary resource in terms of knowledge and skills to undertake effective A&F, such as the ability to understand or respond to a feedback messages (e.g. [104]), or occasionally the related clinical knowledge (e.g. [146]).

Proposition 3 – Identity and culture: Health care professionals and organisations have strong sets of beliefs and behaviours regarding how they provide patient care that influence their interactions with A&F.

Health care professionals and organisations believe they should (and do) provide high quality patient care (e.g. [103]). An important aspect of which involves the autonomy to acknowledge when the scientific evidence does not apply to particular patients based on their specific clinical characteristics (e.g. [123]). Furthermore, health professionals and organisations may have particular goals, such as interest in a certain clinical area (e.g. [181]), and a desire to perform better than their colleagues and neighbouring organisations (e.g. [151]). Accordingly, they have a set of systems and processes to help them achieve those goals, such as methods of care delivery and patient data collection (e.g. [182]). In general, these goals and systems are difficult to change (e.g. [183]), though if recipients receive a feedback message of suboptimal clinical performance, they will generally take corrective action as this aligns with their motivations [7–9,83,85].

Mechanisms: Direct Action and Knowledge/Awareness
A&F interventions facilitate health professionals to improve patient care by enabling them to take Direct Action, or by increasing their general Knowledge/Awareness. Both mechanisms can facilitate behaviour change at both the patient and organisational levels. The mechanisms are not mutually exclusive, and the same intervention will often operate through both mechanisms.

Direct Action means this behaviour is directed by the A&F intervention itself, for example by providing lists of patients that require action, or by highlighting specific organisational issues. Direct Action may be retrospective (e.g. directed at patients who have received suboptimal care in the past), or prospective (e.g. focusing on patients who need an action in future clinical encounters). Knowledge/awareness works by highlighting a particular suboptimal care quality issue to a health professional. This may be related to the knowledge itself (e.g. a clinical guideline recommendation) or their own performance on this particular quality issue. The health professional may already have some degree of knowledge/awareness about the issue, in which case A&F serves to remind them; or they may not, in which case A&F provides them with new learning. Knowledge/Awareness mainly works prospectively by enabling the health professional to take action regarding patients they will encounter in future.

To illustrate, the PINCER intervention [139] (Box 1) used both the Direct Action and Knowledge/Awareness mechanisms: the pharmacists provided lists of patients requiring action (Direct Action), and education on the clinical topic of medication safety (Knowledge/Awareness), to the health professionals. However, it was the patient lists, and the additional resource provided by the pharmacists to undertake action that was the main reason for its success. In contrast, the CRAG indicators [161] relied on the Knowledge/Awareness mechanism, raising awareness of patient outcomes at the hospitals. However, it did not enable direct action, for example by providing clinician-specific data, or lists of patients [161]. To be most effective, CP-FIT suggests that A&F interventions should maximise the way in which they utilise the Direct Action mechanism.

Implications for practice and research
We believe CP-FIT may be useful for guiding the: 1) design and implementation of A&F interventions; 2) evaluation and explanation of observed or predicted A&F intervention outcomes. The vast majority of papers in this review did not explicitly use theory to either design A&F or interpret their results (n=49, 75%), which suggests a need for CP-FIT.

With regard to A&F design and implementation, Box 2 summarises potentially effective features that practitioners may wish to consider. These are only high confidence findings, and other design features are provided in Appendix 5 with lower confidence. However,
these relatively simple statements should be interpreted with caution: Figure 3 demonstrates the number of opportunities (processes) in which A&F can fail, and the number of reasons and their interactions that may influence this failure. This results in tensions between mediating variables within CP-FIT where simply overcoming a ‘barrier’ or implementing a ‘facilitator’ to A&F does not necessarily lead to success [184] (see Appendix 5 for more detail). For example, a central tension exists because effective A&F requires health professional engagement (Proposition 1), which in turn requires resources that health care organisations often do not have (Proposition 2). This is illustrated by the opposing effects of the moderating variable Active delivery (Box 1): a feedback message ‘pushed’ to recipients face-to-face to ensure they receive it may also act as a barrier because they do not have the time to attend the meeting.

In terms of evaluating and explaining A&F outcomes, CP-FIT can guide both data collection and analysis. For qualitative studies, Appendix 5 provides a comprehensive codebook that can be used to code studies, and explain causal patterns. Similarly, in quantitative studies, Appendix 5 provides over 200 falsifiable hypotheses of varying levels of confidence to test. Both qualitative and quantitative data may be derived from empirical studies, literature syntheses, or when assessing an intervention prior to its implementation. In process evaluations [23], CP-FIT may be particularly useful for identifying weak points in a causal chain of events as to why an A&F intervention was unsuccessful, and providing potential explanations [185].

Although developed specifically for A&F, CP-FIT may also be relevant to other quality improvement strategies. Firstly, most quality improvement interventions follow plan-do-study-act cycles based on data collection and quality measurement, similar to A&F [186]. Secondly, CP-FIT’s propositions are wide-ranging and likely applicable to most health care contexts and health professional behaviour change interventions, with parallels to the Capability-Opportunity-Motivation-Behaviour (COM-B) system [121]. Specific examples include interventions that facilitate interpretation of patient data against pre-defined quality standards of care, such as computerised clinical decision support or educational outreach [187].
Box 2: Effective A&F practices according to CP-FIT

High confidence findings only unless stated otherwise. CP-FIT constructs in *italics*.

**Guiding principles**

To be most effective A&F interventions should…

- Induce behaviours in recipients (*Actionability*).
- Provide additional resources to engage with it, or minimise the amount of resource required to do so (*Resource match*).
- Align with characteristics of the recipient (e.g. their beliefs, norms, values, and ways of working), and their organisation (e.g. culture, structures, processes, and technical systems; *Compatibility*).
- Be trustworthy and reliable (*Credibility*).
- Be perceived as better than alternative ways of working (*Relative advantage*).
- Be simple to engage with (*Complexity*).
- Harness the social dynamics between health professionals (*Social influence*).

**Specific examples**

Setting goals that are…

- Within the recipients’ control (*Controllability*).
- Considered important by the recipient (*Importance*).
- Relevant to the recipients’ job (*Relevance*).
- Focuses on areas of suboptimal clinical performance (*Performance level*).

Conducting the audit…

- Without requiring the recipient to collect or analyse the data (*Conducted by recipients*).
- In an automated way (*Manual vs automatic*).
- Ensuring accurate representation of the recipients’ clinical performance by using appropriate source data and analysis methods (including sample size where appropriate; *Accuracy*).
- Allowing the recipients’ to exclude patients they feel are inappropriate to be included in the measurement of their clinical performance (*Exclusions; Choice misalignment; Clinically inappropriate*).

Producing a feedback message that:

- Includes lists of patients used to calculate the recipients’ clinical performance (*Patient lists*).
- Provides the recipients’ individual clinical performance (as opposed to just their team or organisation; *Specificity*).
- Is sent as close to the time of the clinical performance measured in the audit as possible (*Timeliness*).
- Effectively summarises and communicates the relative importance of its contents (*Prioritisation*).
- Has been tested to ensure it is user-friendly (*Usability*).
- Provides historic in addition to current clinical performance (*Trend*).
- Are ‘pushed’ to recipients rather than requiring them to request access, but do not rely solely on face-to-face delivery (*Active delivery*).
- Convinces the recipient that its purpose is to support them improve care rather than punish them (*Function*).
- Appears to come from a source with an appropriate degree of technical or clinical knowledge (*Source – knowledge and skill*).
- Compares recipients’ performance to other health professionals (*Benchmarking*).
- Is delivered to groups of health professionals in a team or organisation rather than just one (*Delivery to a group*).

Implementing A&F in a way that…

- Gains the support of senior managers within the organisation (*Leadership support*).
- Fits with the existing workflows of the organisation (*Workflow fit*).
- Minimises costs in terms of time, human or financial resource (*Cost*).
- Demonstrates its value or benefits for the recipient and organisation (*Observability*).
- Can be tailored to the needs of the health care organisation and its staff (*Adaptability*).
- Targets health professionals with quality improvement skills (*Knowledge and skills – quality improvement*), or provides support and training to recipients regarding how to engage with the intervention (*Training and support*).
• Targets health professionals with adequate knowledge regarding the clinical topic of the A&F, or improves the recipients’ knowledge of the evidence and theory of the clinical topics focused on by the A&F intervention (Knowledge and skills – clinical) – though it is unclear how best to do this at present.
• Makes recipients feel like it is not imposed upon them (Ownership).

Providing additional support to...
• Help recipients interpret and formulate action plans in response to suboptimal clinical performance – especially those that address the organisation (Problem solving; Action planning; Organisational-level behaviour).
• Facilitate recipients to discuss their clinical performance with peers in their own or different organisations, either informally or formally (Social support).
• Resource the intervention e.g. protected time, additional staff or equipment (Resource; External change agent).
• Improve the ability of recipients to communicate with and work towards a common goal with their colleagues (Intra-organisational networks; Teamwork) – this may be helped by Social support and Delivery to a group.
• Help recipients’ organisations communicate with external organisations (Extra-organisational networks) – this may be helped by Social support and Delivery to a group (moderate confidence).
• Address negative attitudes towards A&F (A&F attitude) – this may be helped by Social support and Observability (moderate confidence).
Comparison to existing literature

The latest Cochrane review on A&F found wide variation in its effectiveness [4]. Meta-regression analysis determined some of this variation was caused by: recipients’ baseline performance, who delivered the feedback, how frequently feedback is provided, the feedback format, whether instructions for improvement are provided, and the type of clinical behaviour targeted [4]. Table 5 shows how CP-FIT may explain these findings. In addition, CP-FIT also explains why some hypothesised design features may have lacked outcome effects: for example, the review found conflicting evidence for Benchmarking [4], which CP-FIT predicts may have both positive effects (by harnessing Social influence) and negative effects (by reducing Credibility; moderate confidence – Appendix 5). CP-FIT can also suggest further sources of variation not identified by the Cochrane review’s meta-regression, though their operationalisation may be limited. For example, many of CP-FIT’s variables rely on the context into which an A&F intervention is introduced, which would be difficult to extract and quantify from an RCT’s report. Nevertheless, a number of hypotheses may be easily operationalized (Box 2) and could be tested in future updates of the Cochrane review, for example: if A&F interventions allow Exclusions, provide Patient lists, or recipient’s performance Trend.

CP-FIT aligns well with the tentative best practices posited for effective A&F [2,6] (Table 6). It adds to these recommendations by providing additional evidence-based explanations as to why they work, in addition to suggesting further potential best practices (Box 2) such as delivering feedback to a group, and gaining leadership support. The main advantage of CP-FIT in comparison to these best practice recommendations is it provides a set of generalisable rules and explanations that can be used to design potential A&F interventions. Consequently, the potential design recommendations from CP-FIT can be extended beyond those presented in Box 2 if they conform to its propositions and mediating variables. When used in combination with the causal process model (Figure 3), CP-FIT may be used to evaluate and understand ‘barriers to feedback use’ – one of the main recommendations in [6].
<table>
<thead>
<tr>
<th>Cochrane review finding</th>
<th>Explanation according to CP-FIT*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low baseline performance increases A&amp;F effectiveness</td>
<td>Core to CP-FIT’s <strong>first proposition</strong>. Low <em>Performance</em> level facilitates <em>Intention</em> and <em>Behaviour</em> in recipients because it increases <em>Compatibility</em> with their personal views (i.e. that they want to provide high quality patient care) and <em>Actionability</em> (i.e. low performance implies room for improvement).</td>
</tr>
<tr>
<td>Feedback provided by a supervisor or colleague is more effective than a professionals’ standards review organisation or employer</td>
<td>A supervisor or colleague is likely to be perceived to have greater knowledge and skill (<em>Source – knowledge and skill</em>), which facilitates <em>Acceptance</em> and <em>Intention</em> by increasing <em>Credibility</em> and <em>Relative advantage</em> (as the opportunity to receive feedback from a credible source is valued by health professionals, and aligns with their sense of autonomy). An external feedback source (<em>Source – location</em>) inhibits <em>Acceptance</em> by reducing <em>Compatibility</em> with recipients’ sense of autonomy).</td>
</tr>
<tr>
<td>Feedback provided monthly is more effective than feedback provided weekly, or less than monthly</td>
<td>Feedback that is not frequent enough (<em>Frequency</em>) inhibits <em>Interaction</em> and <em>Intention</em> by increasing <em>Complexity</em> (making the feedback message more difficult to receive). Feedback that is too frequent inhibits <em>Perception</em> by increasing <em>Complexity</em> (too much feedback makes it more difficult to understand) and decreasing <em>Resource match</em> (feedback provided too often gives less time to act on it).</td>
</tr>
<tr>
<td>Feedback provided both verbally and written is more effective than providing either alone</td>
<td>Feedback that is actively ‘pushed’ to recipients (<em>Active delivery</em>) facilitates <em>Interaction</em> by reducing <em>Complexity</em> (making the feedback message simpler to receive). Soley providing feedback face-to-face inhibits <em>Interaction</em> by decreasing <em>Resource match</em>. Inhibits <em>Interaction</em> if solely requires formal face-to-face feedback sessions by decreasing <em>Resource match</em> (as they require significant time commitment from recipients).</td>
</tr>
<tr>
<td>Providing explicit targets and an action plan together are more effective then either alone, or neither</td>
<td><em>Action planning</em> and <em>Problem solving</em> facilitate <em>Intention</em> and <em>Behaviour</em> by increasing <em>Actionability</em> (providing practical support on how to respond effectively to the feedback message) and <em>Resource match</em> (by addressing health professionals’ general lack of knowledge and skills to perform these behaviours). <em>Targets</em> facilitate <em>Perception</em> and <em>Intention</em> by decreasing <em>Complexity</em> (making it easier for recipients to know what constitutes ‘good performance’ and therefore what requires a corrective response).</td>
</tr>
<tr>
<td>Decreasing health professional behaviour is more amenable to improvement than increasing or changing behaviour</td>
<td>CP-FIT has no prediction for the <em>Type</em> of behavioural response required. However, this finding may be explained by the <em>Complexity</em> and <em>Resource match</em> mediating variables, as decreasing behaviour may be more simple and require less resource than increasing or changing.</td>
</tr>
<tr>
<td>Prescribing is more amenable to improvement than test ordering or cardiovascular disease management</td>
<td>CP-FIT has no prediction for specific clinical topics. However, this finding may be explained by the variety of variables related to the A&amp;F goal: <em>Evidence base, Process vs outcome, Controllability, Relevance, Importance</em>.</td>
</tr>
</tbody>
</table>

*CP-FIT concepts are in *italics* and detailed in Appendix 5.*
Table 6: Tentative best practices in literature for A&F compared to CP-FIT

<table>
<thead>
<tr>
<th>Brehaut et al. [6]</th>
<th>Ivers et al. [2]</th>
<th>CP-FIT constructs*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address credibility of the information.</td>
<td>Data are valid</td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td>Delivery comes from a trusted source</td>
<td>Source – knowledge and skill</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Source – Location</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Function</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evidence base</td>
</tr>
<tr>
<td>Provide feedback as soon as possible and at a frequency informed by the number of new patient cases</td>
<td>Data are based on recent performance</td>
<td>Timeliness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Frequency</td>
</tr>
<tr>
<td>Provide individual rather than general data.</td>
<td>Data are about the individual/team's own behavior(s)</td>
<td>Specificity</td>
</tr>
<tr>
<td>Provide multiple instances of feedback.</td>
<td>Audit cycles are repeated, with new data presented over time</td>
<td>Frequency</td>
</tr>
<tr>
<td>Provide feedback in more than 1 way.</td>
<td>Presentation is multi-modal including either text and talking or text and graphical materials</td>
<td>Active delivery</td>
</tr>
<tr>
<td>Choose comparators that reinforce desired behavior change</td>
<td>The target performance is provided</td>
<td>Target Benchmark</td>
</tr>
<tr>
<td></td>
<td>Feedback includes comparison data with relevant others</td>
<td></td>
</tr>
<tr>
<td>Recommend actions that can improve and are under the recipient's control.</td>
<td>Targeted behavior is likely to be amenable to feedback</td>
<td>Controllability Performance level Process vs outcome</td>
</tr>
<tr>
<td></td>
<td>Recipients are capable and responsible for improvement</td>
<td></td>
</tr>
<tr>
<td>Recommend actions that are consistent with established goals and priorities</td>
<td>Goals set for the target behaviour are aligned with personal and organizational priorities</td>
<td>Importance Relevance Workflow alignment</td>
</tr>
<tr>
<td>Recommend specific actions</td>
<td>Goals for target behaviour are specific, measurable, achievable, relevant, time-bound</td>
<td>Action planning Problem solving Social support</td>
</tr>
<tr>
<td></td>
<td>A clear action plan is provided when discrepancies are evident</td>
<td></td>
</tr>
<tr>
<td>Closely link the visual display and summary message</td>
<td>N/A</td>
<td>Usability</td>
</tr>
<tr>
<td>Minimize extraneous cognitive load for feedback recipients.</td>
<td>N/A</td>
<td>Prioritisation Usability Number of metrics Graphical elements</td>
</tr>
<tr>
<td>Provide short, actionable messages followed by optional detail.</td>
<td>N/A</td>
<td>Patient lists Prioritisation</td>
</tr>
<tr>
<td>Address barriers to feedback use.</td>
<td>N/A</td>
<td>All of CP-FIT can be used to assess barriers</td>
</tr>
<tr>
<td>Prevent defensive reactions to feedback.</td>
<td>N/A</td>
<td>Function</td>
</tr>
<tr>
<td>Construct feedback through social interaction.</td>
<td>N/A</td>
<td>Social support</td>
</tr>
</tbody>
</table>

*Detailed in Appendix 5
CP-FIT was developed from a number of existing theoretical models and frameworks, including those specific to feedback [7,8,67,81,82] and intervention implementation in general [86,87,93]. In comparison to the feedback-specific models, CP-FIT provides a wider view of factors that may influence success, particularly in relation to how the audit is conducted, organisational context and how feedback interventions are implemented. It also specifies potential unintended consequences, which are largely missing from most feedback-specific models [173]. Furthermore, CP-FIT differs in its predictions to some of its constituent theories: for example, Feedback Intervention Theory predicts that the presentation of others’ performance (Benchmarking in CP-FIT, Normative information in FIT) decreases effectiveness because it directs attention to meta-task processes [8], whereas described above CP-FIT states the relationship is more complex (Appendix 5). In comparison to the general behaviour change and intervention implementation models [86,87,93], CP-FIT understandably provides more detail and specificity to A&F. This demonstrates the most salient constructs to A&F from these general models: for example, many of CP-FIT’s mediating variables originate from Diffusion of Innovations [87] and the Consolidated Framework for Implementation Research [86] (i.e. Compatibility, Complexity, Relative advantage). In combining both feedback-specific and general implementation models to create CP-FIT, it is comprehensive enough to be both relevant to individual and organisational behaviour change, which is not possible with any of its constituent theories alone [36].

Although a systematic review of qualitative evaluations of A&F interventions to our knowledge has not been previously undertaken, two reviews of the use of patient-reported outcome and experience measures (PROMs and PREMs respectively) for health care improvement have been recently published, both of which included qualitative studies [188,189]. Although neither attempted to develop a conceptual model, their major findings can be mapped to concepts from CP-FIT. Boyce et al. [188] found there were practical difficulties in collecting and managing PROMs data related to the an organisation’s resources (cf. CP-FIT Proposition 2) and compatibility with existing workflows (cf. CP-FIT Proposition 3). Whereas Greenhalgh et al. [189] note ‘actionability’ as a key characteristic in the effective use of PROM data (cf. CP-FIT Proposition 1). Both noted that the ‘credibility’ of the data and source from which it was fed back were essential to securing health professional’s acceptance (cf. CP-FIT’s mediating variable Credibility).

**Strengths and weaknesses**

The main strength of this literature review has been the development of a detailed theoretical model to explain A&F effectiveness [28], drawing on the strengths of existing theories to create a new and potentially valuable framework. Using Realistic Evaluation [73] as a guiding framework in our synthesis enabled us to develop causal pathways that
describe the context, mechanisms and outcomes of A&F, which are reflected as CP-FIT’s moderating variables, mediating variables, and A&F processes and unintended outcomes respectively. As far as we are aware, Realistic Evaluation has not been previously been used to guide a meta-synthesis of qualitative studies, and may therefore present a new useful approach. Further methodological strengths include using CERQual to grade the confidence of our findings, representing one of its first uses in a theory-building metasynthesis [70]; and a systematic approach to inductively selecting appropriate theories. This latter method differs from guidance that advocates deductively applying frameworks that appear on the surface to fit the data ‘best’ [190,191]. We believe our approach may provide an alternative method that maximises both internal and external validity of findings.

The main limitation of this metasynthesis (common to any literature review) is that it reflects only what has been studied and reported by the included studies. Therefore, there may be important aspects of A&F not addressed by CP-FIT because they have been missed or not studied in enough detail: for example, there may be particular features of A&F interventions associated with effectiveness, or address the inherent tensions in the model, that have not yet been studied. This may have been exacerbated by our inclusion criteria (Table 1): for example, excluding interventions not yet implemented into clinical practice may have provided detail on the nuanced effects of feedback message design, though conversely would have provided less insight into how this translated to the real-world. Future research should therefore aim to address these potential gaps by evaluating innovative new designs of A&F that test CP-FIT’s core propositions. Further limitations relate to inherent problems in synthesising qualitative evaluations of interventions. Firstly, when coding causal effects in qualitative data, it may not be assumed (either by authors or participants) that a particular feature X resulted in a particular outcome Y unless stated [192]. This is contrast to quantitative meta-analyses and in particular meta-regressions where intervention features can be coded and causation can be inferred at an aggregated level [193]. Consequently, it is possible that particular features of A&F interventions studied in this review had important effects on their effectiveness, but unless the study participants, or authors explicitly mentioned a causal effect, it may not be included. Secondly, the relative effect sizes of the different moderators in CP-FIT and the complexity of their relationships cannot be easily modeled. For example, it is possible that moderator variables that appear to exert influence with high confidence may in fact have negligible effects on outcomes, or their effects are significantly influenced by or dependent on the presence of other variables. Consequently, future work should aim to quantitatively test CP-FIT hypotheses.

4.6 Conclusion
CP-FIT is a new theoretical model of causal pathways in A&F interventions that can be used to design, implement, and evaluate interventions. It was developed from a systematic search and meta-synthesis of qualitative studies built on existing relevant models and frameworks. CP-FIT sheds light on how A&F interventions may be: designed and implemented to result in health care improvement (Objective 1 and 4); responded to, and interacted with, by health professionals (Objective 2); and responsible for unintended consequences (Objective 3). In developing CP-FIT, this paper has addressed gaps in the A&F literature regarding how to optimise A&F interventions [2], and provides tentative explanations for the variation of effects in the latest Cochrane review [4]. Future research should prospectively test and refine hypotheses in CP-FIT both qualitatively and quantitatively, in addition to addressing areas in which there are tensions or evidence-gaps.

4.7 References


[19] C. Glenton, C.J. Colvin, B. Carlsen, A. Swartz, S. Lewin, J. Noyes, et al., Barriers and facilitators to the implementation of lay health worker programmes to improve access to maternal and child health: qualitative evidence synthesis, Cochrane


doi:10.1017/CBO9781107415324.004.


[104] N. Ivers, J. Barnsley, R. Upshur, K. Tu, B. Shah, J. Grimshaw, et al., My approach to this job is... one person at a time, Can Fam Physician. 60 (2014) 258–266.


[164] A.M. Grant, B. Guthrie, T. Dreischulte, Developing a complex intervention to


[187] B. Brown, N. Peek, I. Buchan, The Case for Conceptual and Computable Cross-


Chapter 4 concluding note

This chapter addressed RO1 by developing a new theory to model causal pathways in A&F effectiveness – CP-FIT. Its findings were used to inform the design of the e-A&F system I developed – the Performance Improvement plaN GeneratoR (PINGR). The finalised model was also used to guide PINGR’s implementation into routine practice and evaluation in Chapter 7. The next chapter describes PINGR’s first usability evaluation with software experts.
Chapter 5

Interface Design Recommendations for Electronic Audit and Feedback: Hybrid Usability Evidence from a Research-led System

Benjamin Brown,1,2 Panos Balatsoukas,1 Richard Williams,1,2 Matthew Sperrin,1 Iain Buchan1,2

1 Health eResearch Centre, Farr Institute for Health Informatics Research, University of Manchester, UK
2 NIHR Patient Safety Translational Research Centre Greater Manchester, University of Manchester, Manchester, UK

Contributions
Author BB intellectually conceived the idea for the study, designed and built the software under evaluation, designed the study, collected and analysed the data, and wrote the manuscript. RW assisted with software development. PB and MS helped with study design and data analysis. All authors reviewed versions of the manuscript, and provided critical feedback on its development.

Publication status

Chapter introductory note
This chapter addresses Research Objective (RO) 2 by presenting the first version of the electronic Audit and Feedback (e-A&F) system I developed based on findings from Chapter 3, and emerging findings from Chapter 4 – the Performance Improvement plaN GeneratoR (PINGR). It addresses RO3 by undertaking a preliminary evaluation of
PINGR’s usability with software experts, and RO1 by using them to derive guidelines for the design of e-A&F systems in general.
5.1 Abstract

**Background:** Audit and Feedback (A&F) is a widely used quality improvement technique that measures clinicians’ clinical performance and reports it back to them. Electronic A&F (e-A&F) system interfaces may consist of four key components: 1) Summaries of clinical performance; 2) Patient lists; 3) Detailed patient-level data; 4) Suggested actions. There is a lack of evidence regarding how to best design e-A&F interfaces; establishing such evidence is key to maximising usability, and in turn improving patient safety.

**Aim:** To evaluate the usability of a novel theoretically-informed and research-led e-A&F system for primary care (the Performance Improvement plaN GeneratoR: PINGR).

**Objectives:** 1) Describe PINGR’s design, rationale and theoretical basis; 2) Identify usability issues with the PINGR; 3) Understand how the issues may interfere with the cognitive goals of end-users; 4) Translate the issues into recommendations for the user-centred design of e-A&F systems.

**Methods:** Eight experienced health system evaluators performed a usability inspection using an innovative hybrid approach consisting of five stages: 1) Development of representative user tasks, Goals, and Actions; 2) Combining Heuristic Evaluation and Cognitive Walkthrough methods into a single protocol to identify usability issues; 3) Consolidation of issues; 4) Severity rating of consolidated issues; 5) Analysis of issues according to usability heuristics, interface components, and Goal-Action structure.

**Results:** A final list of 47 issues were categorised into 8 heuristic themes. The most error-prone heuristics were ‘Consistency and standards’ (13 usability issues; 28% of the total) and ‘Match between system and real world’ (n=10, 21%). The suggested actions component of the PINGR interface had the most usability issues (n=21, 45%), followed by detailed patient-level data (n=5, 11%), patient lists (n=4, 9%), and summaries of clinical performance (n=4, 9%). The most error-prone Actions across all user Goals were: 1) Patient selection from a list; 2) Data identification from a figure (both population-level and patient-level); 3) Disagreement with a system recommendation.

**Conclusions:** By contextualising our findings within the wider literature on health information system usability, we provide recommendations for the design of e-A&F system interfaces relating to their four key components, in addition to their integration with each other in a system.
5.2 Introduction

Audit and feedback (A&F) is an established and widely used technique in quality improvement, employed in health care systems across the world. It consists of measuring a clinician or health care team’s clinical performance over a specified period of time (audit), and reporting it to them (feedback), with the intention of raising awareness and helping them take corrective action [1]. Audit data are obtained from medical records, computerised databases, or observations from patients, and feedback may include recommendations for improvement action [2].

In A&F, clinical performance is measured by adherence to recommended clinical practices (e.g. patients with hypertension receiving regular blood pressure measurements) or the occurrence of particular patient outcomes (e.g. acceptable blood pressure control) [1,2]. A&F relates to care provided to multiple rather than individual patients, and is used to inform improvements at an individual, team, and service level [3,4]. Feedback relating primarily to individual patients, particularly intended for use at the point of care, does not count as A&F, and is classified as a different intervention such as clinical decision support (CDS) [1,2].

A&F is traditionally undertaken using paper medical records, which is laborious and time-intensive. However, widespread use of electronic health records (EHRs) has spawned a variety of electronic A&F systems (e-A&F). These systems usually feed audit results back to provider employees via interactive interfaces such as intranet browser-based portals (e.g. [5]) or desktop applications (e.g. [6]). Users of e-A&F systems are generally clinicians whose performance is being assessed, though may also include managers or administrators [7]. e-A&F systems are distinct from systems where an audit is generated using a computerised infrastructure but feedback is provided on paper, verbally or via a static computerised form such as a screensaver or electronic document (e.g. [8]). Often e-A&F systems are not explicitly termed ‘audit and feedback’, and instead may be called ‘dashboards’, ‘scorecards’, ‘business intelligence’, ‘visualisation tools’ or ‘benchmarking tools’ amongst other names [9]. Conversely, many systems with these names may also not be A&F: for example, many dashboards only provide information regarding individual patients (e.g. clinical dashboards [10]) or may focus on multiple patients but are intended for use solely at the point of care (e.g. [11]); and business intelligence or information visualisation tools may focus primarily on non-clinical performance data such as costs, patient waiting times, or disease epidemiology surveillance (e.g. [12]).

Despite their prevalence, there has been relatively little research into the requirements for designing usable interfaces for e-A&F systems. Prior work has largely focused on the effectiveness of e-A&F systems for improving patient care (e.g. [13]) or their levels of
adoption (e.g. [14]). Some studies have explored factors related to their acceptance and use (e.g. [15]), however, we are aware of only one study that has explicitly focused on usability [16]. Consequently little is known about how best to design e-A&F interfaces.

Ongoing work by our group has identified four key components of e-A&F system interfaces [17]: 1) Summaries of clinical performance; 2) Patient lists; 3) Patient-level data; and 4) Recommended actions. All e-A&F interventions have some combination of these elements; indeed, to qualify as A&F the system must have at least a summary of clinical performance or provide patient lists [1–4]. However, we are unaware of a system reported in the literature that incorporates all four components. Below, we discuss each interface component, and what is currently known about their usability.

1. **Summaries of clinical performance**: A&F interventions generally summarise clinical performance using quantitative measures variably termed ‘quality indicators’, ‘performance measures’ or similar. They usually report the proportions or absolute numbers of patients who have (or have not) received a recommended clinical practice, or experienced a particular outcome [18]. These metrics are the core component of A&F, and are commonly presented either as tables (e.g. [19]), bar plots (e.g. [20]), pie charts (e.g. [21]), or line graphs (e.g. [15]). Sometimes colour coding (e.g. [22]) or comparison with peers (e.g. [23]) are used to highlight progress towards desirable levels of performance (termed targets or goals). In terms of usability, the use of line graphs to monitor trends in performance in an e-A&F system have been found to be useful, in addition to the ability to interactively explore aggregated patient data, and compare performance between departments within an organisation [16]. However, it is unclear how these functions should be optimally designed, or integrated with other formats of data presentation.

2. **Patient lists**: Some e-A&F systems provide lists of patients who have (e.g. [24]) or have not (e.g. [15]) received the recommended clinical practice, or experienced the particular outcome of interest. This is generally supplemental to the summary of clinical performance (e.g. [20]), though occasionally may act as its proxy (e.g. [19]). The intention in providing patient lists is that they can be used to further investigate the care of individual patients and take corrective action where necessary [25]. Patient lists have been identified as a key driver of success in some non-electronic A&F interventions [26], and their absence as a reason for failure [27]. They may simply contain patient names or identifiers, or additional summary data such as demographics or physiological measurements (e.g. [20]). We are unaware of any published studies of e-A&F interventions that have assessed the usability of patient lists, so evidence regarding their optimal design is lacking. For example, it is unclear how they should be integrated with the
summary of clinical performance, or how (and whether) they should include patient-specific summary data as a means of improving information processing and cognitive load during interpretation tasks.

3. Detailed patient-level data: e-A&F systems may occasionally further supplement patient lists with more detailed information about each patient (e.g. historic glycated haemoglobin readings for diabetic control [28]). Access to these data, whether within the e-A&F system itself or the EHR, is key so that individual patients’ care can be reviewed, and action taken where necessary [27]. In e-A&F systems, such information may be presented in tables (e.g. [15]) or graphically (e.g. [16]). From a usability point of view, integrating patient-level with population-level data in an e-A&F system has been demonstrated as desirable to users, and that functionality should support information visualisation over predefined time periods in addition to interactive exploration [16]. Similarly, a usability evaluation of a primary care epidemiological visualisation tool found that providing these data within the system was advantageous as clinicians may not have time to check each patient’s EHR [29]. However, it is unknown how best to present such detailed patient-level data within an e-A&F system, or how much data to present without overwhelming the user and increasing cognitive load during task performance [30].

4. Suggested actions: The definition of A&F states that suggested actions for improvement may accompany clinical performance feedback [2]. There is both theoretical [31] and empirical evidence [1] that suggesting actions increases the effectiveness of A&F. Often A&F recipients do not have the time, capacity or skills to interpret feedback and formulate what improvement action is necessary [27], so providing recommendations increases the likelihood that action is taken [31]. User-needs assessments for e-A&F systems often find that recommended actions are desirable [23]), and some systems provide links to educational materials such as best practice guidelines (e.g. [28]) or templates for users to formulate their own action plans (e.g. [32]), however we are only aware of one e-A&F system in which improvement actions are actually recommended to users (the LPZ Dashboard [23]). The recommendations in this system are generic and target organisational changes only, which the user derives themselves using a decision tree [23]. The usability of this system was not evaluated, so it is unclear how best to present recommended actions within an e-A&F system.

In addition to the knowledge gaps regarding each of the four interface components described above, there is also little insight into how they should be effectively integrated in a single e-A&F system in a manner that aids information processing, minimises technology-induced errors, and reduces cognitive load during interaction. It is therefore important to investigate the usability of e-A&F systems in more depth to produce evidence
that can guide their design. Developing health information systems without regard for user interaction can reduce their effectiveness, with adverse consequences for patient safety and care quality [33,34]. This is particularly important for e-A&F systems, where the non-use or misuse of clinical performance data can lead to suboptimal care on a large scale with important adverse implications for patient outcomes and cost (e.g. [35]). Conversely, effective use of A&F has the potential to vastly improve care quality: the latest Cochrane review of A&F found it can increase desired care processes by up to 70% [1], which if multiplied across large populations can lead to major gains. However if used ineffectively, A&F can decrease desired care processes up to 9% [1]. Whether A&F is effective or ineffective is partly determined by how clinical performance feedback data is presented to users [1].

**Aim and objectives**

The aim of this paper is to address the gaps in the literature identified above by evaluating the usability of an e-A&F system for primary care (the Performance Improvement plaN Generator; PINGR). To the best of our knowledge, PINGR is the first reported system to comprise all four interface components found in e-A&F applications (summaries of clinical performance, patient lists, patient-level data, and recommended actions). Further originality of PINGR relates to its design being informed by existing usability evidence and relevant behaviour change theory (we are aware of only two reported e-A&F systems that explicitly incorporated existing usability guidelines and theory in their design [15,36]). To evaluate the PINGR system, we adopted an iterative approach to system design involving multiple evaluation cycles at different stages of the development process [37]. This paper reports on the findings of the usability inspection study during the first part of our evaluation cycle. The specific objectives of this paper are to:

1. Describe PINGR’s interface design, rationale and theoretical basis;
2. Identify usability issues with PINGR in relation to its four interface components;
3. Understand how these issues may interfere with the cognitive goals of end-users (and therefore the integration of the interface components);
4. Translate these issues into recommendations for the user-centred design of e-A&F systems in general.

The paper is structured as follows: in the next section we present an overview of the PINGR system, and discuss its design and theoretical justification by drawing on relevant design guidelines, usability research and theory (Objective 1). The following two sections then report the methodology and results of the usability inspection study (Objectives 2 and 3). The final section presents a discussion of the results and design recommendations (Objective 4).
PINGR is a web-based e-A&F tool designed to help UK primary care clinicians improve the quality of care they provide to patients. The version presented in this paper focuses on use cases of hypertension and asthma. PINGR was conceived, designed and built by author BB (a UK primary care physician and health informatics researcher), with input from author RW (a software engineer). It is intended for use by clinicians outside patient consultations to assess the care provided by a primary care practice to its patient population, and to inform subsequent improvement actions at an individual, team, and service level. It analyses EHR data in the form of Read v2 codes, though has the capability to handle any type of structured data. These data are processed in a SQL Server database optimised for query execution. PINGR’s interface is built with HTML, JavaScript and CSS, using libraries including Bootstrap (http://getbootstrap.com/), C3.js (http://c3js.org/), jQuery, and Mustache.js (http://mustache.github.io/). Given the paucity of research into e-A&F system usability, its design primarily draws on literature regarding user needs and theory for A&F interventions identified in ongoing work by our group [17], in addition to design guidelines for other types of health information system. These design guidelines were selected based on their similarity and relevance to each of PINGR’s four interface components (summaries of clinical performance, patient lists, patient-level data, and recommended actions). For example: recommendations for displaying quantitative information (e.g. [38]) were used to inform the design of summaries of clinical performance because they contain quantitative quality indicators; EHR design guidelines (e.g. [39]) were used for the design of patient lists and patient-level information because they are common elements of EHRs; and CDS system design guidelines (e.g. [40]) were used to inform the design of recommended actions because they often suggest actions for users [41]. The remainder of this section describes the design and rationale of PINGR’s interface components.

Summaries of clinical performance

Clinical performance summaries for each clinical area within PINGR are organised as separate modules, accessed from an icon-based menu on the left side of the interface (Figure 1). A module-oriented design was employed to enhance information processing as demonstrated in clinical guidelines [42] and general web design [43]. Within each clinical module, there are 2 pages: 1) an Overview page (Figure 1), and 2) a Preview page (zoom and filter; Figures 2a and 2b). Overview and preview has been widely used in the design of applications to support visual information-seeking tasks [44]. After selecting a module from the menu, the Overview page is displayed which presents the primary care practice’s clinical performance as quality indicators. These quality indicators (described in further detail below) convey proportions and absolute numbers of patients who have received a recommended clinical practice, or experienced a particular outcome. Using both relative...
and absolute measures of performance avoids potentially misleading effects of providing isolated measures of relative performance [45]. To create a generic template for all clinical conditions, and consistency of interface design as recommended in EHR [39] and CDS system [40] usability guidelines, PINGR organises quality indicators into four common areas along a clinical condition pathway: diagnosis, monitoring, treatment, and exclusions. To illustrate, the hypertension area of the system displays the following elements: patients with diagnosed hypertension (and other relevant conditions such as chronic kidney disease or diabetes) based on their prior recorded measurements (diagnosis); hypertensive patients who have had their blood pressure measured in the preceding year (monitoring); hypertensive patients whose latest blood pressure measurement is within their recommended personalised target (control); and hypertensive patients who have been excluded from quality standards, such as those with a terminal illness (exclusions).

In accordance with data visualisation design principles [38], and to reduce short-term memory load [46], the four quality indicators are presented as separate panels on a single screen to provide the user with an overview of the practice’s clinical performance in that disease area. Quality indicators are displayed as line graphs for trend visualisation (i.e. in monitoring and control indicators), and bar plots for processing of one-off data points where it was anticipated to be most clinically informative (i.e. diagnosis and exclusions), which is supported by feedback intervention theory [31], cognitive fit theory [47], and evidence from a usability study [16]. Each graph is supplemented with labels to indicate the current level of clinical performance, interactive tool-tips to detail historical performance, and icons to highlight the change in performance from the previous month [38].

Users can request further information regarding their clinical performance by clicking on the quality indicator graphs thereby accessing the Preview page (Figures 2a and 2b). The Preview page is organised with the quality indicator graph in the top left hand corner, with the remaining interface elements (discussed in detail below) arranged with patient lists on the right, and patient-level data and recommended actions at the bottom. This layout mirrors the anticipated reading pattern [48] and workflow that users would follow: reviewing their summary of performance, then list of patients requiring action, followed by detailed patient-level information, and recommended improvement actions. Displaying all 4 interface components on one page was also intended to reduce cognitive load and improve task completion by supporting recognition rather than recall of available user options [46].

**Patient lists**

Patient lists are populated with patients who have not achieved the clinical standard or desired outcome in the quality indicator to help users take corrective action where
appropriate. The lists included patients’ unique identification number, which can be cross-referenced with an EHR system using a ‘copy’ icon to prevent errors [39]. An additional column displays clinical data felt most relevant to the quality indicator (e.g. latest blood pressure reading for the ‘control’ quality indicator), which can be used to order the list and prioritise patients for action. Ordering and prioritisation patient lists has been indicated as valuable in non-electronic A&F interventions [27], and is consistent with design guidelines for EHRs [49]. The current version of PINGR only provides one column of patient attributes based on empirical evidence that displaying multiple clinical variables can adversely affect the usability of primary care epidemiological visualisation tools [29]. The lists can be filtered by selecting sections of the ‘improvement opportunity’ graph (see recommended actions section below), acting as an interactive visual query mechanism as recommended by usability research into e-A&F systems [16], EHRs [49] and quantitative data display in general [44].

**Detailed patient-level data**

When a specific patient is selected from a list on the Preview page, the bottom panel displays an ‘individual patient’ tab (Figure 2B), which can also be accessed by entering a patient’s unique identifier into the search bar located in a menu at the top of the screen. Here detailed patient-level data relevant to the quality indicator is presented, in addition to patient-level suggested actions (which are discussed in detail below). In the hypertension module this information relates to patients’ blood pressure measurements, whereas in the asthma module it relates to their peak expiratory flow rate. Based on previous usability research into e-A&F systems [16], these data are presented as line graphs. Interactive tool-tips provide detail on historical data as recommended in EHR usability guidelines [50]. Further information on relevant non-physiological events are also presented on the graphs as vertical lines such as when a patient had an encounter with the practice, or experienced a change in medication.

**Suggested actions**

PINGR provides suggested actions in the bottom panel of the Preview page (Figure 2). In contrast to the LPZ Dashboard system [23], these recommendations address both the organisation (i.e. the primary care practice) and individual patients, are specific to users’ clinical performance (rather than generic), and are provided automatically (rather than on-demand). These design choices were based on existing evidence and theory: providing two types of recommended action is consistent with health care quality improvement theory [51], whereas providing tailored recommended actions in a user’s workflow is recommended in CDS system design [40,52]. In this sense, PINGR can be viewed as a cross-fertilisation of traditional A&F and CDS systems producing ‘decision-supported
feedback’, which we have previously argued could lead to greater effectiveness of both types of system [41].

Suggested actions are derived through further analysis of contextual data of patients who have not achieved the quality standards, which is supported by CDS system design guidelines [53]. These patients are subsequently grouped into ‘improvement opportunity’ categories that infer potential reasons why patients have not achieved the quality standards or outcomes of interest, and are associated with a specific set of potential solutions in the form of recommended actions (both at the organisational and patient levels). The improvement opportunity categories and bank of recommended actions are generated from clinical guidelines, research literature (e.g. [54]), and empirical analysis of medical records [55]. To illustrate: in the hypertension monitoring quality indicator, improvement opportunity categories relate to patients’ contact with the primary care practice: either face-to-face, non face-to-face (e.g. over the phone), or no contact [55]. An algorithm analyses EHR data from each hypertensive patient who has not met the quality standard, makes inferences regarding the type of contact each patient had with the practice [55], and provides relevant recommended actions to help these patients attain the quality standard.

The proportions of patients in each improvement opportunity category are displayed in a panel to the right of the quality indicator graph as a pie chart (Figures 2a and 2b), which act as the visual query mechanism to filter the patient list described above. Clicking on a section of the pie chart filters the list to display each patient in that improvement opportunity category. The intention was that this would facilitate user action by grouping patients associated with similar improvement tasks, thus minimising cognitive load [30]. The recommended actions are automatically displayed in a table, where users can also add their own actions as free text (Figure 2a and 2b). Users can agree or disagree with them by clicking a ‘thumbs up’ or ‘thumbs down’ icon respectively (Figure 2a and 2a). If a user agrees with a recommended action, it turns green and is saved to their personal bank of actions in the home page in accordance with CDS system usability design [56]. Users can indicate when a task has been completed using a check box, and can download their list of agreed actions as a document to print or share. If a user disagrees with an action, a dialogue box captures the reasons for this using fixed responses or free-text as recommended in the design of EHRs and CDS systems [39,52,53]. Framing recommended actions as advice rather than commands is in accordance with design guidelines for CDS systems [40], and asking for reasons for override has been shown to improve their effectiveness [57].
Figure 1: PINGR’s Overview page (example is hypertension)
2A. Team / Organisational-level recommended actions tab
2B. Patient-level data and recommended actions tab
5.5 Materials and Methods

We evaluated usability issues associated with PINGR using a hybrid usability inspection method, which combined Heuristic Evaluation (HE) and Cognitive Walkthrough (CW). Usability inspection methods involve experienced evaluators assessing a system to identify issues that could potentially hinder user interaction with the software. They are recommended as a cost-efficient initial step in usability evaluation as they can identify a wide range of issues without the need for real end-users (in this case, primary care clinicians) or significant resources [58]. At this stage of our iterative evaluation process, the involvement of experienced evaluators was necessary to identify and correct critical usability issues according to established usability guidelines. In accordance with accepted usability engineering methodology, real end-users will be involved in future evaluation rounds of PINGR to capture any issues that may have been overlooked [34].

HEs and CWs are often recommended to be carried out separately on a system [34], which has both advantages and disadvantages. HEs assess interfaces against a set of well-established design guidelines known to play an important role in user experience, and do not restrict the evaluator to interact with the interface in a specific way, thereby maximising usability issue discovery [58]. This is important for e-A&F systems in general where there is a lack of usability knowledge, and for the PINGR system in particular, which has not been previously evaluated. However, among other limitations (e.g. [59]), HEs do not adequately explore how issues arise during user interaction with a system beyond its static interface features, nor how they relate to the user’s cognitive needs [60]. This is particularly important in health IT systems, such as PINGR, with dynamic user interfaces that require complex interactions to achieve user goals [61]. This limitation can be addressed by the CW method, though traditionally CWs do not take advantage of accepted usability heuristics, which may limit their ability to identify potential issues [62]. Independent HEs and CWs often discover different usability issues in the same system [63], making it cumbersome to combine their relative advantages if used separately. Therefore as the complexity of health information systems progress, there is a need to harness the combined benefits of HEs and CWs into a single hybrid usability inspection technique; though as yet, little progress has been made [61].

Participants and setting

We recruited a convenience sample of health information system evaluators from the Centre for Health Informatics, University of Manchester. Eligible participants were qualified software developers or evaluators with more than five years’ experience in health information system design and development. Using three to five usability experts is recommended in HEs as a balance between costs and benefits, and is expected to detect around 75% of usability issues in a system [64,65]. However, given our objective was to
identify as many unique usability issues as possible, we invited eight potential evaluators, all of whom accepted. None of the evaluators had previously used PINGR, though all had experience of using similar systems such as non-clinical dashboards, and epidemiological surveillance tools. All stages of the evaluation took place at the University of Manchester where evaluators accessed the PINGR application via the Google Chrome web browser on a 17-inch computer screen. To preserve patient privacy, we used simulated data for the purposes of the usability inspection.

Hybrid inspection method

Our hybrid method incorporated elements of both HE and CW, adapting the approach advocated by Kushniruk et al. [61]. It comprised five stages (Figure 3): 1) Development of representative user tasks and their transformation into goals and actions; 2) Combining HE and CW methods into a single protocol to identify usability issues regarding the PINGR application; 3) Consolidation of usability issues identified by evaluators in stage 2; 4) Severity rating of consolidated usability issues; 5) Analysis according to usability heuristics, interface component, and Goal-Action structure. These stages are described in further detail below.

Figure 3: Overview of hybrid usability inspection methodology

![Diagram showing the stages of the hybrid usability inspection methodology]

Stage 1: Development of representative user tasks and their transformation into goals and actions

Initially we followed the typical procedure for a CW evaluation by describing tasks and their associated goals to be used in the evaluation (Figure 4; Task description; User's initial goal/s) [66]. Eight representative user tasks were selected, piloted and refined, to guide interaction with all components of the PINGR interface (Table 1). Each task was decomposed into up to 8 constituent actions, and their optimal sequence determined to minimise cognitive effort to achieve each task’s goal (Figure 4; Action sequence) [66]. There were 44 actions in total across the 8 tasks. Characteristics and needs of intended users were also described for each task (Figure 4; Anticipated users). This information was used to produce a Goal-Action structure document (Appendix 6) to contextualise each task for the interface evaluators in the next stage.
Stage 2: Combining HE and CW methods into a single protocol to identify usability issues

Evaluators worked independently rather than as a group in order to identify a larger and more diverse number of usability issues [67]. Each evaluator met individually with author BB face-to-face; they were introduced to the objectives and methods of the study, and the aims, high-level functionality, and rationale of the PINGR system using a standardised script. A demonstration of how to use PINGR was not provided in order to evaluate the learnability of the system [68]. As in standard CW protocol, each evaluator then investigated the interface following the tasks in the Goal-Action structure document (Appendix 6). For each Action the evaluator applied Nielsen’s 10 usability heuristics to identify usability issues according to its categories [69]: 1) Visibility of system status; 2) Match between system and the real world; 3) User control and freedom; 4) Consistency and standards; 5) Error prevention; 6) Recognition rather than recall; 7) Flexibility and efficiency of use; 8) Aesthetic and minimalist design; 9) Help users recognise, diagnose, and recover from errors; 10) Help and documentation. These generic heuristics were chosen due to a lack of specific heuristics for e-A&F systems. If a usability issue was identified, evaluators took screenshots and described it in detail in an electronic data collection form (Appendix 7). They also recorded the task and Goal-Action(s) in which it occurred, the heuristic category with which it was associated, and their rating of its
severity on a 4-point scale [58] (Table 2). Both the heuristic categories and severity ratings were provided in an electronic document for reference. To make the process more constructive, evaluators also provided suggestions as to how each issue could be improved (if this was not obvious), and once they had completed the tasks listed up to three positive aspects of the system. Any missing data or unclear descriptions were clarified by BB who was present throughout the process. As in a standard HE, participants were encouraged to explore usability issues outside the specified goal-action structure to assess general aspects of PINGR’s functionality and record them under the relevant task. Each participant took on average one hour to perform their evaluation, and in total identified 132 issues with a mean severity of 2.

Table 1: Overview of tasks performed by evaluators during usability inspection

<table>
<thead>
<tr>
<th>Number</th>
<th>Brief description of task and Goal</th>
<th>Interface components assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agree with team/organisation-level action plan</td>
<td>Menu Clinical performance summary Recommended actions</td>
</tr>
<tr>
<td>2</td>
<td>Disagree with team/organisation-level action plan</td>
<td>Menu Clinical performance summary Recommended actions</td>
</tr>
<tr>
<td>3</td>
<td>Agree with patient-level action plan</td>
<td>Menu Clinical performance summary Patient lists Recommended actions</td>
</tr>
<tr>
<td>4</td>
<td>Disagree with patient identification</td>
<td>Menu Clinical performance summary Patient lists Patient-level data</td>
</tr>
<tr>
<td>5</td>
<td>Population-level data interpretation</td>
<td>Menu Clinical performance summary</td>
</tr>
<tr>
<td>6</td>
<td>Patient-level data interpretation</td>
<td>Menu Clinical performance summary Patient lists Patient-level data</td>
</tr>
<tr>
<td>7</td>
<td>Adding action plan</td>
<td>Menu Clinical performance summary Recommended actions</td>
</tr>
<tr>
<td>8</td>
<td>General functionality*</td>
<td>Search box Patient lists Recommended actions</td>
</tr>
</tbody>
</table>

* Included: searching for a specific patient, ordering the lists of patients according to specific criteria, and downloading a summary of activity/actions made using the PINGR system.

Table 2: Usability issue severity rating scale

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cosmetic issue only. Need not be fixed unless extra time is available on project</td>
</tr>
<tr>
<td>2</td>
<td>Minor usability issue. Fixing this should be given low priority</td>
</tr>
<tr>
<td>3</td>
<td>Major usability issue. Important to fix, so should be given high priority</td>
</tr>
<tr>
<td>4</td>
<td>Usability catastrophe. Imperative to fix this before product can be released</td>
</tr>
</tbody>
</table>

Stage 3: Consolidation of usability issues

All usability issues collected from each evaluator in stage 2 were collated into a single document. Two authors (BB and PB) worked independently to consolidate the issues
using an interpretivist approach by: 1) Integrating semantically similar issues into one issue; 2) Removing issues identified by only one evaluator, and rated as a ‘cosmetic issue only’ (Table 2) to reduce the occurrence of ‘false positive’ issues associated with traditional HEs [70]; 3) Removing issues not directly related to the usability of the application, such as suggestions for new system functionality; and 4) Assigning each consolidated issue to the most appropriate heuristic category, component of PINGR’s interface, and task Action(s) in which it arose. The final list was agreed through discussion, which consisted of 47 unique usability issues (Figure 5). Out of these 47 issues, 24 (51%) had been identified by a single evaluator, of which 15 were rated as ‘minor’ (63%), 7 as ‘major’ (29%), and 2 as ‘usability catastrophes’ (8%; Table 2). This suggests our decision to use multiple evaluators working independently achieved our objective of identifying as many diverse usability issues as possible.

**Stage 4: Severity rating of consolidated usability issues**

Each evaluator who participated in Stage 2 was sent by e-mail a list of the finalised usability issues and asked to rate their severity using an electronic questionnaire (Appendix 8). The task number(s) and location in the Goal-Action structure with which each usability issue was associated was provided. The same severity rating scale was used from Stage 2, though to account for issues being identified by only one evaluator in Stage 2 an additional point was added: ‘I don’t agree that this is a usability issue at all’ (non-issues) [58]. Due to the gap of approximately one week between Stage 2 and Stage 4 that could have adversely affected participants’ recall of the issues, a hyperlink to PINGR was provided along with the original list of tasks used in stage 2 (Appendix 6). Participants were encouraged to remind themselves of usability issues they had previously identified, and familiarise themselves with issues that had been identified by others by navigating the system again using the Goal-Action structure.

**Stage 5: Statistical analysis**

We calculated the mean severity rating for each issue to the nearest integer to aid interpretation and prioritisation according to our scale (Table 2). Issues and positive comments were subsequently analysed thematically, and grouped according to interface component, and by their occurrence during user Goals and Actions. For stage 4, we measured inter-rater agreement (IRA), the extent to which evaluators assigned the same value for each item, and inter-rater reliability (IRR), the extent to which different evaluators consistently distinguished between different items on the severity scale [71]. We evaluated IRA by calculating simple proportions of agreement, and the Kendall coefficient of concordance adjusted for ties [71]. We evaluated IRR by calculating intra-class correlation coefficients (ICC) using a one-way model to estimate consistency of single ratings; Light’s weighted kappa; and Krippendorff’s alpha [71]. All measures of IRA and
IRR used range between 0 and 1, with 1 signifying complete agreement or reliability. All analyses were performed using R [72], and the packages ‘irr’ [73] and ‘psy’ [74].

Figure 5: Flowchart of usability issue discovery and finalisation

5.5 Results

The final list of 47 issues were categorised into 8 heuristic themes (Figure 6): ‘Flexibility and efficiency of use’ had no usability issues, and we combined the heuristics ‘Error prevention’ and ‘Help users recognise, diagnose, and recover from errors’ (‘Error prevention and recognition’) due to their issues’ conceptual similarity. The most error-prone heuristics were ‘Consistency and standards’ (13 usability issues; 28% of the total) and ‘Match between system and the real world’ (n=10, 21%). The least violated heuristics were ‘Recognition rather than recall’ (n=1, 2%), and ‘Help and documentation’ (n=1, 2%). Analysis of mean severity ratings revealed 12 (26%) major usability issues, 26 (56%) minor issues, and 9 (19%) cosmetic issues; no usability catastrophes or non-issues were identified. Twenty-four positive comments were made about PINGR, 13 (54%) of which related to the system in general, praising its clean and visually appealing design, responsiveness, intuitive and simple layout, and use of contextual tool-tips. All eight evaluators did not agree on the exact severity of any issues, though within a tolerance of one point agreed on the severity of 8 (17%). Kendall’s coefficient of concordance was 0.44, indicating weak-moderate agreement [75,76]. The ICC and Weighted Light’s Kappa statistic were both 0.33, whilst Krippendorff's alpha reliability coefficient was 0.04, indicating poor-fair reliability [77,78].
Interface components

Suggested actions had the most usability issues (n=21, 45%), followed by detailed patient-level data (n=5, 11%), patient lists (n=4, 9%), and summaries of clinical performance (n=4, 9%). The remaining 13 (28%) issues were associated with other non-unique aspects of the PINGR interface concerned with system navigation. The suggested actions received the most positive comments (n=5, 21%), followed by the summaries of clinical performance (n=4, 17%), and detailed patient-level data (n=2, 8%). Patient lists received no positive comments. Below we present these usability issues in detail organised by heuristic, and discuss positive comments. For brevity we only describe issues in detail with a mean severity rating of ‘minor’ or above.
Summaries of clinical performance

Issues with summaries of clinical performance were categorised under the heuristics ‘Match between the system and the real world’ (n=2), ‘Consistency and standards’ (n=1), and ‘Aesthetic and minimalist design’ (n=1). Under ‘Match between the system and the real world’, issues concerned the use of a cross icon to represent excluded patients as this is generally used to indicate an exit action, and that bar plot data points required clearer labelling. Under ‘Consistency and standards’, it was sometimes unclear what aspects of clinical performance the quality indicators were specifically measuring (rated as at least a ‘major’ usability issue by 4 out of 8 evaluators), whilst in ‘Aesthetic and minimalist design’ it was noted that plots did not re-size well with the internet browser window. Positive comments were made about the use of different colours to indicate the 4 pathways (diagnosis, monitoring, treatment, exclusions), and that although users were presented with a lot of information, it was generally felt to be easy to understand, particularly with the use of tool-tips to find historical performance data on line graphs.

Patient lists

Issues with patient lists were categorised under the heuristics ‘Match between the system and the real world’ (n=2) and ‘Visibility of system status’ (n=2). Under ‘Match between the system and the real world’, issues concerned: a lack of clarity as to what the different lists referred (rated as at least a ‘major’ usability issue by 7 out of 8 evaluators); difficulty in browsing due to the lack of visible ordering options; and the need to use more useful parameters by which they could be ordered. Under ‘Visibility of system status’, issues concerned a lack of feedback when a new patient had been selected, or when a list had been filtered by interacting with the improvement opportunities graph (rated as at least a ‘major’ usability issue by 6 out of 8 evaluators).

Patient-level data

Issues with patient-level data were categorised under ‘Match between the system and the real world’ (n=3), ‘Aesthetic and minimalist design’ (n=1), and ‘User control and freedom’ (n=1). Under ‘Match between the system and the real world’ and ‘Aesthetic and minimalist design’, issues concerned difficulty reading non-physiological data on line graphs, such as when patient medication had been changed. The issue under ‘User control and freedom’ concerned the relatively small size of the graphs, that did not re-size automatically, and which caused occasional difficulty in data interpretation. Positive comments were made about having detailed patient-level data displayed in general, and the use of tool-tips to understand historic physiological data on line graphs.

Suggested actions
Issues associated with the recommended actions were categorised under the heuristics ‘Consistency and standards’ (n=7), ‘Aesthetic and minimalist design’ (n=4), ‘Error prevention and recognition’ (n=3), ‘User control and freedom’ (n=3), ‘Visibility of system status’ (n=2), ‘Help and documentation’ (n=1), and ‘Recognition rather than recall’ (n=1). Under ‘Consistency and standards’, issues concerned conflicting use of language and font sizes, redundant column headers for user-generated actions, illogical ordering of options in dialogue boxes, and the positioning of recommended actions. Under ‘Aesthetic and minimalist design’, issues were deemed cosmetic issues only. Under ‘Error prevention and recognition’, issues concerned being able to edit a user-generated recommended action plan that had been marked “complete”, and technical faults related to deleting and downloading recommended actions (all rated as at least ‘major’ usability issues by 7 out of 8 evaluators). Under ‘User control and freedom’, issues concerned the inability to view, undo or edit reasons for disagreeing with recommended actions, or add user-generated recommended actions to the Home page. Under ‘Visibility of system status’ issues were: a lack of clarity as to whether marking a recommended action as complete had been saved by the system, and a loss of context when the dialogue box for providing disagreement reasons appeared. The issue under ‘Help and documentation’ recommended there should be some explanation of how the suggested actions were generated, whilst under ‘Recognition rather than recall’ it related to clearer signposting of the copy functionality for inputting patients’ unique identification numbers in other systems (e.g. EHRs). Positive comments were made about having the recommended actions in general, in addition to specific features including the interactive improvement opportunity graph to filter patient lists, ability to add user-generated actions, agree or disagree with actions, and provide reasons for disagreement in the form of both fixed responses (radio buttons) or more detailed free text.

**Task and Goal-Action structure**

The 47 usability issues occurred 121 times in total across all tasks (median occurrences per issue of 2, range 1-12). In terms of both frequency and severity of usability issues, Task 4 (Disagree with patient identification) had the most usability violations (26 usability issue occurrences, 21% of the total). This was followed by Task 3 (Agree with patient-level action plan, n=18, 15%), Task 2 (Disagree with team/organisation-level action plan, n=16, 13%), Task 7 (Adding action plan, n=15, 12%), Task 6 (Patient-level data identification, n=14, 12%), and Task 5 (Population-level data identification, n=12, 10%). Task 1 (Agree with team/organisation-level action plan) and Task 8 (General functionality) were the most issue-free with only 10 (8%) issues each. At the Goal-Action structure level (Appendix 6), the most issue-prone actions across all Goals were: 1) Patient selection from a list (which affected Goals 3, 4, and 6); 2) Data interpretation from a figure (both population-level and patient-level; Goals 5 and 6); and 3) Disagreement with a system
proposition (this included: recommended actions [Goal 2] or categorisation of patients into an improvement opportunity group [Goal 4]). The remainder of this section describes how these actions impacted the completion of a given Goal or sub-Goal.

**Patient selection from a list**

In Goals 3, 4, and 6, users navigated to the Overview page of the relevant clinical module (Figure 1), and selected the summary of clinical performance to investigate further in the Preview page (Figure 2a). Evaluators were then required to select a patient from the patient list either directly (Goals 4 and 6), or by first filtering the list using the improvement opportunity graph (Goal 3). At this point, the Goal-Action sequence was likely to be interrupted or become unwieldy due to: a lack of clarity as to what the different lists referred, and why they contained different patients; difficulty in browsing the patient lists due to a lack of visible ordering options and perceived lack of utility of the options by which they could be ordered; and an absence of feedback that a patient list had been filtered, or that different patient-level data was presented. To illustrate, in Goal 3 (Appendix 6) users would expect after selecting an improvement opportunity from the graph on the Preview page (action: ‘select palliative care’) that the filtering of the patient list would be apparent before proceeding to the following action (‘select patient 5556051664 from the list’). This issue was categorised under the heuristic ‘Visibility of System Status’, therefore making the status of the list in this part of the action sequence clearer would make the relationship between the two actions more natural for the user. The remaining issues related to the browsing and ordering of the list would make the process of completing the specific action of selecting a patient from a list less efficient, though were unlikely to disrupt the user’s action sequence.

**Data interpretation**

In Goals 5 and 6, evaluators initially navigated either to the Overview (Figure 1; to interpret population-level data) or Preview (Figure 2a and 2b; to interpret patient-level data) pages respectively. In the Goal-Action structure, evaluators were then required to identify specific data points using the corresponding data visualisations (i.e. line graphs, bar plots, or pie-charts). At this point, the Goal-Action sequence was likely to be interrupted or hinder information processing due to: a lack of clarity regarding the specific aspects of clinical performance the graphs represented; the relatively small text size used for axis labels; unclear explanations for bar plot categories; and the use of non-standard date format (i.e. yyyy/mm/dd). Furthermore, interpreting patient-level non-physiological data (e.g. when medication had changed) were difficult because of misalignment with x-axis dates, an absence of tool-tips, and unclear labelling. To illustrate, in Goal 5, sub-Goal ‘identify how many patients have had face-to-face opportunities to have their asthma monitored’ (Appendix 6), users would expect to easily recognise precisely what the
summary of clinical performance referred to on the Overview page (action: ‘select monitoring’) before proceeding to the following action (‘check the corresponding figure to identify how many patients have had face-to-face opportunities to have their asthma monitored’). This issue was categorised under the heuristic ‘Consistency and standards’, therefore making the specific aspects of clinical performance the graphs represented clearer would reduce the cognitive demands necessary for a user to understand how to access the relevant Preview page. The remaining issues made the identification of specific graph data points less efficient, though were unlikely to disrupt the prescribed action sequence.

Disagreement with a system proposition
In Goals 2 and 4, users initially navigated to the Preview page of the relevant summary of clinical performance (Figure 2a). At this point they were required to either disagree with an organisational-level action plan (Goal 2; Figure 2a), or select a patient and disagree with the improvement opportunity to which it had been assigned (Goal 4; Figure 2b). Despite the fact that several usability issues were identified in the context of recommended actions (Figure 6), all of which could hinder the processing of information and increase the time needed to complete an action, none were likely to disrupt the Goal-Action sequence. To illustrate, in Goal 2 (Appendix 6) users would expect the recommended action plans to be in a conspicuous location and written in prominent font to facilitate their identification (action: ‘check the available option disagree for the action: “nominate an asthma lead … of these changes”’). Furthermore, they would expect to know how the recommended actions were generated in order to judge whether or not they agree. These issues were categorised under the heuristics ‘Consistency and standards’ and ‘Help and documentation’ respectively, therefore improved presentation of recommended actions, and provision of information regarding how they were generated would reduce the attentional and cognitive demands necessary to complete these Goals.

5.6 Discussion
Our results indicate important considerations that are specific e-A&F systems, and which should be taken into account in designing their interfaces. This final section discusses the significance of the usability issues found with PINGR, and translates them into a set of interface design recommendations for e-A&F systems in general by placing them in the context of the wider literature (Box 1). Each of the four components of e-A&F interfaces (summary of clinical performance, patient lists, detailed patient-level data, and suggested actions) are considered in turn, followed by a final section on how they could be integrated. The paper concludes with a discussion of the strengths and weaknesses of this study, and implications for future research.
Interface design recommendations for e-A&F systems

Summary of clinical performance

e-A&F system design should draw on existing usability guidance and theory for the presentation of clinical performance summaries [16,31], in addition to relevant guidance on quantitative information visualisation in general (e.g. [38,44]), and related IT systems including epidemiological surveillance tools (e.g. [12]) and non-clinical dashboards (e.g. [30]). Key recommendations include using line graphs to demonstrate trends over time [16,31,47], and interactive functionality to provide further detail on-demand [16,44]. In addition, our results show that the use of tool-tips can facilitate accurate interpretation of historic performance data on line graphs, and that care should be taken to ensure what performance data specifically refers. If this is not the case, users may disengage with the system, with potentially important implications for patient outcomes and resource-use (e.g. [35]).

Patient lists

Not all e-A&F systems provide lists of patients (e.g. [5]), despite evidence from non-electronic A&F interventions suggesting they are key drivers of success [26,27]. Therefore, a key recommendation is to include patient lists as a core part of e-A&F interface design. These may include patients who have or have not achieved the quality standard or patient outcome of interest. The design of patient lists may utilise existing evidence from non-electronic A&F interventions [26,27], in addition to usability guidelines from related health systems such as EHRs (e.g. [39,49]) and epidemiological surveillance tools (e.g. [29]). Key recommendations include the ability to order and prioritise patients for action [27,39], and providing a manageable number of variables by which to order patients [29]. Our results add that the use of visual querying mechanisms to filter lists, and the ability to order lists can also be helpful, which is supported by studies of other e-A&F systems and wider usability guidelines [16,44]. However, our results also highlight that it should be apparent that lists have changed when they are filtered, thus providing clear feedback of system status. This may be achieved through the use of animation (e.g. self-healing fades) and significant changes to the text in the list header. Furthermore, enough information should be provided regarding what the lists refer to, in addition to making the ordering functionality obvious and using parameters perceived as valuable by users. If patient lists are not designed in a usable way, it may force users to identify patients in an inefficient manner, leading to reduced system effectiveness or disruption of the Goal-Action structure.

Detailed patient-level data

Not all e-A&F systems provide detailed patient-level data (e.g. [23]), therefore a key recommendation from our results is to include this interface component as a core design
consideration. This is supported by other studies of e-A&F systems [16], and ensures the system can efficiently support improvement action [25,27]. The design of this component should draw on existing usability knowledge [16], in addition to design guidelines for systems that summarise individual patient-level data, such as EHRs (e.g. [50]). Important recommendations include the use of line graphs to support trend visualisation over defined time periods, and the ability to interactively explore data further [16]. Our results also suggest that similar to population-level data, tool-tips can be helpful to understand historic data. In addition, provision of non-physiological data, such as contacts with the primary care practice or changes in medication, can improve feedback actionability, though must be displayed as clearly as the medical data to be effective. This could be achieved through novel information visualisation techniques such as Lifelines [79]. Finally, our results also illustrate the importance of highlighting that new patient-level data is displayed when a new patient is selected (e.g. from a list). If this is unclear users may be unsure how to access relevant patient-level data and therefore unable to take relevant improvement action. This may be remedied through the use of animation, or other design features such as presenting individual patient’s data on separate pages from each other.

**Suggested actions**

In addition to PINGR, we are aware of only one other e-A&F system that provides suggested actions to users [23]. This is surprising given such suggestions are part of the definition of A&F [2], that there is both theoretical [31] and empirical evidence [1] they increase A&F effectiveness, and that user-needs assessments for e-A&F systems state they are desirable [16,23]. Our results confirm this need, therefore a key design recommendation is that e-A&F systems should suggest actions for improvement; creating a cross-fertilisation of traditional A&F and CDS systems [41]. The design of recommended actions should draw on guidance regarding CDS systems that regularly provide advice to users (e.g. [40]) and wider quality improvement theory (e.g. [51]). Specifically, suggested actions should address both the individual patient and organisation [51], be specific to user’s context and performance (rather than generic) [40], be provided automatically in the user’s workflow (rather than on-demand) [40], take into account patient contextual data (such as co-morbidities) [53,56], and use concise statements [40,53] with functionality to easily action the recommendation [40,53]. Furthermore, our findings suggest that the ability to agree (and save) or disagree (and provide fixed and free-text response reasons) with recommended actions is well-received by users, which is supported by the wider CDS system literature [53,56], and may improve e-A&F effectiveness [57]. This feedback loop should be used to improve the algorithms driving the e-A&F system [25,80]. Our results also suggest that functionality to action recommendations is made clear, such as using tool-tips to highlight a ‘copy’ function, otherwise their effectiveness may be reduced. Information should also be provided on how the recommendation was generated [40,56],
and users provided with unambiguous and authoritative confirmation that their agreements or disagreements with the recommendation have been saved, otherwise there is risk users would find the system untrustworthy, and potentially ignore the recommendations. Mechanisms for providing disagreement reasons should not lose context from the recommended action, and there should be functionality to view, undo and edit reasons, otherwise users may disengage with this functionality and the gains in algorithm improvement would be lost. However, editing should not be possible once an action is marked complete, otherwise users may fear their work is lost. Suggested actions should be positioned in their own area separate from other data, using the same size font as other information, and in an area of the interface consistent with user workflow (which may not be the bottom of the page). This mitigates the risk they will be overlooked. Finally, users should also have the ability to add their own suggested actions, which should be available throughout the system where actions are displayed.

Integration of interface components

Interface components should be integrated in a way consistent with general software design guidelines (e.g. [46]), in addition to related health IT systems such as CDS systems (e.g. [53,56]) and EHRs (e.g. [39]). For example, components should be arranged in a way that anticipates user’s workflow [48], and patient-level data should be presented on the same screen as recommended actions so they can reliably be evaluated [53,56]. EHR usability guidelines recommend that detailed patient-level information should not be on the same page as patient lists [49], however, we suggest this may not be applicable to e-A&F systems because of the need to rapidly review multiple patients, and an absence of e-A&F functionality related to data input (unlike an EHR). It is also unclear whether detailed patient-level data should be displayed on the same screen as summaries of clinical performance (as in PINGR), as there is a theoretical argument it may motivate action regarding a single patient if a user is made aware of their wider performance [41].
Box 1: Interface design recommendations and future research questions for electronic audit and feedback systems

Summaries of clinical performance, should:
1. Use line graphs to demonstrate trends over time
2. Provide interactive functionality and further detail on-demand
3. Use tool-tips to interpret historic performance data
4. Explain clearly to what performance data specifically refer

Patient lists, should:
5. Be included as a key component of the e-A&F interface
6. Provide ordering functionality that is clearly labelled
7. Use a manageable number of variables (considered valuable by users) by which to order patients
8. Use visual querying mechanisms to filter contents, providing clear indication when the list changes
9. Explain clearly to what the lists refer

Patient-level data, should:
10. Be included as a key component of the e-A&F interface
11. Use line graphs to monitor trends over time
12. Provide interactive functionality to explore data
13. Use tool-tips to interpret historic data
14. Include relevant non-clinical data
15. Indicate clearly when a new patient's data is displayed

Recommended actions, should:
16. Be included as a key component of the e-A&F interface
17. Address both the individual patient and organisation
18. Be specific to a user’s context and performance (rather than generic)
19. Be provided automatically in the user's workflow (rather than on-demand)
20. Take into account patient contextual data (such as co-morbidities)
21. Use concise statements
22. Provide functionality to clearly and easily action the recommendation
23. Provide functionality to agree (and save) or disagree with recommendations
24. Collect reasons for disagreement as both fixed and free-text responses
25. Use disagreement reasons to improve system algorithms
26. Provide information on how recommendations were generated
27. Provided unambiguous and authoritative confirmation that agreements or disagreements with recommendations have been saved
28. Maintain context when asking for disagreement reasons
29. Provide functionality to view, undo and edit previous reasons given for disagreements
30. Not allow editing once an action is marked complete
31. Be positioned in a separate area of the interface, consistent with the anticipated user workflow
32. Use the same size font as other information
33. Provide functionality to add user-generated recommended actions throughout the system

Integration of components:
34. Components should be arranged in a way that anticipates user’s workflow
35. Patient-level data should be presented on the same screen as recommended actions

Questions for further research:
1. What interface components should be visible at the same time?
2. Which patients should be listed and how much information about them should be provided in the lists?
3. How should clinical performance summaries and recommended actions be prioritised?
4. How should patient-level data across multiple clinical areas and quality indicators be presented?
5. How should the clinical performance of other users be incorporated, if at all?
6. How do these guidelines translate into different contexts, clinical areas, and study designs?
7. Do these guidelines remain true when testing with real end-users?
Strengths and limitations of this study

The main strength of this study is that it is the first to develop usability recommendations for clinical e-A&F systems. This was achieved by evaluating a cutting-edge e-A&F system, which to the best of our knowledge is the first to contain all four key e-A&F interface elements (summaries of clinical performance, patient lists, detailed patient-level data, and suggested actions), and whose design was informed by existing relevant usability principles and theory. We used an innovative approach to usability inspection that combined the strengths of both HE and CW. This meant we could evaluate not only the individual interface components (as in an HE; Objective 2), but also how they should be integrated into a single system to support the Goal-Action structure of typical tasks (Objective 3). Combining these two methods maximised usability issue discovery, which was further helped by using eight evaluators working independently rather than the widely recommended three to five in traditional usability inspections [64,65]. The development and use of a thorough usability inspection protocol was important given the relative lack of knowledge in the wider literature regarding usability of e-A&F systems.

The main limitation of this study is that results are based on a usability inspection with expert evaluators, rather than testing with representative end-users (primary care clinicians) [34]. This may result in a number of the issues identified being false positives [70], or potentially missing important issues that may only be apparent with further tests in more naturalistic settings [34]. For example, end-users may not consider being able to edit an action plan marked “complete” a usability issue, as they may wish to retrospectively clarify events; conversely, although end-users may consider the display of actions recommended by PINGR usable, they may not find them useful in improving patient care. Furthermore, expert evaluators may be able to navigate the system more efficiently than target end-users. Specific actions were taken to minimise the impact of these limitations, such as the use of a background document describing the characteristics of intended users (Figure 4), and a thorough hybrid inspection method with multiple data collection and analysis stages, which included a usability issue consolidation stage to ensure the most important unique usability issues were analysed. Nevertheless, our findings should be interpreted with caution, and we consequently deem our usability recommendations for e-A&F systems as ‘preliminary’. To further address these shortcomings, and in accordance with accepted usability engineering methodology [34], we plan to undertake further evaluations of PINGR with target end-users in future.

Further limitations are that our hybrid usability inspection method mainly focused on identification of heuristic violations for each action, rather than the detailed reconstruction of a user’s cognitive goal as in traditional CW [66]. For example, a traditional CW may ask
the evaluator to estimate the percentage of users that will perform a specific action, the level of agreement between a given and actual user’s goal, or the likelihood that users’ goals will change after the performance of an action [66]. However, incorporating all these elements would likely make the hybrid method too cumbersome, which is a criticism of the traditional CW method [81].

Finally, we assessed inter-rater variation, which is recommended for usability inspection studies [82], and can therefore be viewed as a methodological strength. We found weak-moderate agreement and poor-fair reliability between evaluator’s issue severity ratings, which may reflect a useful variety of raters’ abilities to detect a wide variety of potential issues, though conversely may also be viewed as a limitation of our results. However, low levels of agreement are common in usability inspection studies [67], and it becomes increasingly hard to achieve agreement as the number of raters increases [71]. Usually three to five evaluators are recommended for usability inspection studies [64,65], though we used eight in order to detect as many unique issues as possible. To demonstrate, our IRA and IRR metrics improve if re-calculated using the ratings of only three instead of eight evaluators: complete agreement = 32%, Kendall’s coefficient = 0.68, ICC = 0.47, and Weighted Light’s Kappa = 0.46; Krippendorff’s alpha remains at 0.04, which is consistent with a comparable study [82]. Therefore, our relatively low levels of agreement and reliability are a necessary consequence of attempting to achieve our study objectives.

5.7 Conclusion

We have presented the usability evaluation of a modern, research-led clinical e-A&F system (PINGR) using a hybrid usability inspection method. In doing so, we described its design, rationale and theoretical basis (Objective 1), identified usability issues in relation to its four interface components (summaries of clinical performance, patient lists, patient-level data, and recommended actions; Objective 2), and attempted to understand how these issues may interfere with the cognitive goals of end-users of the system (Objective 3). Based on our findings and the wider literature, we have developed a set of recommendations for the user-centred design of e-A&F systems that addresses key interface components, in addition to how they should be integrated (Study objective 4). These recommendations go some way to addressing the gaps in the literature regarding the optimal design of clinical e-A&F systems. Future research should refine and extend this much needed evidence base.

5.8 References


[27] N. Ivers, J. Barnsley, R. Upshur, K. Tu, B. Shah, J. Grimshaw, et al., My approach to this job is ... one person at a time, Can Fam Physician. 60 (2014) 258–266.


[73] M. Gamer, Package “irr” v0.84: Various Coefficients of Interrater Reliability and Agreement, Hamburg, 2015.


Chapter 5 concluding note

This chapter addressed RO2 by using findings from Chapter 3, and emerging findings from Chapter 4 to develop PINGR (version 1). It also addressed RO1 and RO3 by undertaking a usability inspection study of PINGR’s interface, and using the results to not only improve PINGR (to create version 2), but also derive recommendations for the design of e-A&F interfaces in general. The main limitation of this study was that it used software experts rather than target end-users (primary care health professionals). The next chapter seeks to redress this by assessing the usability of PINGR version 2 by studying how General Practitioners interact with it in a lab-based evaluation.
Chapter 6

Multi-method Laboratory User Evaluation of an Actionable Clinical Performance Information System: Implications for Usability and Patient Safety

Benjamin Brown,1,2 Panos Balatsoukas,1 Richard Williams,1,2 Matthew Sperrin,1 Iain Buchan1,2

1 Health eResearch Centre, Farr Institute for Health Informatics Research, University of Manchester, UK
2 NIHR Patient Safety Translational Research Centre Greater Manchester, University of Manchester, Manchester, UK

Contributions
Author BB intellectually conceived the idea for the study, designed and built the software under evaluation, designed the study, collected and analysed the data, and wrote the manuscript. RW assisted with software development. PB and MS helped with study design and data analysis. All authors reviewed versions of the manuscript, and provided critical feedback on its development.

Publication status
A version of this paper has been submitted to the Journal of Biomedical Informatics.

Chapter introductory note
This chapter addresses Research Objective (RO) 2 by presenting the second version of the Performance Improvement plaN GeneratoR (PINGR). In addition to findings from Chapter 3, emerging findings from Chapter 4, its design has also been informed by results from the preceding Chapter 5 usability test with software experts. It also addresses RO1 and RO3 by evaluating its usability with primary care clinicians, and using the findings to refine the guidelines for the design of e-A&F systems introduced in the previous chapter.
6.1 Abstract

**Introduction:** Electronic audit and feedback (e-A&F) systems are used worldwide in care quality improvement. They measure health professionals’ performance against clinical guidelines, and some systems suggest improvement actions. However, little is known about optimal interface designs for e-A&F, in particular how to present suggested actions for improvement. We developed a novel theory-informed system for primary care (the Performance Improvement plaN Generator; PINGR) that covers the four principal interface components: clinical performance summaries; patient lists; detailed patient-level information; and suggested actions. As far as we are aware, this is the first report of an e-A&F system with all four interface components.

**Objectives:** 1) Use a combination of quantitative and qualitative methods to evaluate the usability of PINGR with target end-users; 2) Refine existing design recommendations for e-A&F systems; 3) Determine the implications of these recommendations for patient safety.

**Methods:** We recruited seven primary care physicians to perform seven tasks with PINGR, during which we measured on-screen behaviour and eye movements. Participants subsequently completed usability questionnaires, and were interviewed in-depth. Data were triangulated to derive usability issues.

**Results:** Participants committed a median of 10 errors (range 8-21) in using PINGR’s interface, and completed a median of five out of seven tasks (range 4-7). Errors violated six usability heuristics: clear response options; perceptual grouping and data relationships; representational formats; unambiguous description; visually distinct screens for confusable items; and workflow integration. Eye movement analysis revealed the integration of components largely supported effective user workflow, although the modular design of clinical performance summaries unnecessarily increased cognitive load. Interviews and questionnaires revealed PINGR is user-friendly, and that improved information prioritisation could promote useful user actions further.

**Conclusion:** Comparing our results with the wider usability literature we refine a previously published set of interface design recommendations for e-A&F. The implications for patient safety are significant with regard to: user engagement; actionability; and information prioritisation.
6.2 Introduction

Quality measurement is central to improvement strategies [1]. It identifies where action is needed and monitors the effects of improvement efforts [1]. In health care, this measurement is usually set in the context of ‘audit and feedback’ or ‘clinical performance feedback’, where compliance with clinical standards or patient outcomes is the common metric [2]. Clinical performance is primarily fed back as ‘quality indicators’, ‘performance measures’, or similar quantitative metrics [2]. Electronic audit and feedback (e-A&F) systems use interactive digital interfaces to communicate this information to health professionals such as intranet browser-based portals or desktop applications [3]. They are in use throughout the world, described variously as dashboards, benchmarking tools, scorecards etc [3].

Core to e-A&F systems is the presentation of quality indicators, which may be supplemented by the following additional components: patient lists, detailed patient-level information, and suggested actions [3]. Despite the potential importance of these components for actionable data interpretation [4], relatively little is known about designing usable interfaces for e-A&F to optimise user interaction and reduce errors during decision making [3]. In particular, existing evidence regarding e-A&F usability has been limited to systems without key interface components (e.g. suggested actions), and has largely ignored how interface design can affect user interaction when interpreting clinical performance data [3]. Evidence from the health informatics literature demonstrates that the design of information systems without regard for usability can increase technology-induced errors [5]. In the case of e-A&F systems such errors may have adverse consequences for patient safety by reducing the system’s effectiveness to improve health care outcomes [4]. Therefore poorly designed e-A&F interfaces may result in misinterpretation or ignorance of important information, which could ultimately lead to failings in care quality and efficiency (e.g. [6]).

We have previously reported a usability inspection evaluation of an e-A&F system for primary care – the Performance Improvement plaN GeneratoR; PINGR [3]. PINGR is currently unique among published e-A&F systems in that it possesses all key interface components: clinical performance summaries (i.e. quality indicators); patient lists; detailed patient-level information; and suggested actions [3]. Its design employs existing evidence and theory regarding effective A&F interventions, and is intended to be generic so it can host quality indicators from a range of clinical areas. Consequently, usability findings from PINGR provide valuable insights into how to best design interfaces for e-A&F systems, and the findings are likely to generalise to other settings such as secondary care. The results of PINGR’s usability inspection study enabled us to create a set of generic interface design recommendations for e-A&F systems, covering each of their interface...
components and how they can be integrated [3]. The study also represented the first step in an iterative approach to optimise PINGR prior to deployment in routine clinical practice [5,7]. The present study extends usability testing to target end-users (primary care clinicians) as planned in PINGR’s development framework [5]. We seek to understand how the interface helps or hinders user interaction across a range of information interpretation and decision-making scenarios in clinical quality improvement. The originality of this work lies in studying how primary care clinicians interact with e-A&F systems, using a laboratory-based multi-method usability evaluation.

Aim and objectives
The aim was to understand, through usability testing with end-users and theory-based abstraction, how the design of clinical e-A&F interfaces can facilitate improvements in patient safety.

The objectives were to:
1. Test the usability of PINGR in terms of efficiency, errors, satisfaction, and utility, using a multi-method approach, triangulating data from observations of on-screen and visual search behaviour during task performance, post-test user satisfaction questionnaires, and in-depth interviews.
2. Use these findings to extend and refine our previous set of interface design recommendations for e-A&F systems [3] in relation to their main interface components (clinical performance summaries; patient lists; detailed patient-level information; and suggested actions), whilst comparing them to the wider usability literature.
3. Determine the implications of these interface design recommendations for patient safety by drawing on evidence regarding clinical audit and feedback implementation.

6.3 Materials and methods
The evaluated system: PINGR
PINGR is an e-A&F system for primary care professionals, developed by the authors (Figure 1a and 1b), including a primary care physician/informatician (BB), a software engineer and informatician RW, and a human-computer interaction expert (PB). PINGR is a web-based application that runs outside clinical systems. It analyses structured data extracted from electronic health records (EHRs) on a nightly basis against clinical standards and patient outcomes (e.g. from guidelines).

PINGR’s present interface design was informed by a usability inspection study [3], and an emerging theoretical causal model of effective audit and feedback [8,9]. The use of theory
is recommended in the design of: complex interventions such as e-A&F [7]; and of audit and feedback tools [10]. Our approach is informed by an ongoing systematic meta-synthesis of qualitative studies [8], and draws on: existing theories (such as Control Theory [11] and Feedback Intervention Theory [12]); intervention description frameworks (e.g. [13]); and organisational implementation models (e.g. [14]). The remainder of this section presents a detailed account of the design and rationale of PINGR’s four main interface components.

**Clinical performance summaries**

The PINGR interface (Figure 1a and 1b) employs the overview-preview principle to display information at different levels of detail based on Shneiderman’s visual search mantra [15]. Presenting an overview of clinical performance data with details on demand was found to be an important usability feature in e-A&F systems [3]. The overview is provided as performance summaries at the level of the primary care practice/office (Figure 1a), where quality indicators are grouped into separate data representation modules for each clinical area. This module oriented design was employed to: enhance information processing, as is usual practice with clinical guidelines [16]; and facilitate user workflow [17]. Within each clinical area, quality indicators are further grouped into common care pathways associated with long-term care: diagnosis, monitoring and control [18], with an additional exclusions pathway to track patients excluded from the quality standards for clinical reasons (e.g. terminal illness). The purpose of the pathway groupings is to create a framework for representing a variety of clinical conditions consistently – as recommended in design best practice for EHRs [19] and clinical decision support (CDS) systems [20].

Currently, PINGR supports four clinical areas: hypertension, asthma, chronic kidney disease, and atrial fibrillation (AF). These clinical areas were chosen because they are:

1. managed mostly in primary care, making them familiar to end-users;
2. common components of multimorbidity – a major quality/safety issue in primary care [21] and core to the challenge of summarising patient information across multiple clinical areas [22], which is often not addressed by CDS systems [23];
3. often poorly cared for, resulting in serious adverse outcomes for patients and financial impacts on the health system (e.g. [24]), so address a quality improvement need;
4. associated with different quality indicators from different guidelines (e.g. process and outcome measures [18]; quality problems of overuse and underuse [25]; competing recommendations), so are suitable exemplars from which to generalise findings; and
5. covered by other commonly used existing e-A&F systems, enabling users to make comparisons with PINGR about its relative (dis)advantages.

As prescribed by cognitive fit theory [26], quality indicators are displayed as separate modules for tables and line graphs to support both symbolic and spatial information processing respectively. Both involve perceptual and analytical processes that are fundamental to the interpretation and understanding of quantitative health information [27]. In addition, providing users the option to select between tables and line charts is in accordance with mainstream usability heuristics (e.g. [28]), and effective audit and feedback data presentation theories [4,29]. Tables are presented by default, which users can switch to display time trends as line graphs. The rationale for displaying tables first rather than line graphs was that the addition of rows to tables would more easily facilitate the expansion of PINGR to include further quality indicators, whilst at the same time facilitating users to easily interpret their current clinical performance [26]. Although many e-A&F systems explicitly compare user’s quality indicator scores to targets/goals [3], evidence for their use in improving feedback effectiveness is mixed [2]. Therefore in accordance with actionable feedback design [30], PINGR non-judgmentally presents clinical performance data to users which they can compare with their past performance on line charts, and their internal beliefs regarding care quality [12].

Using a menu on the left side of the screen, users can display performance summaries for a specific clinical module at the overview level (e.g. AF; Figure 1a), or proceed directly to the preview level to view more detail by selecting a care pathway (i.e. diagnosis, control, treatment or exclusion; Figure 1b). Users can also access the preview level (Figure 1b) by selecting one of the pathway’s data representation modules in the overview interface (Figure 1a, blue coloured rectangular areas). At the preview level, information is presented regarding all patients who have not achieved the quality indicator and for whom improvement opportunities have been identified (Figure 1b, left side). For example, for hypertension control, these are patients whose latest blood pressure in the past 12 months is above 139/89 mmHg or above (or 149/89 mmHg if they are > 79 years old) based on national quality standards [31]. An interactive bar chart shows the improvement opportunities identified for patients (Figure 1b, left side above the list). By selecting a bar within the chart, the user can generate a list of all patients in whom the improvement opportunity is relevant, for example, all patients currently prescribed suboptimal medication, or those that may have medication adherence issues. Each improvement opportunity is accompanied by an explanation as to what it refers, and its eligibility criteria, communicated via short static notes at the top of the patient list and tooltips. The use of static notes and tooltips were found to be important for users to complete the goal-action sequence of data interpretation tasks in the context of e-A&F systems [3]. A user can
switch from patients who have \textit{not} achieved the quality indicator ("patients with improvement opportunities"), to those who have ("patients OK") using the corresponding tabs at the top of the bar chart. The user can also use a separate set of tabs to select different quality indicators relevant to the clinical pathway. For example, in hypertension control there are different blood pressure targets recommended by different organisations (e.g. [31,32]). For each generated list of patients, users can view detailed patient-level data by selecting a specific patient identification number (Figure 1b, right hand side). Patient lists, detailed patient-level information and suggested actions components of the PINGR interface are discussed in more detail below.

\textit{Patient lists}

As described above, patients achieving and not achieving each quality indicator are listed in the preview interface for each care pathway module (Figure 1b, left side). These lists can be ordered by patient-level clinical variables to enable users to prioritise high-risk patients, which may improve the effectiveness of e-A\&F [33]. For example, patients with improvement opportunities in their hypertension care can be ordered according to their last systolic blood pressure reading. In addition, following PINGR's usability inspection study [3] the current version includes additional variables including patients’ identification number and the number of quality indicators violated ("improvement opportunities"). As explained in section 2.1.2, the lists of patients not achieving a quality indicator can be filtered by clicking the "improvement opportunity" bar chart (Figure 1b, left side above the list), which displays the number of patients in relevant categories (see section 2.1.5 below for a more detailed explanation). This chart acts as an interactive visual query mechanism to list patients requiring similar improvement tasks, thus minimising user cognitive load by grouping together patients that require the same clinical action [34,35]. Finally, an "All patients" list presents all patients within PINGR across all quality indicators combined.

\textit{Detailed patient-level information}

Detailed patient-level information can be displayed adjacent to patient lists (Figure 1b, right side). Both patient lists and patient-level information are displayed concurrently to facilitate user's anticipated workflow of efficiently selecting new patients requiring improvement action [36]. Patients can be selected to display their information from the lists or via a search box at the top of the page. To improve system status visibility [28] as suggested from our usability inspection study, the patient-level information is separated from the patient list by a border, and when a new patient is selected a self-healing fade indicates their data is presented [3]. At the top of the patient-level information component a dropdown menu provides access to information relevant to each quality indicator. For example, selecting the blood pressure control indicator displays patient's blood pressure measurements, whereas selecting the atrial fibrillation anticoagulation monitoring displays
their International Normalised Ratio (INR) readings. As recommended in our usability
inspection study [3], these data are by default presented using interactive line charts to
help users assess patient readings over time (e.g. tooltips display details in the x and y
axis for each data point), and are contextualised with relevant additional non-clinical
details using tool-tips and vertical markers (e.g. recent medication changes or face-to-face
consultations). A toggle option is available to alternatively present these data as tables
[26]. Further clinical information, including the patient’s current medications and allergies,
is presented below the line charts to improve interpretation of data and suggested actions
for each quality indicator (Figure 1b). This design decision is also supported by research
showing that additional clinical information can improve clinician decision-making [37] and
user experience [20]. As data in the e-A&F system may differ from those in the EHR [36],
functionality is available for users to indicate whether or not PINGR has correctly identified
an improvement opportunity for a patient, and whether patient-level data is correct, using
agree (thumbs up) and disagree (thumbs down) icons.

In accordance with evidence from non-clinical dashboards [34] and CDS systems [20],
quality indicators listed in the dropdown menu are colour-coded and prioritised: clinical
areas in which the patient violates a quality indicator are presented first in red, those they
have achieved are second in green, and indicators that are not relevant to the patient but
are still within PINGR are at the bottom in grey. Colour is a reliable pre-attentive property
that facilitates quick identification of information without sequential searching, which can
reduce short-term memory load [38,39]. The use of colour was identified as an important
element for the unobstructive completion of tasks in the cognitive walkthrough evaluation
of an earlier version of PINGR [3]. The purpose of presenting data related to achieved and
irrelevant indicators is to enable users to highlight if PINGR incorrectly classifies a patient
(false negatives), in order to improve its algorithms [40] and support error prevention [20].

Suggested actions
The defining feature of PINGR is that it suggests actions that users could take, which we
call ‘decision-supported feedback’ [3,9]. PINGR provides two types of suggested actions
for improvement to users based on their specific clinical performance [41,42]:
organisational-level and patient-level. This is because: evidence suggests that both types
are required for effective improvement action [43]; health professionals have difficulty and
limited time to develop their own improvement actions [33]; and providing suggested
actions alongside feedback is shown to improve its effectiveness [2]. Organisational-level
suggested actions relate to steps that the primary care practice/office team could take
collectively to improve, such as introducing a new service or changing the way they work.
In the PINGR interface these are presented at the overview level, on the same page as
the clinical performance summaries showing quality indicators across the four pathways
(diagnosis, monitoring, control, and exclusions), and relate to suggestions for the whole clinical area (e.g. hypertension; Figure 1a, right side). Patient-level suggested actions relate to changes clinicians could make in caring for individual patients, such as introducing new medication, or providing lifestyle advice. They are presented alongside the detailed patient-level information component, with different suggested actions for each quality indicator accessed via the dropdown menu (Figure 1b, right side). Organisation and patient-level suggested actions are positioned to the right-hand side of the overview and preview interface respectively to match the anticipated user workflow of data interpretation and action according to both Control Theory [11], CDS design guidelines [44] and findings from our usability inspection study [3]. Furthermore, this complied with CDS design guidelines of providing relevant patient data alongside alerts [44,45].

We have previously published an early version of our methodology for deriving patient-level suggested actions [24]. In brief, this involves translating relevant clinical guidelines into rule-based algorithms to analyse the EHR data for each patient that has not achieved the quality indicator [24]. For example, in hypertension control one suggested action concerns medication optimisation: data are analysed to derive an up-to-date medication list, which is then compared with published maximal doses [46] and clinical pathways [31]. If a patient’s current medication dose is sub-maximal, then PINGR suggests increasing the dose. Similarly, if their medication does not match the prescribed clinical pathway, then PINGR suggests starting the most appropriate medication. The algorithms also take into account contextual information about patients [44], such as relevant comorbidities and allergies, by not suggesting a medication if the patient has a contraindicated condition (e.g. PINGR would not suggest a beta blocker in a patient with asthma). In this version of PINGR, organisational-level actions were derived from quality improvement actions in the wider literature and clinical guidelines (e.g. [47]).

To improve help and documentation [3,48], information buttons provide explanations for how suggested actions were generated. Hyperlinks to case reports of how other organisations had achieved change and other useful clinical tools (e.g. patient information leaflets) are also provided. These were designed to make the suggestions more actionable by providing further information on demand [44], whilst also drawing on the effects of Social Proof Theory [49]. Users can agree or disagree with PINGR’s suggested actions by clicking thumbs up or thumbs down icons respectively. When the thumbs up icon is clicked, the action is saved to a separate (“agreed actions”) page where it can be viewed, downloaded to share with colleagues, and marked as “implemented”. When the thumbs down icon is clicked users are asked why they disagreed with the action, using optional fixed and free-text responses [19,44]. As guided by CDS literature this is intended to communicate that the recommendations are advisory, in order to improve system
acceptability and potentially system effectiveness [44,50], in addition to collecting information on how PINGR’s algorithms could be improved [40]. Users can also add their own actions in addition to the ones suggested by PINGR, which is intended to provide flexibility user control and freedom [28], and also build a user-sourced bank of generated actions.

Additional functionality suggested by PINGR’s usability inspection study [3] included: use of consistent and concise statements to avoid misinterpretation (all suggested action statements were written by BB and pilot-tested with two additional clinicians); improved visibility of system status (e.g. by showing clearly when a specific action was agreed by turning green, disagreed by turning red and disappearing, or implemented by using strikethrough text); prevention of errors by disabling further editing of an action once marked implemented; supporting user control over actions that have been agreed, disagreed or implemented (including user-generated actions) by enabling undo/redo and edit capabilities; and presentation of all suggested actions in a consistent manner, using the same typographic features and layout characteristics.
Figure 1: The Performance Improvement plaN GeneratoR

1a. Overview level interface displaying clinical performance summaries (light blue border boxes at the centre of the screen, where each box represents a care pathway: diagnosis, monitoring, control and exclusions) and organisational-level suggested actions (light red border box, right hand side of the screen).
1b. Preview level interface displaying the improvement opportunities bar chart, patient lists, detailed patient-level data and suggested actions. The background colour of the detailed patient-level data interface component turns red when an improvement opportunity is present.

AF = atrial fibrillation; A/C = anticoagulation; BP = blood pressure; DASH = Dietary Approaches to Stop Hypertension; NICE = National Institute for Health and Care Excellence; OD = Once daily; QOF = Quality and Outcomes Framework.
**Participants and setting**

To evaluate PINGR’s usability we recruited a sample of primary care physicians (our intended end-user group) to interact with its interface whilst performing a set of tasks. We used purposeful sampling [51] to approach physicians that would typically be expected to use PINGR in the real world through professional networks of lead author BB. A request was made either by phone, email or face-to-face to participate in a study about the evaluation of a novel web-based e-A&F system aimed at improving the quality of primary care. Physicians were eligible if they regularly used: web applications on laptop or desktop computers; EHRs; clinical decision support systems; and e-A&F systems. Eligibility was determined using a short screening questionnaire (Appendix 9), which was sent via email along with an information sheet about the study. A good level of familiarity was determined in terms of number of years in practice (at least three years as primary care physicians), frequency of computer and internet use (at least five to 10 hours per week), and use of specialised health care software at work (at least half the days).

Participant recruitment was conducted concurrently with data collection and analysis. Our target sample size was between five to ten participants to balance costs and maximise usability issue discovery [52,53]. We stopped recruitment when thematic saturation was reached, which we defined as no new usability issues arising after two further participants [54]. Applying this criterion, seven physicians in total were approached and recruited to participate in the study (the sample’s characteristics are presented in section 3.1).

Testing took place at the usability laboratory of the School of Computer Science of the University of Manchester, and was conducted by author BB. At the beginning of each test participants were briefed about the study objectives and usability test protocol, then asked to sign a consent form. During briefing participants were given a short standardised description of PINGR’s functionality, though no demonstration or training was provided. Participants then completed two background questionnaires measuring their level of graphical literacy [55] and numeracy skills [56] as both characteristics could influence participants’ interaction with PINGR and therefore help understand any differences in user interaction. PINGR was accessed via Google Chrome internet browser on a desktop Windows computer and 17-inch screen. For information privacy reasons the version of PINGR used in the tests used only simulated patient data. Participants were offered reimbursement for their time (£50) plus associated travel costs. The study was approved by the UK National Research Ethics Service (Harrow; reference 15/LO/1394) and Greater Manchester Clinical Research Network (reference 187283).

**Tasks and task administration**
Participants completed 7 tasks using PINGR (Appendix 10) in a within-subjects design. As shown in Table 1, tasks were designed to assess participants' interaction with all interface components of PINGR using realistic actions users would perform with an e-A&F system based on existing literature [8]. Specifically, tasks reflected both behavioural and cognitive aspects of user interaction with the interface. To understand the effect of interface design on participants’ cognition, Tasks 1, 2, 3 and 5 required multiple perceptual and cognitive sub-tasks including data interpretation (at both the organisational and patient-levels), and judgment of the appropriateness of PINGR's suggested actions. Tasks 4, 6 and 7 were focused on exposing behavioural aspects of user interaction, such as locating specific information on the screen, entering data and creating and downloading user-generated actions. Tasks were presented in a randomised sequence (using the sample command in R [57]) to mitigate the effects of learning transfer, except for Tasks 6 and 7, which for logical reasons were always last. Each task was presented on-screen with contextual background information about a fictional primary care practice, and a patient as necessary, which participants used to inform their judgments during the tasks. To test the process of participants disagreeing with some of PINGR’s suggested actions and patient-level data, some were phrased to purposefully violate contraindications in the clinical background information (e.g. suggesting a medication to which the patient was allergic, offering a service to which they had explicitly declined, or presenting a wrong list of medications). To minimise participants acting unnaturally because they felt personally judged on how they performed using the software [58], it was made clear that it was PINGR (not they) who were under evaluation.
Table 1: Overview of tasks performed by participants

<table>
<thead>
<tr>
<th>#</th>
<th>Description</th>
<th>Interface components</th>
<th>Evaluated aspects of human cognition/behaviour</th>
</tr>
</thead>
</table>
| 1 | Interpret feedback and organisational-level actions across multiple quality indicators | • Clinical performance summary  
  • Suggested actions | • Identification and interpretation of relevant clinical performance summary  
  • Judgment of organisational-level suggested actions |
| 2 | Interpret feedback and patient-level actions regarding a single quality indicator | • Clinical performance summary  
  • Patient lists  
  • Detailed patient-level information  
  • Suggested actions | • Identification of relevant patient list  
  • Identification of appropriate patient from list  
  • Interpretation of detailed patient-level information (single disease)  
  • Judgment of patient-level suggested actions (single disease) |
| 3 | Interpret feedback and patient-level actions regarding an individual patient | • Detailed patient-level information  
  • Suggested actions | • Identification of relevant patient  
  • Interpretation of detailed patient-level information (multiple diseases)  
  • Judgment of patient-level suggested actions (multiple diseases) |
| 4 | Add a user-generated suggested action | • Clinical performance summary  
  • Suggested actions | • Identification of relevant suggested action area  
  • Data input |
| 5 | Identify the patient with the most improvement opportunities | • Patient lists  
  • Detailed patient-level information | • Identification of relevant patient list  
  • Identification of appropriate patient from list |
| 6 | Download saved actions | • Suggested actions | • Identification of saved actions download function |
| 7 | Indicate an action plan has been implemented | • Suggested actions | • Identification of implemented actions function |

Data collection

We measured usability in terms of efficiency (the time taken for participants to complete each task); errors (task completion rate, and the type and number of errors made); and user satisfaction with the interface design [59]. In addition, we used utility as a fourth outcome [59] based on the number of suggested actions agreed and disagreed with while performing the tasks, and participants’ responses during interviews. Data were collected using a multi-method approach, including observation of user on-screen and visual search behaviour, post-test satisfaction questionnaires, and in-depth debriefing interviews.

User observation

We used Tobii Pro Studio with a Tobii T60 eye tracker to record participants’ on-screen behaviour, eye movements, and time taken for completion of specific tasks. The Tobii T60 eye tracker permits a 60-Hz sampling rate, 0.5 degrees gaze point accuracy, and free
head motion, which was recalibrated before each task. Author BB observed concurrently using a second monitor and took field notes, which permitted identification of interesting aspects of user interaction that were discussed during debriefing interviews.

**Post-test questionnaires**
Following task completion, participants completed two usability questionnaires. We are unaware of any questionnaires specific to e-A&F systems, and therefore used the System Usability Scale (SUS) [60], and developed a questionnaire based on Shneiderman’s Object-Action Interface model (Appendix 11). The SUS is a validated questionnaire that measures user’s overall satisfaction with a system’s interface [60]. It is interface agnostic and consists of 10 items with total scores ranging between zero and 100 [61]. Our Object-Action Interface questionnaire consisted of two parts aimed at evaluating specific aspects of PINGR’s user interface design: the first contained seven items regarding the ease or difficulty participants experienced undertaking actions during tasks; the second contained eight items assessing the clarity of PINGR’s interface objects (e.g. presentation of data or use of colour and terminology). Both parts used a Likert scale from 1 to 5, with 1 representing difficult or unclear, and 5 indicating easy or clear.

**In-depth debriefing interviews**
Finally, participants were interviewed about their experience using PINGR after completing the questionnaires. Interviews were semi-structured (Box 1), and focused on the strengths, weaknesses, opportunities for improvement, and threats of using the software (SWOT). Questions explored concepts from Normalisation Process Theory (coherence, reflexive monitoring, cognitive participation, and collective action), which seeks to understand the work that people do, individually and collectively, surrounding a particular practice (e.g. using PINGR) rather than simply their beliefs or attitudes [62]. Other questions explored problems encountered during completion of tasks, negative responses to questions in the post-test questionnaires or other relevant additional topics that arose during interviews. As necessary, participants were replayed sections of their recorded on-screen interaction to clarify issues, and encouraged to further explore the PINGR interface. Interviews ended when both the interviewee and interviewer agreed all important topics had been covered. Interviews were audio-recorded and transcribed verbatim, and all participants were offered the option of reviewing their transcripts prior to analysis. Field notes were kept throughout the process.
Box 1: Interview schedule

Concepts from Normalisation Process Theory [62] addressed by each question explored in square brackets.

<table>
<thead>
<tr>
<th>Opening question: How did you find PINGR?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strengths</strong></td>
</tr>
<tr>
<td>• What are the advantages of PINGR? [Coherence]</td>
</tr>
<tr>
<td>• How useful or valuable do you think it would be in your primary care practice? [Cognitive participation]</td>
</tr>
<tr>
<td>• What, if anything, does it offer over existing systems you use? [Coherence]</td>
</tr>
<tr>
<td><strong>Weaknesses</strong></td>
</tr>
<tr>
<td>• What are the weaknesses of PINGR? [Coherence]</td>
</tr>
<tr>
<td>• What would be the disadvantages of using it in your practice? [Reflexive monitoring]</td>
</tr>
<tr>
<td><strong>Opportunities</strong></td>
</tr>
<tr>
<td>• How do you think you would use PINGR in your practice? [Collective action]</td>
</tr>
<tr>
<td>• How could it be improved in order to become a routine part of patient care processes? [Reflexive monitoring]</td>
</tr>
<tr>
<td>• How does PINGR differ from audit systems you currently use? [Coherence]</td>
</tr>
<tr>
<td><strong>Threats</strong></td>
</tr>
<tr>
<td>• What are the potential threats to PINGR not being used in practice? [Cognitive participation / collective action]</td>
</tr>
<tr>
<td>• What problems may arise with it being used? [Cognitive participation / collective action]</td>
</tr>
<tr>
<td>• How does PINGR align with the goals of your practice? [Coherence]</td>
</tr>
<tr>
<td><strong>Closing question: What have we missed that you think we should discuss?</strong></td>
</tr>
</tbody>
</table>


Data analysis
Data analysis was concurrent with data collection. This enabled exploration of important emerging concepts in interviews (theoretical sampling) [63], and to recognise when thematic saturation had been reached [54]. Data were triangulated from screen recordings, eye movements, questionnaire responses, interview transcriptions and field notes, in order to identify usability issues with PINGR in relation to its main interface components. This enabled us to confirm or disprove findings from a single source [64]. Figure 2 shows a summary of the data collection and analysis process with respect to the concepts they primarily measured (i.e. efficiency, errors, satisfaction, and utility). However, the triangulation process often used data sources to illuminate findings beyond these primary measures e.g. interview findings often provided insights into errors observed during user observation. To mitigate our results portraying an overly positive view of PINGR [65], our emerging analysis was critically reviewed by and agreed between our entire multidisciplinary research team (i.e. BB – primary care physician and health informatics researcher, PB – human-computer interaction expert, RW – software engineer, MS – statistician, and IB – public health informatician). This encouraged reflexivity, and increased credibility of our findings [66]. We used medians rather than means as our measure of central tendency given the small sample size and presence of outliers [67].

User observation
Videos of participants’ interaction with PINGR (i.e. on-screen activity and visual search behaviour) were imported into NVivo 11 (QSR International) for analysis with respect to efficiency, errors, and utility [59]. Efficiency was calculated as the time taken for participants to complete each task. Errors were defined as deviations of actual from expected user behaviour. A thematic content analysis [68] determined the number and type of errors by categorising them according to CDS system usability design heuristics [20], and the interface component to which they related. We calculated the total number of errors performed by users during each task, in addition to each tasks completion success rate. Utility was calculated as the number of suggested actions users agreed and disagreed with. Eye movement data in error-prone tasks were analysed in Tobii Pro Studio to understand the attention paid to areas of interest (AoSs). We defined six AoSs according to the key interface components of e-A&F systems; two at the overview level (Figure 1a; clinical performance summaries, and organisational-level suggested actions), and four at the preview level (Figure 1b; improvement opportunities bar chart, patient lists, detailed patient-level information, and patient-level suggested actions). We used heatmaps to visualise the number and duration of fixations on the interface, and collapsed fixation sequences to understand how participants transitioned between AoSs. Transition
matrices presented the probability of participants transitioning a fixation from one AoI to another [69]. Because the tasks used in this study included both reading and visual searching a fixation was defined as a stable gaze lasting at least 100ms [70]. Any fixation lasting less than 100ms was recorded as a saccade, i.e. a rapid eye movement between two fixations where no new information is processed by the participant [70]. When interpreted in conjunction with the efficiency, errors and utility data, heatmaps and transition matrices provided insights into participants’ workflow pattern and the appropriateness of how PINGR’s interface components were organised.

Post-test questionnaires
Data from post-test questionnaires were analysed in R [71]. Statistics included median, range, and upper and lower quartiles.

In-depth debriefing interviews
Interview transcripts and field notes kept during the interviews were imported into NVivo 11 (QSR International) for thematic content analysis. Data items were coded line-by-line by author BB to create a set of themes that explained user perceptions of the PINGR tool. These themes were organised into a framework based on the SWOT analysis at the highest level, with lower level codes relating to PINGR’s four interface components, NPT constructs, and usability heuristics [20]. The process was iterative in that each data item was reviewed multiple times to refine theme and code definitions. Findings were discussed with and critically reviewed by PB [66]; any disagreements were resolved through discussion.

Figure 2: Summary of the data collection and analysis process

NPT = Normalisation Process theory, SWOT = Strengths, Weaknesses, Opportunities, Threats
6.4 Results

Participants
Tests took place during September and October 2015, and took between 1.5 and 2 hours per participant. Each participant (two female, 5 male; age range 25–64 years) had between six and 33 years’ experience as a medical doctor, 3 and 28 years’ experience as a primary care physician, and 5 and 25 years’ experience undertaking audit and feedback. All participants used EHRs and CDS systems daily at work, felt at least 70% confident in their numeracy skills (e.g. using fractions, percentages, and graphs) [56], and scored at least 85% on the graphical literacy test [55]. All participants used e-A&F systems, though less often than they used EHRs and CDS systems: one participant used them “nearly every day”, with the rest using them “half the days” (n=3) and “less than half the days” (n=3). None of the participants had used PINGR previously, or had visual impairments that would affect the quality of eye movement recordings.

Efficiency
Figure 3 shows the distribution of time spent by participants on each task. Tasks 1-3 were the most time consuming, which could be explained because they were the most complex. Although Task 5 also required multiple perceptual and cognitive sub-tasks, these were limited to a single data variable (number of improvement opportunities). Conversely, Tasks 1-3 required interpretation and judgment of data relating to either organisational-level performance or patient-level clinical variables, both of which are multi-dimensional, in addition to their corresponding suggested actions. Task 2 had the highest median completion time overall (4.5 minutes), though one participant during Task 1 spent the longest time across all tasks (11.7 minutes).

Figure 3: Time taken to complete each task by participants

Note: Each cross represents one participant.
Errors
Participants committed a median of 10 errors (range 8-21) associated with PINGR’s key interface components during all tasks, and completed a median of 5 out of 7 tasks (range 4-7). Figure 4 shows the distribution of the number of errors made by participants during each task, which mirrors the time taken for each task in Figure 3. Tasks 1-3 were associated with the most errors, and although Tasks 2 and 3 had the joint highest median number of errors (n=4), Task 3 had the highest upper limit (n=7), and Task 1 had the widest range (0-6). This pattern is further reflected in the task completion rates, in which Task 3 had the lowest completion rate (1 participant completed), followed by Task 2 (2 participants), and Task 1 (5 participants); all participants completed tasks 4-7. The same participant who had the longest completion time in Task 1 was also responsible for conducting the highest number of errors during these tasks (6 and 8 respectively). Figure 5 shows which usability heuristic each error violated [20], and which interface component with which they were associated. Out of a possible 38 heuristic categories, six (16%) were violated [20]: Clear response options; Perceptual grouping and data relationships; Representational formats; Unambiguous description; Visually distinct screens for confusable items; and Workflow integration. The ‘Unambiguous description’ category was derived by combining two pre-existing heuristics categories: ‘Unambiguous units’ and ‘Concise and unambiguous language’. Detailed descriptions of the errors and their heuristic categories are provided below.

Figure 4: Number of errors made during each task by participants

Note: The size of the dot represents the number of participants who committed that number of errors.
Eye movement data analysis focused on Tasks 1-3 given they were the most time-consuming and error-prone. An example is illustrated in the heatmap for Task 2 in Figure 6, which is similar to the visual search behaviour observed for Task 3. Both tasks required identification of appropriate patient-level suggested actions at the preview page. During these tasks, participants would be expected to mainly attend to the patient lists, detailed patient-level information and patient-level suggested actions component, which should result in higher visual activity in the preview part of the interface. Yet the heatmaps demonstrated the opposite behaviour, with a greater number of fixations on the overview page (Figure 6). The transition matrix of eye movement sequences for Task 2 (Figure 7) showed high probabilities of transitions between AoIs compatible with optimal workflows for task completion. For example, 83% of transitions from the clinical performance summary were to the improvement opportunities bar chart, 75% of transitions from the improvement opportunities bar chart were to patient lists, 40% of transitions from patient lists were to detailed patient-level information, and 70% of transitions from here were to patient-level suggested actions. This indicates that participants were generally able to drill down to the relevant data for the Task. However, the matrix (Figure 7) also demonstrates unexpected transitions between AoIs at the overview level (Figure 1a): 17% of transitions from the clinical performance summary were to the organisational-level suggested actions, and 100% of transitions from the organisational-level suggested actions were to the clinical performance summary. When considered with the heatmaps, these findings suggest that although the integration of AoIs at the preview level (Figure 1b) largely supported effective user interaction, the overview level (Figure 1a) unnecessarily increased user’s cognitive load making difficult to navigate to the preview page and focus.
on the necessary information to complete these tasks. Typical errors at the overview page relevant to Tasks 2 and 3 included participants selecting the wrong clinical performance summary data representation module or care pathway, which often led to selecting the wrong patient list and ultimately the wrong patient. These errors violated the unambiguous description heuristic and accounted for 12 errors at the clinical performance summary and 11 at the patient list interface components. During interviews, most participants (n=5) explained this was because the module-oriented design of each clinical area, and the use of clinical pathways to organise quality indicators in PINGR made it difficult to prioritise which area to focus. Although users could view all quality indicators within a selected clinical area at the same time, they could not view indicators across different clinical areas. Therefore, judging which area required the most urgent attention required accessing each module individually and comparing performance across different pathways. This was exacerbated because scores were not explicitly compared to desirable levels of performance (targets/goals), so making value judgments regarding their performance required further information processing and could lead to errors in accessing the wrong module. Participants spontaneously suggested the presentation of benchmarking data (peer performance) as their preferred target/goal.

The most frequently violated heuristic was the workflow integration heuristic (n=40), all of whose errors were associated with the detailed patient-level information (n=24) and suggested actions (n=16) interface components. At the detailed patient-level information component, most participants (n=4) did not interact with the thumb icons to indicate whether PINGR had correctly identified an improvement opportunity for a patient, nor whether the patient-level data were correct despite the Task 2 description highlighting a discrepancy regarding the patient’s medical information in PINGR. During interviews participants explained this was because using the thumb icons in this way was unnecessary and time-consuming, since they would not have time to check these during their busy clinical work and would not expect the accuracy of patient-level data within PINGR to differ from the EHR. Suggested actions did not integrate with participants’ workflows as most (n=5) did not indicate agreement with the actions they added to the PINGR system using the thumb icons, which on questioning was because they felt should be automatic to save time – for although they may wish to disagree with an action they added to demonstrate it had been considered, the default would likely be agreement. Participants suggested the ability to manually order and prioritise actions in the “agreed actions” page may help with this issue and integrate PINGR within their existing workflows, with functionality to assign actions to other staff members with deadlines and reminders to track progress.
The third most frequently violated heuristic was clear response options (n=9), all of whose errors were associated with the suggested actions interface component – either when undoing a prior agree or disagree response, or when marking an action implemented. When undoing an agree or disagree response, participants expected clicking the opposing thumb icon, in addition to the same thumb icon, would also undo their prior agreement or disagreement response. In this situation, participants repeatedly clicked the opposing thumb icon to undo their prior response. When marking an action implemented, participants found it difficult to identify the button due to its small size and non-contrasting colour, which in one case resulted in non-completion of Task 6. Furthermore, when an action had been marked as implemented, some misinterpreted the self-healing fade to present the undo option as a slow system response, and the use of strikethrough text that the action had been deleted. Consequently, some participants suggested that requiring users to indicate an action had been implemented was superfluous, and it would reduce cognitive load if this could be automatically detected by the system.

Errors regarding perceptual grouping and data relationships (n=8) related to the dropdown menu on the detailed patient-level interface component. Participants (n=5) often did not use it to access the correct information related to different quality indicators, which subsequently impacted their ability to view all suggested actions for a patient. Four participants subsequently could not complete the task because they had not accessed the required patient-level data. During interviews, most participants (n=4) reported being unclear of the menu’s purpose (n=3), felt it prevented an overall impression of the patient’s clinical status in order to prioritise action (n=2), and it contained too many options to operate efficiently (n=4). The remaining errors for this heuristic category (n=3) related to the search box.

Errors violating the visually distinct screens for confusable items heuristic (n=6) were confined to the detailed patient-level interface component during Task 3 where participants selected a patient either via the search functionality or “all patients” list. In this situation, it was unclear a new patient had been displayed because the dropdown menu first had to be accessed to display information relevant to a specific quality indicator. This contrasts with where a patient was selected via a quality indicator-specific patient list, where the relevant information is displayed automatically. Interviews revealed this was exacerbated by concurrently displaying patient lists and detailed patient-level data on the same page, as it was confusing to display both population and patient level data at the same time.

Under the representational formats heuristic, one participant was unable to correctly interpret a patient’s latest INR reading on the detailed patient-level interface component.
During their interview, they revealed this was because it was difficult to visualise in the default line chart. A further participant felt the detailed patient-level information should be default presented as tables, as it was more difficult to identify one-off or low readings in a chart.
Figure 6: Visual attention heatmap for Task 2

Overview page is on the left, and preview page on the right. Red represents greater number of fixations whereas orange, yellow and green represent progressively less.
Figure 7: Transition matrix of eye movement sequences between areas of interest during Task 2

Number of visual transitions from one AoI to another as a proportion of transitions to all AoIs.

Note: AoI = Area of Interest
Satisfaction

The median SUS score was 73 (range 58-88), indicating a “passable” level of satisfaction with PINGR’s usability [61]. This was supported by interviews, where all participants (n=7) volunteered that PINGR was easy to use. Nevertheless, some felt a tutorial module would be helpful, particularly to highlight some of the novel features within PINGR such as the suggested actions. Despite the number of errors observed, Figure 8 shows that overall participants felt tasks were easy to complete: the lowest median Likert difficulty rating for any task was 3, and no participant gave the lowest score of 1. Task 2 was reported the most difficult (median Likert rating = 3, range = 3-4), with Tasks 1 and 3 joint second (median Likert ratings = 4, and ranges = 2-5), which mirrors the findings in the efficiency and errors sections (3.2 and 3.3. respectively) described above. Participants who committed the most errors in Tasks 1, 3 and 5, also rated them lowest.

Figure 8 also shows that participants felt in general that PINGR’s interface was clear, with only one participant giving the lowest Likert rating score of 1 to the font because it was felt too small. The patient lists interface component was considered the least clear (median Likert rating = 3, range = 2-5), which during interviews participants (n=3) explained that despite the addition of more relevant clinical variables to the lists from the previous version of PINGR, further information to effectively prioritise patients for action was needed. Suggested additional variables included patients’ age, as younger patients would likely gain most benefit from improvement actions; risk of a related outcome (e.g. cardiovascular disease event in the blood pressure control quality indicators); and highlighting patients who had violated a different particularly high risk quality indicator (e.g. inappropriately untreated AF). Three participants commented that the number of patients in each list was overwhelming, and that in clinical practice they would likely not have the time or resources to go through all of them. They stated they would attempt to deal with this by sharing the list with other staff in their practice, for example by printing the list and asking a nurse to go through it on their behalf. Further problems with the patient list was confusion over the use of red flag icons to denote improvement opportunities, because in clinical medicine this usually refers to important clinical signs. In this case, they suggested using a different colour, or alternative icon such as a lightbulb, star or lightening flash.

Despite being associated with the most errors during tasks, the detailed patient-level interface component was rated the most clear by participants (median Likert rating = 5, range 2-5), with all but one participant giving it the highest score of 5. During interviews, participants (n=3) made positive comments about the benefits of visualising patient’s physiological data as line charts as it facilitated interpreting the clinical significance of data, and prioritisation of patients for action.
Both the clinical performance summaries and suggested actions interface components received a median Likert rating of 4. With respect to the clinical performance summaries, participants stated that framing clinical performance in a positive rather than negative fashion was preferable (e.g. number of correctly treated AF patients, rather than the number incorrectly treated) because they felt “rewarded” for their efforts, which consequently encouraged further action. However, they suggested presenting the clinical performance trend data concurrently with the current performance information as it helped aid their data interpretation and prioritisation, though was not clear how it could be accessed via the toggle buttons. Similar suggestions were also received regarding the improvement opportunities bar chart.

With regards to suggested actions, all participants stated they were acceptable because they were presented as suggestions and could express disagreement. The majority (n=4) expressed concern they would not have the time or resources to evaluate and implement every suggested action, and would only be able focus on the most important. They felt PINGR’s design made it difficult to prioritise its suggested actions because of its modular clinical performance summaries format that prevented viewing organisational-level actions across multiple quality indicators, and dropdown menu in the detailed patient-level information component dividing the patient-level suggested actions across different clinical areas. Participants stated they wanted to be presented with only the top three or four most important improvement actions they could implement at any one time and felt that ideally PINGR should do this automatically. Views regarding what they considered important criteria varied and included clinical safety (e.g. high-risk drug-drug interactions requiring urgent attention), potential effectiveness (e.g. predicted impacts on patient outcomes), financial value (e.g. how much money the organisation could save from unnecessary tests, avoiding adverse outcomes, or by aligning with pay-for-performance schemes), and quick wins (i.e. the perceived ratio of implementation effort to potential benefit – either clinical or financial). Almost all participants stated the position of suggested actions on the right of the page was satisfactory as it mirrored their workflow of data interpretation, then formulation of subsequent action. However, one participant could not complete a task because they did not visualise any of the suggested actions. In the same vein, participants recommended moving the “agreed actions” page (where their agreed actions were saved) to the end of the navigation menu to fit with their workflow sequence of data interpretation, action formulation, and action plan review. Other recommendations for improving the suggested actions included using less prose and only providing detail on demand; though conversely also providing specific reasoning as to why each action was suggested – despite the inclusion of information buttons, which were not visible, or detailed enough.
Figure 8: Participant responses to the Object-Action Interface questionnaire
Actions are on the left; Objects on the right. The size of the dot represents the number of participants who committed that number of errors.
Utility

During tasks, each participant viewed a median of 12 suggested actions (range 5-13), of which they agreed with a median of 7 (range 4-8), disagreed with a median of 3 (range 0-6), and did not respond to a median of 0 (range 0-8). Reasons for disagreeing with organisational-level suggested actions were because they had already been implemented in the participant’s organisation (n=4), participants perceived lack of resources to implement the suggestion (n=1), or they were felt to be an inefficient use of resources (n=1). Reasons for disagreeing with patient-level suggestions related to user disagreement with clinical guidelines to avoid over-medicalisation and unnecessary use of resources (n=3), an absence of local services to carry out the action (n=2), a lack of further patient-level information on medication adherence to judge the actions’ appropriateness (n=1), and a perceived need to clinically assess the patient first (n=1). Two participants did not respond to suggested actions because they wanted to defer a decision following either discussion with colleagues (regarding organisational-level actions) or individual patients (for patient-level actions), or carrying out other actions first. They suggested the ability to manually order and prioritise actions in the “agreed actions” page may help with this issue and integrate PINGR within their existing workflows, with functionality to assign actions to other staff members with deadlines and reminders to track progress. Worryingly, two participants agreed with patient-level actions for which the patient had clear contraindications (i.e. patient refusal of a home blood pressure monitoring service, and prescription of a medication to which there was an allergy); on further questioning, both felt they would not expect the system to suggest actions to which there were documented reasons against. Of their own accord, three participants added 10 actions to PINGR they had formulated themselves. These covered organisational-level actions (n=2), such as services or quality improvement ideas not currently suggested by PINGR, and patient-level (n=8) actions relating to lifestyle advice and medication safety issues. A summary of findings relating to PINGR’s utility from the NPT-driven SWOT analysis interviews is presented in Box 2.
Box 2: Summary of NPT-driven SWOT analysis interviews related to PINGR's utility

Strengths

Clinical performance summaries
- Covers multiple clinical areas and quality indicators relevant to primary care.
- Includes quality indicators relating to undiagnosed patients, over-treated patients (e.g. over-anticoagulation in AF), and patients who may benefit from exclusion (e.g. palliative care patients), thus addressing issues of quality of over-medicalisation.
- ‘Improvement opportunity’ chart provides unique insights into reasons for poor performance, guided improvement action, and can save time by filtering patient lists to those requiring the similar actions.

Patient lists
- Lists of patients requiring action facilitates quality improvement.

Suggested actions
- Shifts focus from data interpretation to improvement action.
- Has potential to save time by negating the need for users to formulate their own action plans.
- Functionality to save, download and mark actions as implemented helps track those yet to be implemented, enables communication with other staff, and could be used as evidence for annual appraisals.
- Links to case reports (organisational-level) and patient information leaflets (patient-level) can aid implementation.
- User-added actions are a good way to share best practice between organisations.

Detailed patient-level information
- Ability to “drill-down” from population-level data via interactive links is intuitive and user-friendly.
- Provides non-clinical in additional to clinical data (e.g. contacts with the primary care practice), which are useful for contextualising improvement action.

Weaknesses

Clinical performance summaries
- Inclusion of quality indicators with differing guidance can be confusing (e.g. different blood pressure targets [31,32]).
- ‘Improvement opportunity’ chart functionality could be unclear and categories too numerous for action.

Opportunities

Clinical performance summaries
- Addition of quality indicators in areas important to primary care e.g. chronic obstructive pulmonary disease, diabetes, and general health checks.
- Tailoring of which quality indicators could be displayed based on user-preference.

Detailed patient-level information
- Inclusion of further information including demographics, medication adherence or prescription frequency, historical medication prescriptions with dates and reasons for cessation, details on improvement opportunity categories, and additional relevant physiological parameters and comorbidities.

Suggested actions
- Integration with existing information systems – e.g. write-in functionality to EHRs to facilitate care documentation, direct communication with patients via text messages/letters/emails, medication prescribing, ability to access individual patients’ records in EHRs and the e-A&F system directly from each other.
- The ability to view other users’ agreed actions within their organisation to aid action planning and prevent duplication of work.
- Inclusion of patient decision aids into suggested actions where appropriate (e.g. regarding specific recommended treatments).
- Alignment with local clinical pathways.
- Link organisational-level suggested actions to specific reasons for under-performance.
- Present actions not previously considered by users.
- Patient-level actions may be valuable for nurses conducting chronic disease clinics.
**Threats**

*Detailed patient-level information*

- Should not be too detailed because:
  - Users will only want to view data relevant to the quality indicators
  - Users could believe it replicates the EHR, which may give false reassurance they are viewing a complete medical history, leading to safety issues (e.g. if all currently prescribed medications or laboratory tests are not displayed).

**Suggested actions**

- Some users may formulate their own actions and ignore the suggestions.

Note: EHR = Electronic health record

---

### 6.5 Discussion

This study identified usability issues with a novel actionable e-A&F system for primary care, regarding efficiency, errors, satisfaction, and utility. The main strength of this work was to address all four components of e-A&F interfaces (clinical performance summaries, patient lists, detailed patient-level data, and suggested actions). The main limitation was the restricted number of clinical scenarios and clinicians that was practical to include.

In the following discussion we combine our findings with wider usability literature in order to refine a previously published set of preliminary interface design recommendations for e-A&F systems [3] (Box 3). We also discuss the implications of these findings for patient safety by drawing on findings from an ongoing systematic meta-synthesis of qualitative research studies of A&F interventions performed by our research group [8]. Finally, we discuss the limitations of our research methodology in more detail.

**Refined interface design recommendations for e-A&F systems**

*Clinical performance summaries*

Clinical performance summaries should cover multiple clinical topics relevant to users. Where possible they should address issues of over-treatment, missed diagnoses, and situations where it may be inappropriate to treat patients (e.g. those receiving palliative care). To align with users’ workflows they should offer the functionality to include quality indicators addressed in different quality programmes (e.g. [31,32] in PINGR’s case), and for users to select those in which they are most interested. Where appropriate, clinical performance should be framed positively rather than negatively (e.g. by presenting the number of patients attaining a quality standard rather than the number not attaining). This is supported by studies of presenting quantitative information to clinicians [63,64], and may leverage principles of positive psychology making users feel recognised for their performance efforts, and creating a positive loop of further action and interaction with the system [65].
Quality indicator results from all clinical domains should be presented concurrently in one display screen. This enhances information processing by making data comparison efficient, and is supported by studies of health-related dashboards [66,67]. Data prioritisation can be further helped through the presentation of clinical performance trends over time, and by explicitly comparing users’ scores with desirable levels of performance (targets/goals). Both elements should be presented simultaneously for each quality indicator to increase the chances of visualisation by users; the use of space-saving tools such as sparklines [68] may help. Although users may prefer peer performance data as the target/goal by which to judge their performance (e.g. average peer performance), other options for choosing targets/goals exist (e.g. those set externally by experts), all of which have potential advantages and disadvantages. Systems may further reduce cognitive load by automatically prioritising quality indicators on behalf of users and communicating this via colour (e.g. Red Amber Green [RAG] rating) or the order in which they are presented [34]. Criteria for prioritisation may include current levels of performance or predicted numbers of patient adverse outcomes [40,44].

Like PINGR’s “improvement opportunity” charts, e-A&F systems should automatically undertake further analysis and visualisation of clinical performance. This is supported by findings from evaluations of other e-A&F systems [66,69], and may include patients not achieving quality standards grouped according to similar actions (as in PINGR), or other patient or organisational variables. Where possible, these visualisations should highlight relevant patients for action via the patient list interface component. To facilitate cognitive processing, these analyses should be presented concurrently with overall quality indicator results, and provide a limited number of analysis options (e.g. those focusing on the largest groups of patients). They should include clear explanations as to what the analyses refer, with instructions as to how they may be used to facilitate action planning. Such detailed tutorial instructions have also been found necessary in similar population-level data exploration tools [70].

Patient lists

Patient lists should present sufficient information to enable users to efficiently prioritise patients for action or review, which may include patients’ age, physiological measures relevant to the quality indicator, the number of quality indicators the patient has violated and whether they have violated a particularly high-risk quality indicator (e.g. untreated high risk AF), and where relevant, their predicted risk of an adverse outcome (e.g. cardiovascular disease event). Given the many potential variables that could be included in patient lists, and variation in user preference demonstrated in our study, it may be appropriate for users to maintain freedom and control by customising which variables are displayed [28]. The ability to order and filter patient lists is essential for prioritisation, and
the availability of this function should be made clear. User control over patient list variables and ordering is supported by wider EHR design guidelines [71]. To avoid the volume of patients in lists from overwhelming users, e-A&F systems may display a manageable number of patients at any one time (e.g. 10) starting with the highest priority, with further patients displayed on-demand. This conflicts with EHR design guidelines that states that all patients in a list must be visible on one page [71], though is acceptable because EHR systems are not primarily intended for quality improvement purposes unlike e-A&F systems where this may be less important. Icons used in patient lists should be appropriate to the clinical context in which they are used. For example, radiological systems should avoid using red dot icons, which are also used to highlight abnormal findings on a radiological image [72]. This agrees with more general usability guidelines regarding the appropriate use of icons [73], though may only be recognised as problematic through user testing.

**Detailed patient-level information**

Detailed patient-level information should be accessible via interactive links to “drill-down” from population-level aggregated data via patient lists [15]. It should be as comprehensive as possible to provide users with sufficient information to interpret suggested actions, formulate their own, and in situations where the e-A&F system is stand-alone, avoid inefficiencies of accessing other information sources (e.g. EHRs). Similar recommendations have been made with primary care population-level data exploration tools [74]. As a minimum, patient-level information should include demographics, their diagnoses and physiological measures, and current and past medications (including data on adherence and reasons for stopping medications). Both clinical (e.g. physiological measurements, medication prescriptions) and non-clinical (e.g. contacts with the primary care practice) information should be presented to illustrate how patients interact with the health system, which can in turn facilitate action implementation. However, only data directly relevant to taking action should be presented in order to aid information prioritisation. Line charts should be used to display patients' physiological data where possible, to facilitate data interpretation and patient prioritisation. This is supported by evaluations of patient-level dashboards [66], and graphical representations of patient-reported outcome measures (PROMs) [75]. To accommodate user control and freedom [28], and maintain appropriate representational formats [20], users should also have the option to display data as tables. This is supported by EHR design guidelines [76], and studies of PROM [77] and laboratory data [78] graphs. To facilitate perceptual grouping and relationships between data [20], patient-level information should be displayed on a single page. This in turn helps prioritise which clinical areas require the most urgent attention, and reduce cognitive overload [19]. Visualisation techniques exist that help efficiently summarise patient-level data on one page [79] e.g. LifeLines display data as
multiple charts over a common timeline [80]. To provide visually distinct screens for confusable items [20], such detailed patient-level information should be completely separate from other interface components. Although this may intuitively interfere with users’ anticipated workflows, it is supported by EHR design guidelines [71]. A warning should be provided that the system is not attempting to replicate the patient’s EHR nor provide a comprehensive medical history, to ensure users are not given a false impression of its completeness. Users should also have the ability to validate the accuracy of EHR data within the e-A&F system, though to fit their workflow [20] this should only require them to highlight inconsistencies.

**Suggested actions**

Suggested actions may be derived from clinical guidelines, and the wider quality improvement literature. Suggestions should align with local clinical pathways, and address specific reasons for user’s under-performance based on detailed analysis of their clinical performance data [41,42]. They should strive to present ideas users may not have previously considered, and where possible adapt to contextual features of organisations (e.g. whether actions have already been implemented, or resource availability) and individual patients (e.g. potential contraindications or individual preferences) [44]. This mirrors CDS design guidelines that specify patient-level alerts should incorporate contextual data into decision logic to improve specificity [44]. Suggested actions should be written concisely e.g. using bullet points, with further detail on-demand regarding why it was suggested (i.e. algorithm logic) [44,49], reports of how other organisations achieved change (case reports), and patient-facing information such as leaflets or decision aids. They should account for the first step a user may take during implementation such as consulting a patient, or discussing an organisational change with colleagues. This could be achieved through the phrasing of the suggested action text (e.g. using goal-setting theory [81]), encouraging users to add their own actions to reflect this workflow, or providing an option to defer the decision about an action [45,82]. Suggested actions should be located to align with user workflow, which may be influenced by whether they are presented on the left or right of the page [68,83]. They should be prioritised, displaying only the most important three or four options across all quality indicators concurrently, with the option to view more if desired. Prioritisation criteria for suggested actions in e-A&F systems may vary between users, and should be accommodated with user preference settings where possible. Specific examples include patient safety issues [40,44], potential effectiveness, financial value, and quick wins. Prioritisation may be communicated through ordering or use of colour [20]. To improve acceptance, suggested actions should be advisory [44,50], and allow users to indicate disagreement, which in turn should be captured and used to improve the system’s algorithms [40,44]. This may be re-enforced by a warning to users that suggestions do not over-rule clinical judgment. Users should be
able to add their own actions, which should be saved automatically to integrate with their workflows [20], in addition to also being used to optimise the system’s own suggestions. Users should be able to save suggested actions, mark them implemented, and view other users’ saved actions to facilitate intra-organisational teamwork. Saved actions should be located at the end of the navigation menu to align with user workflows [40]. It should be possible to order and manually prioritise saved actions, with the facility to set deadlines and reminders, assign them to users, and export them for wider sharing. Response options to suggested actions should be clear [20]; e.g. undoing (dis)agreements should consider how users may respond to thumb up/down icons, and action implementation may be effectively communicated with a simple checkbox. Ideally, to reduce cognitive load e-A&F systems should automatically detect when an action has been implemented. Where possible, suggested actions should provide functionality to easily action the recommendation, which may be facilitated through integration with existing health information systems [44] including write-in functionality to EHRs, or direct communication with patients.

Box 3: Summary of interface design recommendations for electronic audit and feedback systems, and questions for further research, refined from [3]

Clinical performance summaries, should:
• Cover multiple clinical topics relevant to users.
• Address over-treatment, missed diagnoses, and situations where it may be inappropriate to treat patients (e.g. those receiving palliative care).
• Allow users to select which quality indicators to display.
• Be framed positively where appropriate to emphasise achievement e.g. patients achieving a quality standard, rather than those not achieving.
• Be presented across all clinical domains concurrently in one display screen.
• Use line graphs to demonstrate trends over time with tooltips to interpret historic performance data.
• Compare user’s scores to desirable levels of performance (targets/goals).
• Automatically prioritise quality indicators e.g. through the use of colour or ordering.
• Undertake further data analysis and visualisation related to improvement action.
• Explain clearly to what performance data specifically refer.

Patient lists, should:
• Present sufficient information to efficiently prioritise patients e.g. age, physiological measures, number of quality indicators violated etc.
• Allow users to control what information is used to prioritise patients.
• Clearly allow users to order and filter patients.
• Display a limited number of high-priority patients (e.g. 10), with more on-demand.
• Use appropriate icons to communicate patient variables.
• Explain clearly to what the lists refer.

Detailed patient-level information, should:
• Be accessible via interactive links to “drill-down” from population-level data.
• Be as comprehensive as possible.
• As a minimum include demographics, diagnoses, physiological measures, medications.
• Include both clinical and non-clinical data (e.g. contacts with the primary care practice).
• Only include information directly relevant to taking improvement action.
• Use line charts to display physiological data with tooltips to interpret historic performance data.
• Provide users the option to display data as tables.
• Be displayed on a single page.
• Be completely separate from other interface components.
• Provide a warning the system is not attempting to replicate a patient’s health record.
• Enable users to highlight inconsistencies with health record data.

Suggested actions, should:
• Address both the individual patient and organisation.
• Be derived from acceptable sources e.g. clinical guidelines or the wider quality improvement literature.
• Align with local clinical pathways.
• Address specific reasons for user’s poor performance based on detailed analysis of clinical performance data.
• Adapt to contextual features of organisations (e.g. whether or not actions have already been implemented, or resource availability) and individual patients (e.g. potential contraindications or individual preferences).
• Strive to present ideas users may not have previously considered.
• Be written concisely e.g. using bullet points.
• Provide details on-demand regarding on why it was suggested, case reports, and patient-facing information.
• Account for the first step a user may take during implementation.
• Be located in a separate interface component aligned with user workflow e.g. on the same screen as clinical performance summaries, or detailed patient-level information.
• Display only the most important three or four options at a time, though provide the option to view more if desired.
• Have prioritisation criteria that may include patient safety issues, potential effectiveness, financial value, and ‘quick wins’, and should be accommodated through user settings where possible.
• Have prioritisation that may be communicated through action ordering or colour.
• Be advisory, by allowing users to indicate disagreements (using both fixed and free-text responses), and indicating this via a warning message.
• Use data from user disagreements to improve its algorithms.
• Allow users to add their own actions, which should be saved automatically and used to optimise the system’s own suggestions.
• Allow users to clearly save, mark them implemented, and view those of others within their organisation.
• Provide functionality to view, undo and edit previous reasons given for disagreements.
• Allow users to order and manually prioritise saved actions, set deadlines and reminders, assign them to users, and export for wider sharing.
• Have clear response options, and ideally automatically detect when an action has been implemented.
• Provide functionality to easily action the recommendation, which may be facilitated through integration with existing health information systems e.g. write-in functionality or direct patient communication.

Questions for further research:
• How much interface adaptation should be user-controlled or automated? What methods can be used to optimise automated interface adaptation?
• What methods are most appropriate to adapt suggested actions to contextual features of organisations and patients?
• Which are the most effective types of targets/goals to use in clinical performance summaries?
• What additional methods for deriving suggested actions are possible, acceptable, efficient, and effective?
• What are the most appropriate criteria and methods to prioritise clinical performance summaries, patient lists, and suggested actions? What are the optimal ways to communicate and display this prioritisation?
• What is the optimal position of suggested actions within the user interface?
• How do findings from this study translate to more naturalistic settings outside the laboratory?
Implications for patient safety

Considering our e-A&F design recommendations in the context of literature on causal pathways in A&F interventions [8] we can frame their implications on patient safety through three main concepts: user engagement, actionability, and information prioritisation.

User engagement may be influenced through system compatibility, relative advantage, and satisfaction. Compatibility may refer to the clinical areas addressed by the system; existing systems, policies and workflows with which they align; and user preferences. Ensuring e-A&F systems address clinical areas users deem important and relevant, and that align with existing quality or financial incentive programmes, is essential to ensuring compatibility with their goals and motivation [84]. Similarly, e-A&F systems should align with user workflows, and integrate with existing information systems. Compatibility with user preferences can be improved by enabling user-directed customisation and tailoring [19,28]. Incompatible e-A&F systems might not be used by health professionals (e.g. [85]), or they will dismiss its feedback as trivial (e.g. [86]). PINGR’s relative advantages relate to its provision of detailed patient-level information, suggested actions for improvement, and user-friendliness. Ways in which other e-A&F systems can provide relative advantages depends on the individual system and the environment into which it is implemented [14]. If a system has a relative advantage, it is more likely to be used and implemented (e.g. [84]). User satisfaction with e-A&F systems can be influenced by its efficiency and tendency to induce user errors. e-A&F that are more satisfying and positive to use encourage further engagement (e.g. [87]). Where there is non-engagement with an e-A&F system – whether due to incompatibility, relative disadvantage, or dissatisfaction – potentially important clinical performance information is ignored, which could lead to failures in patient safety (e.g. [6]), whereas continued engagement generally leads to improved patient care [2].

Users must take action based on the information from e-A&F systems in order for improvements in patient care and safety to take place [30]. However, health professionals often do not have the time or skills to translate clinical performance information into improvement action [33]. Therefore making the information in e-A&F systems as actionable as possible increases this likelihood [2]. Our design recommendations suggest this can be achieved by providing additional clinical performance data analysis and visualisations, suggested actions, patient lists, and detailed patient-level data. Clinical performance data analyses (such as the improvement opportunities charts in PINGR) can help users understand potential reasons for low performance, and identify specific patients requiring action [87]. Suggested actions can help users formulate plans for improvement [88]. This is especially true for organisational-level actions, which may have
the greatest effect on patient care but that clinicians often struggle with most [33,89]. Optimising suggested action algorithms through user feedback shares best practice between organisations [88] and harnesses positive deviance [90]. Patient lists and detailed patient-level data direct the user’s attention to patients requiring action or further investigation [36]. Not providing patient lists makes it difficult for users to understand who to target or how to improve [33]. Similarly, an absence of detailed patient-level information may mean users fail to take action [89], and warnings that highlight the system does not provide a comprehensive medical history should reduce the risk of patient safety events (e.g. a limited medication list may risk drug-drug interactions if new ones are prescribed).

Our design recommendations highlight the importance of information prioritisation in all key e-A&F interface components. Health professionals have limited time to work on quality improvement due to various competing demands from both clinical and non-clinical responsibilities [91]. If their attention is not directed to the most important feedback information, the most appropriate improvement action will likely not occur [92]. For example, focusing on a quality indicator that is not the worst performing may not result in improvement action or the greatest population health gain [93]; reviewing a patient violating multiple quality standards would be more effective than one that violates only one [33]; being unable to view the most important areas of a patient’s information may miss additional, more important areas requiring attention [89]; and implementing an evidence-based action may be more effective than one that is not [84]. An additional benefit of information prioritisation is to prevent users feeling overwhelmed and disillusioned, similar to the phenomenon of alert fatigue in CDS systems [94], in which users may reject the feedback or abandon improvement work [11]. Nevertheless, prioritisation techniques should be used with caution as they may have unintended consequences e.g. using average peer performance as a target may not comprehensively raise standards.

Study limitations
The team who designed and built the system conducted this evaluation, the lead investigator of which held a position recognisable to study participants as a primary care physician. This posed risks to the trustworthiness of our findings including how participants behaved [58], our interpretation of this behaviour [95], and our degree of positivity in communicating our results [96]. We took specific steps to address these potential problems (section 2.6 above), and in doing so believe we present a balanced and detailed critique of PINGR. On reflection, author BB found that his position may have afforded him insider status [97], providing a potential advantage by gaining more honest insights from participants than a non-medically qualified researcher could elicit. The study’s small sample size (n=7 participants) may be perceived as a further weakness,
though was guided by the achievement of thematic saturation of usability issues [54]. This implies that the cost of including further participants would have been unnecessary [52,53]. Interview transcripts were coded by one researcher, which some may view as a threat to credibility [98]. However, we took explicit steps to mitigate this by triangulating findings from multiple data sources [99], in addition to conducting critical analytic discussions between authors to challenge any potential biases or assumptions [100]. Finally, our interface design recommendations for e-A&F systems (Box 3) have been derived from empirical studies of one system (PINGR). Although PINGR’s design has been informed by existing usability [3] and theoretical [8,9] evidence relevant to e-A&F, and its evaluations contextualised in the wider usability literature, effective alternative e-A&F designs may exist. Given the paucity of evidence on e-A&F usability, our recommendations are a reasonable starting point though should continue to be tested and refined in future.

6.6 Conclusion

This paper provides important insights into how to design usable e-A&F systems according to their four key interface components (clinical performance summaries, patient lists, detailed patient-level information, and suggested actions) [3], and how the components can be integrated. Although our study focused on primary care and long-term conditions, many of our findings may be relevant to other clinical areas and settings. We used a combination of quantitative and qualitative methods to evaluate the usability of a novel e-A&F system with target end-users (Objective 1) to refine our previous set of design recommendations [3] (Objective 2), and determined their implications for patient safety (Objective 3). As far as we are aware, this is the first published study of an e-A&F system to use eye tracking, and therefore presents unique insights into the visual search behaviour. We previously identified research questions for e-A&F system design [3], which have been partly answered by this paper including how to provide information in patient lists, summarise detailed patient-level data across multiple clinical areas, and incorporate clinical performance of other users. This study raises further questions (Box 3), for example regarding how to most effectively prioritise and communicate clinical performance summaries, patient lists, and suggested actions – a grand challenge of CDS [22] – and how these findings translate to more naturalistic settings. Our future work will seek to address these questions in further phases of our iterative development framework [5].

6.7 References


[33] N. Ivers, J. Barnsley, R. Upshur, K. Tu, B. Shah, J. Grimshaw, et al., My approach to this job is ... one person at a time, Can Fam Physician. 60 (2014) 258–266.


Chapter 6 concluding note

This chapter evaluated the second version of PINGR with primary care clinicians in a usability lab. In doing so, it refined the e-A&F design guidelines I introduced based on the results from Chapter 5. One criticism of this study design is that it is not a real-world setting, and may miss problems when PINGR is used in clinical practice. To address this shortcoming, the next chapter presents the third version of PINGR that has been improved based on the findings in this chapter. It implements it into a number of GP practices, and assesses its utility, acceptability, and barriers and facilitators to its use.
Chapter 7

Implementing Actionable Clinical Feedback in UK Primary Care: a Longitudinal Optimisation Study Using Clinical Performance Feedback Intervention Theory

Benjamin Brown,1,2 Thomas Blakeman,2 Matthew Sperrin,1 Richard Williams,1,2 Iain Buchan1,2

1 Health eResearch Centre, Farr Institute for Health Informatics Research, University of Manchester, UK
2 NIHR Patient Safety Translational Research Centre Greater Manchester, University of Manchester, Manchester, UK

Contributions
Author BB intellectually conceived the idea for the study, designed and built the software under evaluation, designed the study, collected and analysed the data, and wrote the manuscript. RW assisted with software development and data analysis. All other authors helped with study design and data analysis. All authors reviewed versions of the manuscript, and provided critical feedback on its development.

Publication status
A version of this paper is being prepared for submission to Implementation Science.

Chapter introductory note
This chapter addresses Research Objective (RO) 2 by using the model derived from Chapter 4 (CP-FIT), and the usability design guidelines for electronic Audit and Feedback (e-A&F) systems derived in Chapter 6, to design the third version of the Performance Improvement plaN GeneratoR (PINGR). It also addresses RO3 by implementing PINGR into clinical practice and evaluating how it is used to derive wider learning. CP-FIT is used to frame the results and translate findings to e-A&F systems in general.
7.1 Abstract

Introduction: Electronic health record (EHR) data are often used to quantify the quality of care provided to patients in electronic audit and feedback (e-A&F) systems. These systems vary widely in their effectiveness at improving patient care, and there is a lack of clarity as to the mechanisms by which they work. We developed Clinical Performance Feedback Intervention Theory (CP-FIT) to address this gap, and used it to inform the design of a novel e-A&F system for UK primary care that suggests improvement actions to users (the Performance Improvement plaN GeneratoR [PINGR]).

Objectives: 1) Understand how health professionals and primary care practices respond to, and interact with, actionable e-A&F; 2) Determine facilitating and inhibiting factors to explain findings from objectives 1.

Methods: We recruited 15 GP practices in one city in the North of England, to use PINGR between November 2016 and June 2017 inclusive. Quantitative data were collected remotely on how health professionals used the system, and its potential effects on patient care. Analyses included visual inspections, descriptive statistics, process mining, and comparisons with patients not viewed in the software. Qualitative data were collected from semi-structured interviews guided by Normalisation Process Theory and field notes. Framework Analysis was used informed by CP-FIT.

Results: Forty-eight users (11 GPs, 6 nurses, 3 pharmacists, and 5 non-clinicians) participated for a median of 4 months (range 1–8). Thirty-eight were interviewed at baseline, with twenty-five interviewed a second time. PINGR was used on 227 separate occasions. GP practices adopted the system to different extents and in different ways, though largely used it to take patient-level, rather than the organisational-level action. Patients viewed in the system were 1.6 (95% CI 1.5-1.7) times more likely to improve in at least one quality indicator in comparison to patients not viewed. Barriers and facilitators to its success included the resources available to use it; its perceived advantages of user friendliness, ability to suggest actions, and educational elements; how compatible it was with pre-existing beliefs and ways of working; the credibility of its data; the complexity of the clinical problems it highlighted; and the ability to act based on its recommendations.

Conclusion: CP-FIT has good applicability and explanatory power for e-A&F. Our results add to it by providing examples on how: particular phenomena may occur and inter-relate in practice, additional concepts that lie outside the model but align with its propositions, and how novel A&F designs may address some of its recommendations. Future research
should continue to investigate novel ways to deliver e-A&F and test CP-FIT’s assumptions.
7.2 Introduction

Recent years have seen exponential growth in the availability of patients’ electronic health record (EHR) data for secondary purposes [1]. One key use typically involves quantifying patients who have received (sub)optimal care, reporting results to health professionals as ‘performance measures’ or ‘quality indicators’ with the intention of improving care quality [1]. Tools are often developed to automate these processes, which may be called ‘dashboards’ (e.g. [2]) or ‘surveillance systems’ (e.g. [3]), though are direct descendants of traditional audit and feedback (A&F) [4]. Similar to traditional A&F interventions, these ‘electronic’ A&F (e-A&F) systems vary widely in their effectiveness at improving patient care, with a lack of clarity as to the mechanisms by which they work [4,5]. Experts have long called for the use of behaviour change theory to inform their design and implementation [6], however, this rarely happens, and may be partly explained by a lack of consensus regarding which may be most appropriate [5,7].

To address this need, we developed Clinical Performance Feedback Intervention Theory (CP-FIT) – a theoretical model of causal pathways in A&F derived from a systematic search and metasynthesis of findings from 65 qualitative studies (Figure 1) [8]. In brief, CP-FIT has three core propositions: 1) A&F interventions exert their effects by inducing action in health professionals; 2) Health care organisations have limited capacity to engage with and respond to A&F interventions; and 3) Health care professionals and organisations have a strong set of beliefs and behaviours regarding how they provide patient care that influence their interactions with A&F. It posits that effective A&F is a cyclical process of Goal setting, Audit, Feedback, recipient Interaction, Perception, and Acceptance of the feedback, followed by Intention, Behaviour and improved Clinical performance. Progress round this cycle is influenced by moderating variables relating to characteristics of the Goal, Audit methods, Feedback message, Implementation process, Organisational context, Co-interventions, Health professional, and Patient population. These moderators exert their effects via mediators relating to Actionability, Resource match, Complexity, Relative advantage, Compatibility, Credibility, and Social influence. Unintended consequences of Gaming and Tunnel vision may also occur. CP-FIT suggests that A&F interventions have two key mechanisms by which they exert effects: facilitating improvement action directly (Direct Action) and increasing knowledge or awareness of a particular quality issue (Knowledge/Awareness). It states that A&F interventions that maximise their effects via the Direct Action are most effective. Based on the above, CP-FIT makes a number of specific recommendations regarding how A&F interventions could be designed and implemented to be most successful.

In the systematic literature search that we used to develop CP-FIT [8], only 10 (15%) included papers studied e-A&F systems, two of which were in primary care.
Consequently, there is an evidence-gap with regard to how these systems are used, and what factors may facilitate their adoption and impacts on patient care. Therefore, we developed the Performance Improvement plaN GeneratoR (PINGR) – a novel e-A&F system for use in UK primary care. PINGR’s design has been informed by CP-FIT [8], the unique feature of which is to align with its main propositions by suggesting tailored improvement actions to health professionals regarding individual patients and their organisation. PINGR’s usability has previously been evaluated and optimised in simulated settings as part of a multi-stage iterative development process [9,10]. The next phase of its development was to implement and test in real-world clinical settings [11,12], which is the focus of this paper.

Aims and objectives
The aims of this study were to: 1) Optimise PINGR to increase its likelihood of adoption prior to its wider implementation and evaluation; 2) Derive wider learning for e-A&F systems in general; and 3) Empirically test and refine CP-FIT. In doing so, the study had two objectives: 1) Understand how health professionals and primary care practices respond to, and interact with, actionable e-A&F; and 2) Determine facilitating and inhibiting factors to explain findings from objective 1.
Figure 1: Clinical Performance Feedback Intervention Theory

Notes: Boxes – Blue boxes = moderating variables, orange boxes = mediating variables, Green boxes = processes and outcomes. Arrows – Unbroken = essential event sequence, dotted = non-essential.
7.3 Materials and methods

Approach
We followed guidance on the evaluation, and optimisation of complex interventions [12,13], and used mixed-methods to address all research objectives. We followed reporting [14] and quality [15] guidelines for qualitative research, evaluations of health informatics applications [16], and intervention description [17]. The study was approved by the UK National Research Ethics Service (Harrow; reference 15/LO/1394) and Greater Manchester Clinical Research Network (reference 187283).

2.2 Participants and setting
The study was set in a city in North West England served by 48 primary care practices, and was conducted between November 2016 and June 2017 inclusive. Fifteen practices highlighted by the local Clinical Commissioning Group (CCG) were targeted for recruitment, each with a history of trialling health care innovations as is usual when trialling new software [18]. Lead contacts at each practice (either a General Practitioner [GP] or practice manager) were sent a study information sheet via e-mail by lead author BB (a GP and health informatician). All 15 practices agreed to participate, none of whom were previously known to the research team. Practices volunteered staff for the study according to whose job it was most relevant. All proposed users provided written consent, and practices were offered £90 per member of staff interviewed as part of the study. In the year prior to the study, two different e-A&F systems were introduced to practices that served as comparisons to PINGR. The first was introduced by the CCG to all 48 practices to facilitate the implementation of a local pay-for-performance scheme. It was integrated into the EHR and practices were required to use it in order to participate in the pay-for-performance. The second system was an e-A&F system developed as part of a research project focusing on medication safety, which was used mainly by pharmacists and used in a limited number of practices. Both systems differed from PINGR in that they did not suggest actions for users to take.

Intervention: the Performance Improvement plaN Generator
PINGR [9,10] is an e-A&F system for UK primary care developed by authors BB and RW (software engineer), based on principles in CP-FIT [8] (Table 1). It is accessed as a web site and uses coded EHR data to calculate quality indicators. The user interface has three main areas, which are intended to be accessed sequentially to facilitate users taking direct action for individual patients and their wider organisation (i.e. PINGR’s ‘Direct Action’ mechanism; Figure 2):

- Overview (Figure 2a): Practice current and past performance on quality indicators across all clinical areas, including fixed and relative (based on peer performance) targets against which to aim. Indicators are ordered by descending performance.
• Indicator (Figure 2b): Practice performance on a specific indicator, in addition to sortable lists of patients not achieving the indicator. Detail on past performance (Trend) and comparison to other practices’ current performance (Benchmarking) is also provided.

• Patient (Figure 2c): Detailed information on a specific patient regarding their diagnoses, physiological measures, and medications relevant to the quality indicators. All information is on one page rather than separated according to quality indicator.

Additional pages include: Save actions, which collates the agreed practice improvement actions in response to quality indicator performance; All patients, which lists all patients in a practice prioritised by the number of quality indicators they violate; Search patients, where users can search for individual patients; and Help and Contact, where guidance on how to use the system is provided.

PINGR aimed to induce health professional behaviour change by suggesting tailored actions to users – i.e. ‘decision-supported feedback’ [19]. These relate to how care may be organised in the practice (organisational-level) or provided to specific patients (patient-level). Actions are generated by comparing the EHR data of patients not achieving quality indicators to National Institute for Health and Care Excellence (NICE; www.nice.org.uk) guidelines. For example, a typical quality indicator for patients with hypertension is whether they have a recent blood pressure measurement below a target level [20]. In this example, PINGR initially performs patient-level EHR analyses to determine whether specific actions could be taken such as optimising their medication or re-measuring their blood pressure at home (Improvement opportunities; Figure 2b) [21]. These are then aggregated to derive organisational-level actions, for example a high proportion of patients prescribed sub-optimal medication may indicate clinical staff are unaware of NICE treatment guidelines, so an educational session or laminated copy of the guidelines in each treatment room may help [21]. Organisational-level actions appear on the Overview and Indicator pages, whereas patient-level actions appear solely on the Patient page. Only three actions are displayed at a time to avoid cognitive overload [10], and are prioritised by how many patients to which they relate. They are presented as advisory using thumb up/down icons to improve acceptance [22,23]. If users click ‘thumbs up’ the action is saved to the practice’s Action plan page and is viewable by all PINGR users in the practice, if they click ‘thumbs down’ they can provide a free-text response to explain why. Users can also add their own actions, which are automatically saved. When a suggestion is made, hyperlinks to the relevant clinical guidelines are provided to support the rationale and enabled users to read the evidence further.
PINGR is a generic e-A&F system that can analyse any type of coded EHR data, to derive quality indicators from any clinical topic. Consequently, to increase the likelihood of its adoption, in accordance with CP-FIT’s notion of Compatibility and Relative advantage [8] the indicators used in this study were determined by the CCG based on which they deemed most important and worst performing (Table 2). These were informed by NICE guidelines, and some of which replicated those in the local pay-for-performance scheme running from 1\textsuperscript{st} April to 31\textsuperscript{st} March. The CCG’s priorities changed over time, so additional indicators were added as the study progressed: they initially started with chronic kidney disease, then grew to include other chronic diseases. Furthermore, the CCG had an EHR data repository that includes coded data from secondary care. Therefore to align with their existing systems, PINGR analysed EHR data from both primary care practices and the hospital.

Members of staff in practices identified to use PINGR were provided with unique log-in details and a brief (30-60 minutes) standardised demonstration of how it could be used by author BB. They were subsequently advised to use PINGR however and as much (or as little) as they wished. Users received weekly emails generated by PINGR that highlighted five patients that were violating the most quality indicators at their practice in an attempt to focus action on patients that most urgently required attention and would increase the most quality indicator scores.

Table 1: How Clinical Performance Feedback Intervention Theory (CP-FIT) informed the design and implementation of PINGR

<table>
<thead>
<tr>
<th>CP-FIT recommendation</th>
<th>Relevant PINGR design feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus on clinical areas that are:</td>
<td>How these aspects were addressed:</td>
</tr>
<tr>
<td>• Within recipients’ control.</td>
<td>• Focused on clinical topics routinely cared for in primary care guided by the CCG where topics are believed to be poorly cared for.</td>
</tr>
<tr>
<td>• Considered important and relevant.</td>
<td></td>
</tr>
<tr>
<td>• Poorly cared for.</td>
<td></td>
</tr>
<tr>
<td>Conduct the audit:</td>
<td>How these aspects were addressed:</td>
</tr>
<tr>
<td>• Without requiring the recipient to collect or analyse the data.</td>
<td>• Data were analysed automatically from EHRs, which are routinely used in the UK to measure clinical performance.</td>
</tr>
<tr>
<td>• In an automated way using appropriate source data and analysis methods.</td>
<td>• Quality indicator definitions were informed by existing guidelines.</td>
</tr>
<tr>
<td>• Allowing recipients’ to exclude patients they feel are inappropriate to be included in the measurement of their clinical performance.</td>
<td>• Health professionals could record standard reasons for exclusion in the EHR to remove patients from the quality indicators.</td>
</tr>
<tr>
<td>Produce a feedback message that:</td>
<td>How these aspects were addressed:</td>
</tr>
<tr>
<td>• Includes lists of patients used to calculate the recipients’ clinical performance.</td>
<td>• Included lists of patients.</td>
</tr>
<tr>
<td>• Provides recipients’ individual clinical performance.</td>
<td>• Data was updated daily.</td>
</tr>
<tr>
<td>• Is sent as close to the time of the clinical performance measured in the audit as possible.</td>
<td>• Prioritisation was used to order quality indicators on Overview page in terms of performance; suggested actions by number of patients and points to which they referred; All patients list by the number of quality indicators violated.</td>
</tr>
<tr>
<td>• Prioritises the relative importance of its contents.</td>
<td>• Been through two rounds of usability testing.</td>
</tr>
<tr>
<td>• Has been tested to ensure it is user-</td>
<td>• Includes reminder emails.</td>
</tr>
</tbody>
</table>
friendly; provides historic in addition to current clinical performance.
- Are ‘pushed’ to recipients rather than requiring them to request access.
- Convinces the recipient that its purpose is to support them improve care rather than punish them.
- Appears to come from a source with an appropriate degree of technical or clinical knowledge.
- Compares recipients’ performance to other health professionals.
- Is delivered to groups of health professionals in a team or organisation rather than just one.

| Focuses on quality improvement rather than performance management. |
| Developed by a primary care clinician. |
| Includes Benchmarking. |
| Allows all practice staff to access it. |

How these aspects were not addressed:
- Could not provide recipients’ individual clinical performance because such data was not available.

Implementing A&F in a way that:
- Gains support of senior managers.
- Fits with the existing workflows.
- Minimises time, human or financial costs.
- Demonstrates its value or benefits to the recipient.
- Can be tailored to the needs of the health care organisation.
- Targets health professionals with quality improvement skills, or provides support and training to recipients regarding how to engage with the intervention.
- Targets health professionals with adequate knowledge regarding the clinical topic of the A&F, or improves the recipients’ knowledge of the evidence and theory of the clinical topics focused on by the A&F intervention.
- Makes recipients feel like it is not imposed upon them.

| Implemented with the support of the CCG and senior members of staff in each practice. |
| Focused on clinical topics relevant to primary care chosen by the CCG, and included the ability to download lists of patients. |
| Support and training was provided on how to use software. |
| Clinicians were mainly targeted to use the software, and the software included links to evidence to improve clinical knowledge. |
| Users were advised the software was solely for quality improvement and could use it as much or as little as they liked. |

How these aspects were addressed:
- As not integrated with existing systems used by practices.

Providing additional support to:
- Help recipients interpret and formulate action plans.
- Facilitate recipients to discuss their clinical performance with peers.
- Resource the intervention e.g. protected time.
- Improve the ability of recipients to communicate with and work towards a common goal with their colleagues.
- Help recipients’ organisations communicate with external organisations.
- Address negative attitudes towards A&F.

| Suggested actions were provided for individual patients and the wider practice. |
| Actions could be saved, and viewed by other users in the practice. They could also be added by users and downloaded to share with others. |
| Additional resources were not provided. |
| Communication with external organisations were not facilitated. |
| Negative attitudes to A&F were not addressed. |

How these aspects were not addressed:
- Additional resources were not provided.
- Communication with external organisations were not facilitated.
- Negative attitudes to A&F were not addressed.

Key: CCG = clinical commissioning group; EHR = electronic health record.
Figure 2a: PINGR Overview page (simulated data)
Figure 2b: PINGR Indicator page (simulated data)
Figure 2c: PINGR Patient page (simulated data)
Table 2: Quality indicators included in PINGR (ordered by dates introduced)

<table>
<thead>
<tr>
<th>Indicator name</th>
<th>Description</th>
<th>In local pay-for-performance scheme?</th>
<th>Date introduced</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKD monitoring</td>
<td>Patients with CKD (Stage 3 and above) who have received the correct frequency of eGFR monitoring based on their latest eGFR and ACR readings.</td>
<td>No</td>
<td>November 2016</td>
</tr>
<tr>
<td>CKD diagnosis</td>
<td>Patients with evidence of CKD (stage 3 or above) based on their eGFR and ACR readings with a diagnosis in their records.</td>
<td>No</td>
<td>November 2016</td>
</tr>
<tr>
<td>CKD stage coding</td>
<td>Patients with CKD who have the correct coding of their CKD stage based on their latest eGFR and ACR readings.</td>
<td>No</td>
<td>November 2016</td>
</tr>
<tr>
<td>CKD and proteinuria BP control*</td>
<td>Patients on the CKD register with ACR 70mg/mmol or more with a BP recorded in the last 6 months since 1st October or 1st April where the latest BP is &lt;130/80 mmHg.</td>
<td>Yes</td>
<td>January 2017</td>
</tr>
<tr>
<td>Hypertension BP control*</td>
<td>Patients on the hypertension register with a BP recorded in the last 12 months since 1st April where the latest BP is &lt;140/90 mmHg for patients under 80 years and &lt;150/90 mmHg in patients 80 years or older.</td>
<td>Yes</td>
<td>January 2017</td>
</tr>
<tr>
<td>CKD and DM BP control*</td>
<td>Patients on both the CKD and diabetes registers with a BP recorded in the last 6 months since 1st October or 1st April where the latest BP is &lt;130/80 mmHg.</td>
<td>Yes</td>
<td>January 2017</td>
</tr>
<tr>
<td>CKD BP control*</td>
<td>Patients on the CKD register with a BP recorded in the last 6 months since 1st October or 1st April where the latest BP is &lt;140/90 mmHg.</td>
<td>Yes</td>
<td>January 2017</td>
</tr>
<tr>
<td>Hypertension casefinding</td>
<td>Patients with persistently raised blood pressure that are on the hypertension register.</td>
<td>No</td>
<td>February 2017</td>
</tr>
<tr>
<td>COPD exacerbations and PR*</td>
<td>Patients with COPD identified as MRC 2 in last 5 years with an exacerbation (coded or uncoded) recorded after 1st April who have been offered or declined Pulmonary Rehabilitation within 2 months of their latest exacerbation.</td>
<td>Yes</td>
<td>February 2017</td>
</tr>
<tr>
<td>AF casefinding acute*</td>
<td>Proportion of patients aged 55 years and over who present with one or more of the following: shortness of breath, palpitations, chest pain, syncope, dizziness, stroke, TIA or heart failure since 1st April, and have had a pulse rhythm assessment afterwards</td>
<td>No</td>
<td>May 2017</td>
</tr>
<tr>
<td>AF casefinding chronic*</td>
<td>The proportion of patients aged 65 years and over diagnosed with one or more of the following conditions: hypertension, diabetes, CKD, PAD, stroke or COPD, and have had a pulse rhythm assessment since 1st April.</td>
<td>No</td>
<td>May 2017</td>
</tr>
</tbody>
</table>

Key: AF = atrial fibrillation; BP = blood pressure; CKD = chronic kidney disease; COPD = chronic obstructive pulmonary disease; DM = diabetes mellitus; EHR = electronic health record.

*Indicator scores reset on 1st April
Data collection

Quantitative data

Quantitative data were collected remotely from PINGR regarding records of its use, in addition to how practices and their individual patients performed on the quality indicators. Every time a user accessed PINGR, data were captured on the pages visited, patients viewed, and whether they agreed or disagreed with any actions or added their own. The data PINGR contained regarding individual patients and practices across all 48 practices in the CCG was saved as a daily snapshot.

Qualitative data

Qualitative data were collected from semi-structured interviews conducted in person, observations of meetings, and reasons for disagreeing with suggested actions entered into PINGR by users. PINGR users in GP practices were interviewed by author BB during the demonstration of how to use the software. A second round of interviews was conducted at least two months later in order to allow for adequate system use and insight into sustainability. During this round at least one frequent and one less frequent user were purposively sampled at each practice (based PINGR’s usage records), to provide insight into factors affecting its use [24]. The first interview focused on how the participant and their practice currently conducted A&F (if at all) and their initial reactions to PINGR, whereas the follow-up interview focused on their experiences of using PINGR in practice. All interviews explored facilitators and inhibitors to A&F in general and PINGR, how it compared to other e-A&F systems, and how processes could be improved through asking questions guided by Normalisation Process Theory (NPT) [25] to explore working practices described in CP-FIT [8] (Appendix 12). NPT seeks to understand the work that people do, individually and collectively, surrounding a particular practice (e.g. implementing PINGR), rather than simply their beliefs or attitudes [26]. Additional questions probed topics arising from ongoing data analysis, such as important emerging concepts from interviews, or relevant findings from PINGR usage records [27]. A smaller number of managers and leaders at the CCG were also sampled for interviews and demonstrations of PINGR to provide a broader context in which to interpret findings from individual practices [24]. These were supplemented by observer-as-participant observations of CCG policy and committee meetings by BB where field notes were taken [28]. Interviews were generally conducted one-to-one in a private room, though if preferred by the practice, were conducted as a group. During interviews PINGR was accessed to enable both interviewer and interviewee to demonstrate issues when appropriate. Interviews lasted between 30 and 120 minutes, and stopped after both parties agreed all important topics had been covered. All were audio recorded, transcribed verbatim, and supplemented with field notes. Participants were offered the opportunity to
check their transcripts [29]. Data collection continued until saturation was felt to be achieved [30].

**Data analysis**

As recommended in the evaluation of complex interventions [13] data analysis was iterative and complementary in that qualitative analyses attempted to explain quantitative findings, and quantitative analyses tested hypotheses generated by qualitative findings. Analyses were concurrent with data collection in order for emerging findings to be tested in future interviews [13]. Both types of data analysis focused on comparing and contrasting differences between GP practices, users, and quality indicators in a case study approach [31], as these are sources of potential variation of A&F effectiveness in CP-FIT [8].

Quantitative data analysis

PINGR’s usage records were initially assessed using visualisations and descriptive statistics in R software (version 3.4.1) [32]. To assess how use between practices and quality indicators differed, aspects of PINGR’s usage records were mapped to concepts in CP-FIT [8]: user logging into PINGR (Interaction); specific indicator or patient viewed (Perception); suggested action disagreed with, or suggestion made regarding an individual patient (Non-acceptance); user agreements with PINGR-generated actions or user-added actions (Intention); patients changing improvement opportunity category (Behaviour); and patients achieving quality indicators they had previously violated (Clinical performance improvement). We determined relative risk ratios of patients viewed in PINGR experiencing Behaviour or Clinical performance improvement versus those not viewed in intervention practices. Usage records were also imported into Disco (version 1.9.9; Fluxicon) [33] and analysed using a Fuzzy mining algorithm [34] to understand the different sequence of pages users accessed.

Qualitative data analysis

All qualitative data (transcriptions and field notes) were loaded into Nvivo (version 10; QSR International) to aid triangulation [35]. Each item was coded line-by-line by author BB using Framework Analysis [36] according to concepts from CP-FIT [8]. Inspired by Realistic Evaluation [37], the focus of the analysis was to understand what factors influenced outcomes expected during A&F, and by which mechanisms. Consequently, each passage was coded to include A&F processes, moderating and mediating variables from CP-FIT where possible [8]. Concepts from CP-FIT were refined and new ones added as necessary. Author TB (a GP and qualitative researcher, not involved with PINGR’s development) independently read each interview transcript, and met author BB on a monthly basis to critically discuss and challenge the emergent analysis. In particular, TB’s
sought to ensure the analysis included sufficient reflexivity regarding BB’s potential influence on participants as a GP and creator of PINGR. The final analysis was then discussed and agreed with all authors.

7.4 Results

**Practice and participant characteristics**

Table 3 details characteristics of GP practices and study participants. We recruited 48 users (16 GPs, 6 nurses, 5 pharmacists and 21 non-clinicians) in 15 practices who participated in the study for a median of 4 months (range 1–8). Thirty-eight were interviewed at baseline (15 GPs, 7 nurses, 5 pharmacists, and 11 non-clinicians), and 25 (11 GPs, 6 nurses, 3 pharmacists, and 5 non-clinicians) were interviewed a second time a median of 3 months later (range 2-7) before data saturation was reached [30]. Three practices (Over Peak, Grand Oak, and Swan River) were not interviewed a second time because they had not used PINGR for more than two months, and two practices withdrew from the study (Hope Garden, and King’s Way). At the CCG, six members of staff were interviewed once (1 GP and 5 non-clinicians), and five hours of observations conducted.

**Table 3: GP practice (pseudonyms) and study participant characteristics (ordered by recruitment date)**

<table>
<thead>
<tr>
<th>Practice</th>
<th>Approximate list size (1000 patients)</th>
<th>Deprivation</th>
<th>Recruitment date</th>
<th>PINGR users (n)</th>
<th>Repeat interview (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oak Dale</td>
<td>4</td>
<td>Middle</td>
<td>November 2016</td>
<td>GP (1)</td>
<td>GP (1)</td>
</tr>
<tr>
<td>Golden Valley</td>
<td>22</td>
<td>High</td>
<td>November 2016</td>
<td>GP (1) Nurse (1)</td>
<td>GP (1) Nurse (1)</td>
</tr>
<tr>
<td>King’s Way</td>
<td>2</td>
<td>Middle</td>
<td>November 2016</td>
<td>GP (1)</td>
<td></td>
</tr>
<tr>
<td>Hope Garden</td>
<td>14</td>
<td>Low</td>
<td>January 2017</td>
<td>GP (1) Non-clinician (1)</td>
<td></td>
</tr>
<tr>
<td>Kindred Medical</td>
<td>18</td>
<td>High</td>
<td>February 2017</td>
<td>GP (1) Nurse (2)</td>
<td>Pharmacist (2)</td>
</tr>
<tr>
<td>Park View</td>
<td>11</td>
<td>High</td>
<td>February 2017</td>
<td>GP (1) Non-clinician (4)</td>
<td>Pharmacist (2) Non-clinician (2)</td>
</tr>
<tr>
<td>Rose Petal</td>
<td>10</td>
<td>High</td>
<td>February 2017</td>
<td>GP (2) Non-clinician (1)</td>
<td>GP (2)</td>
</tr>
<tr>
<td>Still Water</td>
<td>7</td>
<td>High</td>
<td>February 2017</td>
<td>GP (1)</td>
<td>GP (1)</td>
</tr>
<tr>
<td>Sapphire Lake</td>
<td>5</td>
<td>Low</td>
<td>February 2017</td>
<td>GP (1) Non-clinician (1)</td>
<td>GP (1) Non-clinician (1)</td>
</tr>
<tr>
<td>Green County</td>
<td>3</td>
<td>High</td>
<td>February 2017</td>
<td>GP (1) Nurse (1)</td>
<td>GP (1) Nurse (1)</td>
</tr>
<tr>
<td>Maple Grove</td>
<td>11</td>
<td>High</td>
<td>March 2017</td>
<td>GP (2) Nurse (2)</td>
<td>Non-clinician (2)</td>
</tr>
</tbody>
</table>

214
Research objective 1: Understand how health professionals and primary care practices respond to, and interact with, actionable e-A&F

Figure 3 demonstrates how GP practices accessed PINGR over time, and Table 4 shows how PINGR’s use varied across practices according to successful A&F processes in CP-FIT [8]. Overall, PINGR was used on 227 separate occasions (28.4 times per month) for a median of 67.0 minutes per session (range 1.0 – 383.0 minutes). Individual users accessed it a median of 3 times in total (range 1-23) and 1.3 times per month (range 0.1 – 18.7). Practices overall accessed it a median of 9 times in total (range 1-44) and 3.7 times per month (range 0.1 – 37.3). There was no discernible pattern of use over time (Figure 3). In terms of frequency of use, GPs, nurses, pharmacists, and non-clinicians used PINGR for a median of 0.8 (range 0.1-3.2), 1.5 (0.8-5.8), 1.4 (0.6-2.0), and 1.6 (0.2-18.7) times per month respectively (p=0.3). In terms of session length, GPs, nurses, pharmacists, and non-clinicians used PINGR for a median of 50.0 (range 1.0-383.0), 52.0 (1.0-127.5), 123.0 (5.0-255.0), 69.5 (0-358.0) minutes respectively (p=0.5).

Figure 4 demonstrates how each quality indicator was accessed over time, and Table 5 shows how patient-level A&F processes detailed in CP-FIT [8] varied between indicators compared to intervention and controls. Again, there was no discernible pattern of use over time (Figure 4). Overall, 725 patients were viewed across all quality indicators (Table 5): users disagreed with PINGR’s suggested actions in 15.3% (Non-acceptance), made plans to improve on 72.8% (Intention), and ultimately took action on 60.9% (Behaviour). Of those who were viewed in PINGR, 347 (47.8%) improved in at least one indicator after being viewed, compared to 6453 (29.6%) of those not viewed in PINGR (Table 5; relative risk ratio of improvement [RRR] = 1.6, 95% CI 1.5-1.7). This contrasts to only 64 (8.8%) patients who improved in all indicators they were violating after being viewed in PINGR, compared to 3291 (15.1%) not viewed (RRR = 1.7, 95% CI 1.4-2.2). Regarding individual indicators, patients viewed in PINGR were more likely to improve in the CKD staging and CKD diagnosis quality indicators when compared to patients not viewed. In ‘Hypertension...
casefinding’ and ‘AF casefinding acute’, rates of improvement after being viewed in PINGR were also higher but statistically non-significant. In all other indicators, rates of improvement were less likely when patients were viewed in PINGR.

Indicators were accessed with different frequencies \( (p=0.03) \): Hypertension and AF casefinding (acute symptoms) indicators were accessed most frequently (median 9.5 and 8.0 times per month, range 3.0-13.0 and 5.0-12.0 respectively) along with COPD exacerbations and PR (8.0, 2.0-13.0); the least frequently accessed were CKD monitoring (2.0, 1.0-5.0) and CKD diagnosis (3.0, 1.0-9.0). Users accessed the lowest performing indicator in only 40 (17.6%) sessions using PINGR, and the second, third and fourth lowest indicators 74 (32.6%), 48 (21.1%) and 24 (10.5%) sessions respectively.

Figure 5 is a process map of how participants used PINGR. The most common transitions were between the Overview page, Indicator, and Patient pages, suggesting that users mainly used PINGR in the way it was intended i.e. starting with population-level data, and moving to individual patient-level data. The second most common involved accessing the All patients list and Saved actions or Search patient pages. In 39 (19%) sessions performed mostly by GP users, only the Overview page was accessed. Trend and Benchmarking pages were accessed infrequently, with the Help and Contact pages accessed least. In total, 515 emails were sent during the course of the study. Most sessions occurred on the day an email was sent \( (n=69, 30.4\%) \), which declined steadily throughout the rest of the week.
Figure 3: GP practices use of PINGR per month

- Golden Valley
- Grand Oak
- Green County
- Hope Garden
- Kindred Medical
- King's Way
- Maple Grove
- Oak Dale
- Over Peak
- Park View
- Rose Petal
- Sapphire Lake
- South Shore
- Still Water
- Swan River
Figure 4: Quality indicators accessed per month

Key: AF = atrial fibrillation; BP = blood pressure; CKD = chronic kidney disease; COPD = chronic obstructive pulmonary disease; DM = diabetes mellitus.
Table 4: Processes detailed in Clinical Performance Feedback Intervention Theory in practices using PINGR (ordered by recruitment date)

<table>
<thead>
<tr>
<th>Practice</th>
<th>Audit Patients with suggested actions</th>
<th>Interaction Sessions per month</th>
<th>Perceptio n Patients viewed (% of Audit)</th>
<th>Non-acceptance Patients with disagreed actions</th>
<th>Intention Patients with planned actions (% of Perception)</th>
<th>Behaviour Patients with relevant changes in their EHR after being viewed in PINGR (% of Perception)</th>
<th>Improvement (any) Patients with improvements in any indicator after being viewed in PINGR (% of Perception)</th>
<th>Improvement (all) Patients with improvements in all indicator after being viewed in PINGR (% of Perception)</th>
<th>Qualitative description of PINGR adoption (scale: full, partial, limited or failed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oak Dale</td>
<td>798</td>
<td>0.5</td>
<td>10 (12.5)</td>
<td>2 (20.0)</td>
<td>7 (70.0)</td>
<td>10 (100)</td>
<td>5 (50.0)</td>
<td>1 (10.0)</td>
<td>Initial agreement for 1 GP to access. Limited adoption due to workload issues.</td>
</tr>
<tr>
<td>Golden Valley</td>
<td>2075</td>
<td>0.8</td>
<td>24 (11.6)</td>
<td>21 (87.5)</td>
<td>10 (41.2)</td>
<td>23 (95.8)</td>
<td>18 (75.0)</td>
<td>0 (0)</td>
<td>Initial agreement for 1 GP and 1 nurse to use. Partial adoption by one nurse interested in CKD. Staffing issues prevented ongoing use.</td>
</tr>
<tr>
<td>King's Way</td>
<td>714</td>
<td>0.1</td>
<td>4 (0.6)</td>
<td>1 (25.0)</td>
<td>2 (50.0)</td>
<td>4 (100)</td>
<td>1 (25.0)</td>
<td>0 (0)</td>
<td>Initial agreement for 1 GP to access. Failed adoption due to workload issues.</td>
</tr>
<tr>
<td>Hope Garden</td>
<td>2808</td>
<td>0.5</td>
<td>2 (0.07)</td>
<td>1 (50.0)</td>
<td>1 (50.0)</td>
<td>2 (100)</td>
<td>2 (100.0)</td>
<td>0 (0)</td>
<td>Initially agreed to be used by 1 GP, 1 nurse, and 1 non-clinician. Failed adoption due to workload issues.</td>
</tr>
<tr>
<td>Kindred Medical</td>
<td>2306</td>
<td>3.7</td>
<td>71 (3.1)</td>
<td>24 (33.8)</td>
<td>21 (29.6)</td>
<td>45 (63.4)</td>
<td>24 (33.8)</td>
<td>1 (1.4)</td>
<td>Large multi-site practice with strong internal processes for quality improvement. Partial adoption by both nurses and pharmacists.</td>
</tr>
<tr>
<td>Park View</td>
<td>1459</td>
<td>8.6</td>
<td>189 (13.0)</td>
<td>10 (5.29)</td>
<td>187 (99)</td>
<td>96 (50.8)</td>
<td>67 (35.5)</td>
<td>28 (14.9)</td>
<td>Initial agreement for 1 GP and 5 non-clinicians to use. Full adoption into working practices of 3 non-clinicians as part of job plan.</td>
</tr>
<tr>
<td>Location</td>
<td>CKD</td>
<td>EHR</td>
<td>Patients</td>
<td>Non-clinicians</td>
<td>Nurses</td>
<td>GPs</td>
<td>Adoption Details</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-----</td>
<td>-----</td>
<td>----------</td>
<td>----------------</td>
<td>--------</td>
<td>-----</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rose Petal</td>
<td>1911</td>
<td>3.3</td>
<td>15 (0.8)</td>
<td>0 (0)</td>
<td>5 (33.3)</td>
<td>6 (40)</td>
<td>5 (33.3) 1 (6.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Still Water</td>
<td>1332</td>
<td>1.6</td>
<td>12 (0.9)</td>
<td>0 (0)</td>
<td>1 (8.3)</td>
<td>8 (66.7)</td>
<td>6 (50.0) 2 (16.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sapphire Lake</td>
<td>834</td>
<td>1.3</td>
<td>26 (3.1)</td>
<td>4 (15.4)</td>
<td>15 (57.7)</td>
<td>15 (57.7)</td>
<td>11 (42.3) 3 (11.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green County</td>
<td>594</td>
<td>7.4</td>
<td>27 (4.5)</td>
<td>3 (11.1)</td>
<td>13 (48)</td>
<td>15 (55.6)</td>
<td>6 (22.2) 1 (3.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maple Grove</td>
<td>1481</td>
<td>5.6</td>
<td>34 (2.3)</td>
<td>4 (11.8)</td>
<td>23 (67.6)</td>
<td>18 (52.9)</td>
<td>13 (38.2) 2 (5.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Shore</td>
<td>1718</td>
<td>4.1</td>
<td>95 (5.5)</td>
<td>21 (22.1)</td>
<td>54 (56.8)</td>
<td>51 (53.7)</td>
<td>42 (42.2) 7 (7.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grand Oak</td>
<td>1151</td>
<td>2.5</td>
<td>5 (0.43)</td>
<td>0 (0)</td>
<td>1 (20.0)</td>
<td>1 (20.0)</td>
<td>1 (20.0) 0 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swan River</td>
<td>1061</td>
<td>34.2</td>
<td>185 (17.4)</td>
<td>12 (6.5)</td>
<td>146 (78.9)</td>
<td>132 (71.4)</td>
<td>128 (69.2) 9 (4.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over Peak</td>
<td>2842</td>
<td>7.3</td>
<td>26 (0.9)</td>
<td>1 (3.8)</td>
<td>24 (92.3)</td>
<td>8 (30.8)</td>
<td>8 (30.8) 1 (3.8)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Key: CKD = chronic kidney disease; EHR = electronic health record.
Note: numbers of patients may not tally with Table 5 due to patients moving practices.

Initial agreement for GPs to access. Limited adoption limited by workload issues.
Used by 1 GP. Limited adoption limited due to compatibility with workflows.
Initial agreement for 1 GP, 1 non-clinician and 1 nurse to access. Limited adoption limited by staffing issues.
Initial agreement for 1 GP, 1 nurse, and 1 non-clinician to use. Full adoption by nurse and partially adopted by GP; non-clinician did not adopt.
Initial agreement for GPs and non-clinicians to use, but failure to adopt. Partial adoption later by nursing staff.
Initial agreement for GPs, nursing staff, pharmacists, and non-clinicians to use. Partial adoption by GPs and nursing staff into job plans, but not by others. Adoption limited by staff absence.
Recruited later. Partial adoption due to workload issues.
Recruited later. Full adoption by non-clinicians as part of job plan.
Recruited later. Partial adoption by non-clinicians.
Table 5: Patient-level processes detailed in Clinical Performance Feedback Intervention Theory according to quality indicator (ordered by dates introduced)

<table>
<thead>
<tr>
<th>Quality indicator</th>
<th>Audit Patients with suggested actions</th>
<th>Interaction Sessions per month</th>
<th>Perception Patients viewed</th>
<th>Non-acceptance Patients with disagreed actions</th>
<th>Intention Patients with planned actions (% of Perception / Audit)</th>
<th>Behaviour Patients with relevant changes in their EHR after being viewed in PINGR where appropriate (% of Perception / Audit)</th>
<th>Improvement (any)* Patients with improvements in any indicator after being viewed in PINGR where appropriate (% of Perception / Audit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All indicators</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>21784</td>
<td>28.4</td>
<td>725</td>
<td>111 (15.3)</td>
<td>521 (72.8)</td>
<td>442 (60.9)</td>
<td>12913 (59.3)</td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKD monitoring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>1017</td>
<td>2.1</td>
<td>112</td>
<td>42 (22.7)</td>
<td>112 (60.5)</td>
<td>32 (28.6)</td>
<td>499 (49.1)*</td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKD diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>554</td>
<td>3.4</td>
<td>53</td>
<td>6 (7.2)</td>
<td>60 (72.3)</td>
<td>41 (77.4)</td>
<td>213 (38.4)*</td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKD staging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>2583</td>
<td>3.4</td>
<td>303</td>
<td>51 (16.8)</td>
<td>189 (62.2)</td>
<td>160 (52.8)</td>
<td>657 (25.4)*</td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKD and proteinuria BP control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>54</td>
<td>4.8</td>
<td>41</td>
<td>14 (27.0)</td>
<td>23 (44.2)</td>
<td>12 (29.3)</td>
<td>34 (63.0)*</td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension BP control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>14068</td>
<td>6.8</td>
<td>317</td>
<td>68 (15.5)</td>
<td>310 (70.6)</td>
<td>100 (31.5)</td>
<td>6395 (45.5)*</td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKD and DM BP control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>685</td>
<td>7.8</td>
<td>126</td>
<td>35 (23.5)</td>
<td>80 (53.7)</td>
<td>34 (27.0)</td>
<td>350 (51.1)*</td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKD BP control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>2426</td>
<td>5.7</td>
<td>209</td>
<td>50 (17.3)</td>
<td>175 (60.6)</td>
<td>79 (37.8)</td>
<td>1283 (52.9)*</td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Patients NOT viewed in PINGR</td>
<td>Patients viewed in PINGR</td>
<td>Patients NOT viewed in PINGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------</td>
<td>--------------------------</td>
<td>------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hypertension casefinding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>842</td>
<td>10.6</td>
<td>133</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td>9 (6.5)</td>
<td>97 (70.3)</td>
<td>42 (31.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>238 (28.3)</td>
<td>38 (28.6)</td>
<td>218 (25.9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>COPD exacerbations and PR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>315</td>
<td>7.0</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td>7 (11.5)</td>
<td>45 (73.8)</td>
<td>32 (60.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>264 (83.8)*</td>
<td>26 (49.1)</td>
<td>214 (67.9)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AF casefinding acute</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>642</td>
<td>12.5</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td>9 (25.0)</td>
<td>31 (86.1)</td>
<td>5 (25.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>101 (10.3)</td>
<td>4 (20.0)</td>
<td>73 (11.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AF casefinding chronic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients viewed in PINGR</td>
<td>9753</td>
<td>10.0</td>
<td>269</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients NOT viewed in PINGR</td>
<td>52 (14.0)</td>
<td>262 (70.4)</td>
<td>55 (20.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7929 (81.3)*</td>
<td>13 (4.8)</td>
<td>809 (8.2)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Difference from intervention practices is statistically significant at 0.05 level.

Note: Totals do not tally between individual indicators and ‘all’ indicators because patients may be in multiple indicators.

Key: AF = atrial fibrillation; BP = blood pressure; CKD = chronic kidney disease; COPD = chronic obstructive pulmonary disease; DM = diabetes mellitus; EHR = electronic health record.
Figure 5: Process map of PINGR use by individual users

Key = Numbers represent counts of sessions. Boxes represent the pages accessed in PINGR. Arrows represent transitions between pages. Dotted green and red arrows represent a start and end of a session respectively.
Research objective 2: Determine facilitating and inhibiting factors to explain findings from objective 1

Interviews and observations revealed three inter-related pairs of concepts from CP-FIT that explained the differences observed between practices, users, and indicators in research objective 1. These were: available resources to engage with PINGR (Resources) and PINGR’s perceived advantages (Relative advantage); PINGR’s compatibility with practices’ and users’ ways of working (Compatibility), and how this impacted on its credibility (Credibility); and the ability to act on PINGR’s data (Actionability), related to the complexity of the patients it highlighted (Complexity). We discuss each in turn below.

**Resources and relative advantage**

Despite most participants felt PINGR had advantages over existing e-A&F systems concerning its usability, suggested actions, and educational elements, they often struggled to find the time to use it. UK general practice is under increasingly significant workforce and workload pressures [38], and because additional resource was not provided alongside the intervention practices often struggled to find time to use it and take the necessary corrective actions. This was compounded by the perceived relatively short time period practices had to use the software and act on its feedback (median 4 months). Those that fully or partially adopted the software tended to provide their clinicians with dedicated time for administrative or quality improvement work in which they used PINGR, whereas practices with limited or failed adoption generally did not (Table 4). Reasons for whether practices had dedicated non-clinical time were not necessarily a function of their beliefs regarding quality improvement, attitude towards PINGR or ability to organise themselves, but as a consequence of understaffing and resulting increased workload. In this situation, competing demands were often prioritised above PINGR use, such as direct patient care and using the e-A&F system that had been designed for the local pay-for-performance scheme. This appeared to be a particular problem for smaller practices, where fewer staff resulted in less capacity to cope with understaffing. The type of user was also important: in general, non-clinicians had more time to use PINGR because their working days were more flexible than clinicians who had fixed times to provide patient care, whereas GPs tended to solely access the Overview page in order to monitor overall scores by virtue of their leadership position within a practice:

*To be honest with you, the main thing is we’re so busy at the moment, we’ve not got a lot of time even for the stuff we’re doing on a day to day basis. And part of that is because of staff shortages due to being under capacity in terms of doctors and also doctor illness recently as well, and lots of calls on our time. So I’d say that’s one of the main reasons [for not using PINGR]*

GP, practice 9, second interview
A further resource issue concerned the large number of patients highlighted as receiving suboptimal care. Because PINGR focused on quality indicators related to high prevalence conditions, usually with low baseline scores that reset on 1st April, often hundreds of patients were highlighted in its patient lists (see the ‘Audit’ column in Table 4). This could discourage users from accessing PINGR because of the large amount of work it suggested was required, particularly in comparison to other quality improvement projects where only a few patients may be flagged for action. Furthermore, these larger numbers of patients meant that any work they did undertake per patient had less impact on their overall score:

My only downside is trying to juggle the numbers that it's generating for me because, yes, it's just more work… But it's just like the [other quality improvement programme], when we first started that there was tons wasn't there, but then you start working through [and numbers reduce quickly]… I think you look at [PINGR] and you are like, oh my God, there's a lot there…

Pharmacist, practice 2, second interview

Compatibility and Credibility
In relation to PINGR’s compatibility with practices’ and users’ goals, most had positive attitudes towards quality improvement work, consequently its perception as a quality improvement rather than a performance management tool was key in its acceptance. In general, users felt the quality indicators in PINGR were both important and relevant to primary care, and because many of them aligned with the local pay-for-performance scheme it encouraged use by helping practices achieve their financial and clinical goals. Indicators that were not aligned with the pay-for-performance scheme, or that were felt to be less clinically important were accessed less frequently (Table 5; e.g. CKD monitoring), whereas those felt to be most important and not offered in other e-A&F systems were accessed most (Table 5, e.g. hypertension casefinding, CKD diagnosis):

So my slight thing say with adding that [CKD staging] code, I don’t want to over-complicate things when we’re so pushed for time. I know that’s really good to have the proper code on, but does that really benefit anyone? Do you know what I mean?

GP, practice 12, first interview

Fitting with practice’s existing technical infrastructure prevented PINGR from being used more frequently. Most practices stated that having PINGR separate from their EHR increased complexity with regard to accessing the software (they meant they often forgot)
and taking patient action (as this had to be done separately in the EHR). Email reminders alleviated this problem to some extent, though all participants stated they would prefer PINGR integrated with their EHR: appearing as an alert, with the ability to open and write into patients’ records directly.

And [the other e-A&F system] is quite good, because if I open it up in the background at the same time, I am getting these triggers that are relevant to my patients. If I went to a different patient, you know, that trigger would pick up that I've gone into a new patient, and do it. So that's nice, it ties it in, but it does slow down the programmes as well. There's a definite lag, a definite slowing down, I think. But it does mean I'm not flipping in, it is integrated into my data. So if it isn't integrated into my tool, my daily tool, which is [EHR system], I won't use it.

GP, practice 10, first interview

Misalignment with practices’ workflows reduced the number of patients viewed in PINGR on whom subsequent action was taken. All practices conducted annual reviews with patients in which they addressed aspects of chronic disease care highlighted in PINGR. However, when seven indicators reset their scores on 1st April in accordance with the local pay-for-performance scheme, PINGR started to suggest patients required action out of sync with their annual review date. For example, a patient with hypertension who had a blood pressure measurement in March appeared in PINGR as receiving suboptimal care in April. Consequently, when such patients were viewed in PINGR, users took no action because it was most appropriate to wait until their next annual review (Table 5). This sometimes affected whether users accessed these indicators because they thought they would be most valuable in the last few months prior to April:

The bit I was probably having slight difficulty in, and it's sort of difficult with a lot of them because [local pay-for-performance scheme] is the same difficulties again at the moment because when you start from April is it runs from April to April. It becomes very relevant when we get beyond December…Whereas as doctors we were tidying up in the second half of the year but ultimately you just get bombarded hugely and you think, well we're not actually doing the reviews because they're coming in on a recall [for chronic disease annual review].

GP, Practice 8, second interview

In general, PINGR’s use of secondary care patient data was considered advantageous because it provided useful information not usually available to primary care. However, using these data to determine care quality in chronic disease was sometimes incompatible with participants’ beliefs. For example, most agreed that blood pressure measurements
taken in secondary care should not be used to assess hypertension control because patients are often acutely unwell in hospital, leading to transient changes in blood pressure:

One thing that I noticed and one thing that I said is that on some of the hypertension ones some of the blood pressure readings are feeding in from hospital admissions or clinic. And I suppose the only comment I would have for that is whether that's the best time to be doing prevention. Because we had one guy there where he was in renal clinic and his blood pressure was measured at like 190/100 or something, so it does home BP so it was 125/85, so there's quite a discrepancy there.

GP, practice 13, second interview

**Actionability vs Complexity**

Most users felt that PINGR’s suggestions were useful, and helped them take improvement action. In general nurses, pharmacists, and non-clinicians found them more useful than GPs, which appeared to be a function of their perceived level of clinical knowledge. The patient-level actions were seen as more helpful than the organisational-level actions because they were demonstrably tailored to patients’ needs and directly relevant to clinical care. This was supported by the relative low levels of disagreement with PINGR’s suggestions (Table 5, Non-acceptance 15.3%). In contrast, they felt in general the organisational-level suggested actions were more generic, despite them being prioritised based on each practice’s specific reasons for suboptimal performance. This was reflected in reasons for disagreement collected by PINGR, where the most common reason for rejection was the action had been previously been attempted. The two practices that implemented some of PINGR’s organisational-level suggested actions (Kindred Medical and Green County) were both in highly deprived areas and shared an enthusiasm for PINGR, though differed significantly in terms of size and culture. In general, the organisational-level actions appeared more difficult to implement, requiring more time and resources, though the ones that were implemented were the most simple and involved distributing education material to staff:

Interviewer: Okay. And then, finally, I know that you've all been using it for the patient level and looking at these recommendations on the patient level, but we also have these recommendations at the organisational level. Have you used those much, are they helpful?
Respondent 1: I've not, I've not looked at them.
Respondent 2: I've not really looked at them either, I think they will be helpful but, again, it's just that you can only concentrate on one thing at one time, can't you?
Respondent 1: At the moment, yeah, we don't think of the bigger view, we're just looking at individual patients.

Non-clinicians, practice 6, second interview

You can see that one [suggested action], we have done that, it's in every room. We have...laminated the [hypertension NICE guideline] pathway and put that in every room… But, what I am considering very, very seriously is a healthcare assistant for the reasons that I have just given you. Because when we start bringing [undiagnosed hypertension patients] back and doing all those things, I could just see more time required.

GP, practice 14, second interview

Quality indicators in which taking patient-level action was less complex saw greater improvements at both the practice and patient levels (Table 5). For example, indicators that measure clinical processes (e.g. AF casefinding) or recorded diagnoses (e.g. CKD staging) are easier to improve upon compared to those that measure whether patients achieve specific outcomes (e.g. blood pressure control) or require further investigation (e.g. hypertension casefinding). In the former, improvement action is largely within control of the clinician, whereas in the latter it requires more time and depends on patient engagement and response to management. Furthermore, users suggested that patients highlighted in PINGR were often more difficult to treat. This may be expected: by definition the system flags patients who have not achieved care quality standards as part of their routine care. However, because PINGR also prioritises patients violating multiple indicators for review this effect may be exacerbated. Consequently for many patients viewed in PINGR it was difficult to take action because they were medically complex or were often under secondary care consultants in which case primary care clinicians would not intervene. This was supported to some extent by quantitative analyses that found patients viewed in PINGR tended to be older (median 70 vs 66 years, p<0.001) and violate more quality indicators (median 3 vs 1, p<0.001) than those not viewed. This in part may explain why patients viewed in PINGR (particularly for outcome indicators) had lower levels of Behaviour and Improvement than those not viewed in PINGR in intervention practices, and why patients viewed in PINGR overall had a lower chance of improving in all indicators they violated (Table 5):

Of all the patients the CKD ones tend to be the hardest to get their blood pressure within the goals and they're usually under renal and they're usually under optimum therapy anyway. So it's not the most realistic thing to try and do that for them, to bring them in and give them the blood pressure monitors.

Nurse, practice 8, second interview
And so… people like her, so she… I think she’s under the renal team anyway. So we’d be really wary about… she’s like CKD stage four, I think… and we probably wouldn’t be tinkering with her medication for fear of making her kidneys worse.

GP, practice 12, second interview

7.5 Discussion

Summary and explanation of findings
Our results demonstrate that an e-A&F system that suggests tailored improvement actions for organisations and individual patients can be developed and implemented into clinical practice, whilst being both acceptable and useful to health professionals in UK primary care. GP practices adopted the system to different extents and in different ways, though each individual user largely used the software as intended: taking direct action based on its feedback rather than generally raising awareness of clinical performance. However, this action was mainly at the patient-level, rather than the organisational-level. The perceived advantages of the system lay in its user friendliness, ability to suggest actions, and educational elements. Patients who were viewed in PINGR were overall more likely to receive improvements in at least one quality indicator. This was particularly true with indicators that were less complex to action, and not addressed by other e-A&F systems e.g. those focusing on CKD diagnosis and staging, and AF casefinding (Table 5). Patients viewed in PINGR tended to violate more quality indicators than those that were not, therefore these improvements may simply be a function of more chances to improve. The main barriers and facilitators to the software’s success related to the resources each practice had available to use it; how compatible it was with practices’ and users’ beliefs and ways of working; and the ability to take action based on its recommendations.

Comparison to existing literature
Two recent systematic reviews of e-A&F systems have suggested that further research is required to understand the mechanisms underlying how these systems work [5,8]. The in-depth interviews conducted as part of this study, in combination with analysis of usage logs go some way to address this, and can largely be explained by concepts in CP-FIT [8]. The main mechanism of how PINGR was used by practices can be explained by CP-FIT’s ‘Direct Action’ mechanism, whereas it was used less via the ‘Knowledge/Awareness’ mechanism [8]. Reasons for its adoption can be explained by its perceived Relative advantages, high Credibility and low Complexity, whereas the main barriers and facilitators to success are directly related to the constructs of Resource match, Compatibility, and Actionability [8]. The only mechanism in CP-FIT that was not found to be important in this study was Social influence, which relates to interpersonal processes that cause health professionals to change their thoughts, feelings, or behaviours in
response to A&F [8]. Social influence has been found to be a key mechanism of influence in A&F studies (e.g. [39]), and was operationalised in PINGR by comparing practices’ scores to their neighbours and the wider city in the Benchmarking and Overview pages (Table 1). Although these comparisons were described as interesting and motivating during interviews, they were not found to be a key driver of success. This was supported by PINGR’s usage analysis, which showed the Benchmarking page was accessed relatively infrequently (Figure 5).

Two other studies have also examined the usage records of e-A&F systems [40,41]. Similar to our findings, Gude et al. found that health professionals in intensive care often ignored their current performance level, and opted to improve on quality indicators in which they were already performing optimally because they were considered essential aspects of care [40]. Whereas Wahid et al. found that use of a neonatal intensive care e-A&F system was inversely related to units’ baseline performance on health care associated infections, though directly proportional to their subsequent improvements [41]. In the current study, quality indicators considered most clinically important, and that were part of the health professionals’ workplan were accessed most frequently.

The e-A&F system studied in this paper is one of only three of which we are aware to suggest tailored actions that users could take based on their feedback results. This is surprising given that A&F interventions have been shown to be more effective when action plans are provided [4]. The actions provided by other reported e-A&F systems solely provided patient-level actions on adverse drug events [42] and chronic care [43], whereas those in PINGR also included organisational-level actions based on the hypothesis that organisational change is required to most effectively improve patient care [8]. Evaluations of these other actionable e-A&F systems focused only on the quantitative outcomes of the systems [42,43]. Consequently, the present study provides unique insights into how health professionals react and respond to suggestions. Only one of the other actionable e-A&F systems studied processes of the system in which they found the positive predictive value of its suggestions to range between 0-67% [42]. This may be compared to PINGR’s rates of non-acceptance of suggestions of ranging between 6.5-27.0% (Table 5), though this figure refers to numbers of patients rather than individual suggestions (an individual patient in PINGR often had multiple suggestions).

A further unique feature of the system studied was the use of emails to users. This has been used in a relatively few number of other e-A&F systems in order to solely remind users to access the e-A&F system with variable success [44–46]. PINGR’s approach differed in that it also fed back potentially high priority patients on which action could be taken. Our results demonstrated this was both useful and acceptable to health
professionals, and may have had important effects on their likelihood of accessing the software.

**Implications**

In addition to providing specific improvements for PINGR, this study has wider implications for theory in relation to e-A&F systems in general. With regards to PINGR, its optimisation will address the barriers identified in research objective 2, including alignment of its quality indicators with practice’s provision of patients’ annual reviews, integration with EHR systems, changes to how it handles secondary care data, and new designs to encourage organisational-level action and harness social influence.

We used CP-FIT as a model to design and implement PINGR, and guide the collection and analysis of evaluation data. A limited number of e-A&F systems were used to generate CP-FIT [8], however as described above it is able to explain most of the this study’s findings providing support for its relevance to and explanatory power for computerised as well as non-computerised A&F interventions. Furthermore, this study adds to the model by: demonstrating how some of its concepts may inter-relate in practice, providing example detail to its concepts, in addition to testing novel features of A&F that align with its propositions but have not been previously investigated [8].

Different mediating variables from CP-FIT interacted with each other to explain the key facilitating and inhibiting factors described in objective 2: although PINGR had perceived benefits over existing e-A&F systems (*Relative advantage*), practices often did not have the resources to use it as much as they wanted (*Resource match*); when PINGR was incompatible with practice’s ways of working (*Compatibility*), this impacted its credibility by reducing the perceived accuracy of its data (*Credibility*); and because PINGR often highlighted complex patients or improvement actions to be taken (*Complexity*), this reduced the ability for users to action based on its feedback (*Actionability*). Some of these concepts address the same propositions in CP-FIT – for example, both *Resource match* and *Relative advantage* relate to Proposition 2 (Capacity limitations) – suggesting they are linked. However, this study provides important learning with regard to how these links may arise in practice.

In relation to providing conceptual detail, three key findings related to the: numbers of patients highlighted as receiving suboptimal care; complexity of the target behaviour; and how e-A&F systems integrate with existing technical infrastructure. Large numbers of patients led to less user action due to limited organisational resources and negative responses from health professionals. This is partially captured in CP-FIT as *Performance level* [8], however its description of low performance is unclear and its effects are not
demonstrated with high confidence. The findings from this study therefore clarify that performance may relate to relative performance on a quality indicator score, or absolute levels relating to the number of patients that require action: the most important aspect is the amount of resource required to improve the performance on the quality indicator. In contrast, the complexity of the target behaviour for improvement is not captured in CP-FIT, though may be inferred from the Complexity mediating variable [8]. Consequently, this study has provided an additional illustration of how a general rule from CP-FIT may exist in clinical practice. Interestingly, a recent trial of an e-A&F intervention focusing on medication safety where absolute numbers of patients requiring action were relatively low, and the target behaviour relatively simple (e.g. stopping medication), was shown to be effective [47]. Finally, integration of e-A&F systems with existing software (i.e. practice’s EHRs) is captured as Compatibility and Workflow fit in CP-FIT [8], though our findings provide an important illustration of how this occurs in practice that is likely applicable to all e-A&F systems.

The main novel feature PINGR tested in accordance with CP-FIT was the provision of tailored suggested actions, though others include the use of email to increase Active delivery of feedback reports, and the use of in-built evidence to increase users’ clinical Knowledge and skills [8]. All these features appeared to be acceptable to health professionals and potentially effective, providing support for CP-FIT’s assertions and examples of how they may be applied in practice. However, attempts by PINGR to induce organisational-level action and harness social influence were less effective. Consequently, future research should investigate further ways in which this can be achieved, whilst minimising the resources required to engage with the intervention as per CP-FIT’s other propositions [8].

**Strengths and limitations**

The strengths of this study lie in its use of both qualitative and quantitative data to test a theoretical model (CP-FIT) used to design and implement an e-A&F system. Triangulating both types of data enabled a richer and more credible understanding of study findings [35], and provided greater detail on how PINGR could be optimised and wider learning established than if either data set was used alone [13]. This was further supported by the granularity of data collected regarding user interactions with PINGR, and longitudinal qualitative data from repeat interviews, which provided detailed objective and subjective insights into potential mechanisms of success and failure [48,49]. The two practices that dropped out of the study (Hope Garden, and King’s Way) could not be interviewed in-depth about their reasons, though contact via email suggested they were similar to factors encountered in other practices (i.e. workload and not enough resource).
The main weaknesses of this study relate to the influence of the research team, potential bias, and limitations of the quantitative data collection and analyses. Author BB developed PINGR and held a position recognisable by participants (as a GP), consequently during interviews there were issues of reflexivity and trustworthiness relating to: imposing ‘preconceived notions’ of how PINGR should be used by participants [50]; ‘shared conceptual blindness’ in which common experiences shared by participants and the researcher go unquestioned [51]; participants feeling obliged to portray an overly positive view of PINGR to please the research team [13]; and because PINGR evaluated participants’ clinical performance, they may have felt judged or threatened leading to a less honest disclosure of events [51]. Therefore, we took three steps to address these issues. Firstly, we involved a multidisciplinary research team to provide a detached and critical input into the data analysis (including TB, RW, and MS [statistician] and IB [public health informatician] both whom were removed from the software development process). Secondly, we focused qualitative data collection on working practices and processes of PINGR’s implementation (embedding and integration), rather than participant’s subjective views or personal performance by using NPT to inform our interview guide (Appendix 12) [26]. Thirdly, we conducted repeat interviews to enable users to develop a relationship with author BB in which they felt comfortable ‘opening up’ and critiquing the software [49,51]. Although these steps were taken, author BB’s position may have conversely also had positive effects. His position as a fellow health professional may have afforded ‘insider’ status that may not have been available to a non-clinician qualified [52]. Furthermore, as creator of the software and a target user, his detailed knowledge of the system and context may have meant his interview questions and interpretation were more relevant [53].

Sources of bias relate to GP practice selection and the Hawthorne effect. GP practices who had a track record of trialling health care innovations were recruited. This is a widely accepted method of optimising new software, as it increases the likelihood it will be used and that information is gained regarding how it could be improved [18]. However, it means that our findings may not be generalisable to all primary care practices and should be interpreted with caution. For example, quantitative estimates of PINGR’s effects on patient care may simply be a function of how these practices work, and their acceptance of PINGR may not be shared by practices without a track record of innovation. Generalisability may also be exacerbated by practices being recruited from only one region in the UK. On the other hand, our sample included practices with a wide range of sizes and levels of deprivation, so our qualitative findings may have wider relevance [24]. Participants were aware their interactions with PINGR were remotely monitored and they may be interviewed in future about their experiences using the software. Consequently they may have either consciously or subconsciously used the software more often or
changed the way they provided care to patients in order to appear more favourable to the research team [54]. However, given our objectives were to refine hypotheses about PINGR’s implementation and causal pathways, not estimate intervention effects, this may not necessarily be judged a weakness of study design [55]. Nevertheless, further testing of actionable e-A&F systems should occur in studies designed to test their effects on patient care and acceptability in more representative primary care populations.

Further limitations relate to the usage records where for patient-level effects (Table 5), those that received care processes via PINGR’s emails or were viewed in PINGR’s lists without accessing their individual patient page, or via general raised awareness of clinical issues from using PINGR would not have been captured. Future research will use the results of this study to inform a robust study design to evaluate the effectiveness of the intervention.

7.6 Conclusion
This study has demonstrated the feasibility, acceptability and potential value of implementing a theoretically informed e-A&F system that suggests improvement actions into UK primary care. In doing so, we have understood how health professionals and primary care practices respond to, and interact with, actionable e-A&F (research objective 1); and determined facilitating and inhibiting factors (research objective 2). Our findings provide not only specific guidance on how the system could be optimised in future, but also wider learning related to e-A&F systems and theory in general. We used CP-FIT [8] to design and implement the e-A&F system, in addition to guiding data collection and analysis. We found CP-FIT had good applicability for e-A&F, explaining our study findings in relation to the relative advantage of the system, the resources each practice had available to use it, how well it fitted with practices’ and users' beliefs and ways of working, and the ability to take action based on its recommendations. Furthermore, our results provide practical illustrations of how particular phenomena describe in CP-FIT inter-relate (e.g. Relative advantage and Resource match) and occur in practice (e.g. Performance level), additional concepts that currently lie outside the model but align with its propositions (e.g. goal complexity), and how novel A&F designs may address some of its recommendations (e.g. suggested actions). Future research should continue to investigate novel ways to deliver e-A&F and test CP-FIT’s assumptions.

6.7 References


Chapter 7 concluding note

This chapter addressed RO2 and RO3 by using CP-FIT and the usability design guidelines for e-A&F systems from Chapter 6 to design and evaluate the third version of PINGR. This represented a progression in PINGR’s evaluation process by testing its feasibility and acceptability in practice. In doing so, it derived generalisable learning about how CP-FIT’s concepts arise in clinical practice. This is the final ‘results’ chapter the thesis – the next chapter discusses the findings and significance of all results chapters (Chapters 3-7).
Chapter 8

Discussion

8.1 Introduction
This chapter focuses on the overall findings and approach of this thesis. Although it is not intended to address specific issues of individual studies, it does first summarise their methods and findings. Next, it discusses the strengths and weaknesses of the overall approach taken by the research programme and the tension between its ‘micro’ and ‘macro’ focus, before comparing its findings to the wider literature on e-A&F systems. Finally, it addresses the implications of this thesis, plans for future research, and patient and public involvement.

8.2 Summary of methods and findings
In accordance with Medical Research Council (MRC) guidance on the development of complex interventions [1], Chapters 3 and 4 initially established both a theoretical and evidence base for using EHR data for quality improvement, with a focus on audit and feedback (A&F). Chapter 3 examined the literature on electronic A&F (e-A&F) and Clinical Decision Support (CDS) systems, and made arguments for combining features from both in order to increase their potential effectiveness. One key recommendation was that e-A&F systems could provide recommendations for clinical action to health professionals. Chapter 4 systematically searched the literature for qualitative evaluations of A&F interventions, and from 65 included papers developed a new model of causal pathways in A&F effectiveness – Clinical Performance Feedback Intervention Theory (CP-FIT). CP-FIT has three propositions: 1) A&F interventions exert their effects by inducing patient-level behaviours in health professionals; 2) Health care organisations have limited capacity to engage with and respond to the demands of A&F interventions; and 3) Health care professionals and organisations have a strong set of beliefs and behaviours regarding how they provide patient care that influence their interactions with A&F. CP-FIT suggests that effective A&F is a cyclical process of Goal setting, Audit, Feedback, recipient Interaction, Perception, and Acceptance of the feedback, followed by Intention, Behaviour and Clinical performance improvement, and that progress round this cycle is influenced by key moderating and mediating variables. Based on these findings, Chapter 4 makes specific recommendations for the design and implementation of A&F interventions to improve their effectiveness. The remaining results chapters (Chapters 5–7) used outputs (emerging and final) from Chapters 3 and 4 to develop, evaluate and optimise an electronic A&F (e-A&F) system for UK primary care – the Performance Improvement plaN GeneratoR (PINGR), the unique feature of which was its suggestion of improvement
actions to users based on analysis of EHR data. These chapters followed Borycki et al.’s framework of sequential Heuristic Evaluation and Cognitive Walkthrough (Chapter 5), Usability Testing (Chapter 6), and Naturalistic Testing (Chapter 7), with defect correction after each stage [2]. Based on the findings of their studies, Chapters 5 and 6 introduced and refined a set of design guidelines for the interfaces of e-A&F systems relating to: summaries of clinical performance, patient lists, patient-level information, and suggested actions. Chapter 7 explicitly tested CP-FIT, demonstrating how some of its concepts are borne out in the real-world, in addition to illuminating other factors that may be beyond the scope of the original theory (though still align with its propositions). It found that key variables of CP-FIT may inter-relate, for example the Complexity of patients highlighted by PINGR in turn reduced the system’s Actionability, and its low Compatibility with pre-existing practice workflows reduced its Credibility. It also highlighted the importance of new moderating variables, previously not captured in the model but that align with its propositions, such as the complexity of the target behaviour. During the course of this thesis, my thinking regarding how e-A&F systems should use EHR data to drive quality improvement has evolved. Consequently, so has PINGR, which has been presented as three different versions (Table 1 and Figure 1), with a fourth version planned for post-doctoral work.

**Table 1: How PINGR has evolved during the course of this thesis**

<table>
<thead>
<tr>
<th>PINGR version</th>
<th>Design based on Chapter(s)</th>
<th>Presented and evaluated in Chapter</th>
<th>Main design features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3, 4 (emerging findings)</td>
<td>5</td>
<td>Suggest actions to users</td>
</tr>
<tr>
<td>2</td>
<td>3, 4 (emerging findings), 5</td>
<td>6</td>
<td>Summaries of clinical performance, Patient lists, Patient-level information, Suggested actions</td>
</tr>
<tr>
<td>3</td>
<td>4 (emerging findings), 6</td>
<td>7</td>
<td>Prioritisation of information, Displaying all data on one page, Compatibility with local priorities, Benchmarking, Leadership support</td>
</tr>
<tr>
<td>4</td>
<td>4, 7</td>
<td>N/A</td>
<td>Alignment of performance measures practice workflow, Integration with EHR systems, Changes to how secondary care data is handled</td>
</tr>
</tbody>
</table>
Figure 1: Relationship between thesis studies and chapters, PINGR development, research output, and guiding methodological frameworks

Key: blue boxes = theoretical chapters, orange = empirical studies.
8.3 Strengths and weaknesses

The strength of this thesis has been its utilisation of a range of methods, evidence, and theory to iteratively develop and test an approach to using EHR data to drive quality improvement. By choosing to ground the work in an established technique (i.e. A&F), it drew on an existing evidence base and theoretical literature, rather than starting from scratch. Using both qualitative and quantitative research methods enabled the identification of multiple issues and triangulation of findings that may not have been apparent using either type of technique alone [3]. By incorporating multiple sequential steps of evaluation of increasing ‘cost’, it enabled significant failures with PINGR to be identified and rectified early in its development cycle, increasing the likelihood it would be adopted into practice in Chapter 7 [2,4,5]. This reflects a Popperian approach [6], where the usability design guidelines and CP-FIT were developed and refined in Chapters 5 and 6, and 4 and 7 respectively. Using multiple methodologies also helped address specific areas that may have been missed by other chapters. For example, Chapter 4 focused on synthesising evidence from implemented A&F interventions only. Consequently it does not provide extensive detail on interface design features, though these are covered in detail in Chapters 5 and 6. Similarly, using CP-FIT to inform PINGR’s design and evaluation in Chapter 7 allowed it to address design features that have not previously been tested, and explore concepts not previously evaluated.

The intended approach to the PhD was intended to be sequential i.e. Chapter 4 would be finished first and its finalised outputs subsequently used to inform the subsequent chapters. However, although Chapters 5-7 followed this plan, as work progressed on Chapter 4 it became apparent the metasynthesis was such a substantial piece of work that it would not be finished in time to develop and evaluate PINGR after it was finished. Consequently, Chapters 5 and 6 used only its emerging findings. Non-linearity and flexibility of PhD research is common [7], though in this particular instance this was helped by the methodology used in Chapter 4 in which an iterative approach was used to develop an early theoretical model, and test and apply it to the next literature set. In doing this, an emergent model was available early on in the process to be used in Chapters 5 and 6. Furthermore, rather than a weakness, iteration is explicitly acknowledged as a part of complex intervention development [1].

A key weakness is that I led both the development and evaluation of PINGR throughout Chapters 5-7. As identified in Chapter 2, this posed the risk of a positivity bias in their results [8]. Specific steps were take in the evaluation methodologies to address this potential risk including the use of other researchers from different background to provide a detached and critical input into the research process [9], and focusing data collection and analysis on working practices and processes of using PINGR rather than subjective views.
This last step was facilitated by the use of Normalisation Process Theory [10] as a sensitising framework for interview questions in Chapters 6 and 7, and explore different types of work around the use of PINGR. As testament to these mitigating steps, a wide range of issues were found during PINGR’s evaluation in each chapter, demonstrating this risk was satisfactorily addressed.

Further weaknesses of this PhD relate to the delimitations discussed in Chapter 2, which specifically include: a narrow focus on the definition of care ‘quality’, and a lack of research involving patients, and a focus on A&F as an approach to using EHR to drive quality improvement. The first two points are related, for as discussed in Chapter 1, often the aspects of care quality that are ignored by quality improvement interventions are those regarding patient experience and satisfaction [11]. Although patients were not explicitly part of the research sample of this thesis, presentation of emerging findings happened to a number of patient and public involvement (PPI) groups. No specific concerns were raised regarding the definitions of quality used by PINGR, and they were often reassured that it used clinical guidelines as its standard of ‘best’ care. Nevertheless, these groups did make suggestions to the research, which are discussed in section 8.5 below. Finally, focusing on A&F rather than any other method to use EHR data for quality improvement may have limited the generalisability of my findings. However, as discussed throughout the thesis, there are significant overlaps between CDS and A&F that challenge this, and may warrant further investigation.

8.4 ‘Micro’ vs ‘macro’ intervention development

The approach to develop PINGR by this thesis could be described as having a ‘micro’ perspective: it focused on the optimisation of the intervention’s microscopic features (i.e. PINGR’s usability in artificial settings) before broadening out to evaluate its ‘macro’ impacts in the real world, rather than the other way round. In this section, I argue this was a function of: the overarching methodological frameworks used; and a potential artifact of how the thesis is presented. In doing so, I will also summarise: the arguments for and against a micro or macro approach to intervention development; and ultimately why I believe a micro approach is preferable.

As detailed in Chapter 2 and illustrated in Figure 1, this thesis used both guidance from the MRC [1] and Borycki et al. [2] to develop PINGR. The MRC advocates first identifying relevant evidence and theory (addressed by Chapters 3 and 4), followed by feasibility and pilot testing of complex interventions prior to their implementation and evaluation (Chapters 5-7) [1]. Boyckci et al. recommend sequential and iterative evaluation of increasing ‘cost’ of health information technologies prior to their deployment, starting with lab-based studies of Heuristic Evaluation and Cognitive Walkthrough (addressed by
Chapter 5), then Usability Testing (Chapter 6), followed by Naturalistic Testing in clinical practice (Chapter 7) [2]. Both frameworks clearly fit the micro approach to intervention development, by paying attention to an intervention’s inherent features before its implementation into clinical practice. Furthermore, both were born as a proposed solution to perceived problems with historic methods of intervention development [2,12]. For example, the MRC state that complex interventions are often implemented into practice without knowing their precise mechanisms of action, so cannot be easily translated to other settings (if found to be effective) [12]. This is pertinent to A&F where the historic approach to investigation has been ‘stagnant’ due to a lack of theorising as to how it may effect care [13], and interventions are often implemented into practice based on ideas that seemed good at the time [14]. However, the converse is also problematic: if a health information system is found to be ineffective and has not been developed using a micro approach, it may be unclear if poor usability contributed, or whether the results would have been different if usability had first been optimised [15]. Similarly, Borycki et al. argue that poor usability of health information systems have been implicated in a wide range of patient safety events (e.g. medication errors) because little attention has been paid to their optimal interface design [2]. This is particularly relevant to e-A&F where there has been little prior research that could have been drawn upon when designing PINGR (see section 5.2 on p.102). The main disadvantage of the micro approach is that problems related to implementation are not discovered until relatively late in the development process. An example of which in this thesis are issues related to the lack of integration between PINGR and existing EHR software (Chapter 7). Such issues can still be rectified, as the iterative development process continues following implementation [2,12] – it just happens later.1

An alternative macro approach to intervention development may start with implementing it straight into practice. For example, in this thesis this I could have first conducted Naturalistic Testing (Chapter 7), and dropped the Heuristic Evaluation / Cognitive Walkthrough (Chapter 5) and Usability Test (Chapter 6). This may have identified potential implementation issues sooner, such as EHR integration, though would have ignored the role of usability. As testament to its importance, it is clear to see how much PINGR’s interface and usability evolved during Chapters 5 and 6, and that its user-friendliness was deemed an advantage over existing e-A&F systems by participants in Chapter 7. I believe that if PINGR’s initial interface (see Figures 1 and 2 on p.110 and p.111 respectively) had been given to clinicians in clinical practice it would not only have limited adoption further

1 It is important to note that the integration of PINGR with existing EHR software was never planned during this thesis because the aim was to test whether EHR data could be used to develop a system like PINGR (i.e. that suggested actions to clinicians) before embarking on a lengthy integration process that involved technical infrastructure beyond my control (i.e. relying on computer code owned by a third party).
but also had potential impacts on patient safety [2]. Finally, implementing an intervention without sufficient prior theorising or modeling risks post-hoc rationalisation, where justification for a particular outcome may be lead to cognitive biases [16].

An additional explanation for the apparent ‘micro’ focus of this thesis is due to its ‘journal’ format. One of my aims was to maximise the number of publishable findings to share with the wider scientific community, and bolster my curriculum vitae. Consequently, I reported each step taken during the development process as a scientific paper. This may have given the impression that micro issues were focused on more, or that I believe it is more important than the role of wider context. This is not necessarily true, and belies the time I spent on these issues. For example, the qualitative metasynthesis (Chapter 4) includes a large component of the wider context (which informed Chapters 5 and 6), and Chapter 7 took the longest to plan and execute of all. The thesis may also be criticised for a relatively ‘late’ introduction of the views of the main ‘audience’ (GPs and service managers) in PINGR’s development process. However, this may be due to the presentation of chapter orders: GPs were the study subjects in the usability studies (Chapter 6), which were conducted early on (end of the first year of PhD studies). An alternative may have been to have clinician input prior to PINGR’s development, for example as a focus group. However, this would likely have not been publishable, and the approach is not recommended in the development of new technologies because customers often find it difficult to appraise new products without a working prototype [17].

On balance, I believe that micro approaches have more advantages over macro approaches, though on reflection it would be best to complete the early studies more quickly so that implementation problems can be identified and resolved sooner. This could be achieved either by: adopting a less rigorous approach (with the risk the studies are not publishable); or deploying enough resource (e.g. a large enough team). Given the limited resources available in a research setting (and particularly at doctoral level) this final option may only be viable in commercial settings.

8.5 Comparison to existing literature
There are few published examples of the systematic development and sequential evaluation of e-A&F systems as has been demonstrated in this thesis with PINGR. Here I discuss two examples ultimately evaluated in randomised controlled trials (RCTs) with differing fortunes, and discuss their findings and relevance to CP-FIT and other findings from this thesis: the data-driven quality improvement in primary care (DQIP) [18–21], and the cardiac rehabilitation decision support system (CARDSS) Online [22–27] tools. Clearly

\[^{2}\text{For example see the acceptance of incorrect recommendations made by PINGR to GPs in Chapter 6, which were subsequently corrected prior to Chapter 7.}\]
PINGR has not (yet) been evaluated in an RCT, so comparisons regarding potential effectiveness cannot be made.

DQIP was an e-A&F system supplemented with educational outreach visits. It was a recently developed and evaluated in Scotland to address medication safety issues in primary care regarding the prescription of non-steroidal anti-inflammatory medications (NSAIDs). It was based on extensive epidemiological work into the prevalence and effects of medication safety errors in primary care [21], though the rationale for the use of an e-A&F tool to address this issue appeared to be more pragmatic than theory-based. The inspiration for the initial design of DQIP is unclear, though it was initially evaluated in a mixed methods pilot study, which was included as part of the metasynthesis in Chapter 4 [18]. Results from this study influenced the development of the tool for the RCT by including a financial incentive, providing ongoing feedback on progress to GP practices via newsletters, and re-prompting review of patients whose high-risk prescribing was restarted after a decision to stop [18]. The trial was ultimately successful, demonstrating a reduction in high-risk NSAID prescribing from 3.7% to 2.2% in intervention practices (odds ratio [OR], 0.63; 95% confidence interval [CI], 0.57 to 0.68). The intervention was most successful in practices that were able to organise the work between staff members, and had the staff resources to dedicate to doing the work [19]. All findings can be explained by the CP-FIT moderating variables: Importance, Clinical education, Financial reward, Teamwork, Workflow fit, and Resources. Furthermore, in accordance with CP-FIT’s propositions, the main mechanism of action appeared to be via direct action using the e-A&F system to identify patients for action, whereas the financial incentives appeared help with initial practice recruitment rather than ongoing motivation, and the educational outreach visits re-enforced the importance of the topic rather than impart new knowledge to be used prospectively (‘Knowledge/Awareness mechanism’) [20]. Of particular relevance to this thesis (as highlighted in Chapter 7), the initial DQIP tool included additional quality indicators covering high-risk prescribing in patients with asthma and atrial fibrillation [18]. However, in the pilot study they were found to target patients who were often too complex to action because they were difficult to contact, or the decision regarding what action to take often lay with secondary care [18]. These findings echo those from Chapter 7 where PINGR highlighted patients for action that were medically complicated and often under the care of renal physicians (CP-FIT concepts: Complexity and Actionability). Interestingly, the response with the DQIP system was to address this issue by not addressing it, and simply not including these quality indicators in the RCT.

CARDSS Online was an e-A&F system developed and evaluated in The Netherlands for teams in cardiac rehabilitation supplemented by quality improvement support outreach visits. Its development arose from the failure of a CDS system built by the same team and
addressing the same clinical problems had failed to address organisational barriers to change [22]. Similar to DQIP, CARDSS was based on extensive work to develop and publish a set of 18 quality indicators in collaboration with an expert and patient panel [23]. It was informed by existing theory, and had limited naturalistic pilot testing prior to the RCT [23,24], which was ultimately negative [25]. Two process evaluations, one of which was a quantitative analysis of usage logs [26] and another was a concept-mapping study [27], found that potential reasons for the failure of effect related to participants disagreeing with the importance or relevance of quality indicators, their perception of not being able to improve, choosing to focus on quality indicators they thought were important even though they already scored highly, and low team commitment or organisational readiness [26,27].

Similar to DQIP, these findings can be explained by the following concepts in CP-FIT: Importance, Relevance, Controllability, and Teamwork. Unfortunately, a more in-depth qualitative study to explore the detail of these barriers was not undertaken to discuss in more depth.

Interestingly, neither DQIP nor CARDSS Online underwent formal usability testing prior to their implementation. This is clearly different to the approach with PINGR, and regarding the wider evidence base on health informatics interventions [2,28].

8.6 Implications

This thesis has implications that are theoretical, practical, policy, and methodological. From a theoretical point of view, the main findings arise from Chapters 3-6 and relate to arguments for combining A&F and CDS (Chapter 3), CP-FIT (Chapter 4), and the design of e-A&F interfaces (Chapters 5 and 6). As discussed in Chapter 3, the case that A&F and CDS systems should not be considered separate entities has implications for their development, design, and research. It suggests that characteristics from CDS should inform A&F design and vice versa: for example, A&F should recommend actions for users to take, and CDS should address team and organisational barriers to change. However, the ultimate conclusion is that these systems should not be either A&F or CDS, but be both together – i.e. ‘decision-supported feedback’. It also suggests that their research evidence and theories could be drawn upon to generate specific hypotheses to test in future research. CP-FIT clearly has theoretical recommendations regarding A&F design, specifically with regard to its three propositions, mediating variables, and set of design recommendations. Though given the arguments in Chapter 3, it is then reasonable to also test its relevance to CDS. This could be achieved by prospectively using it to design or evaluate CDS systems, or by retrospectively applying it existing literature by informing either meta-analyses of RCTs (e.g. variables in a meta-regression) or syntheses of the qualitative literature. This latter example could involve using CP-FIT as an initial coding framework for a metasynthesis of findings from qualitative evaluations of CDS systems,
with appropriate changes made as necessary [29]. Wider applications of CP-FIT may also be relevant for as noted in Chapter 3, arguments regarding A&F and CDS cross-fertilisation could be extended to other computerised interventions that facilitate clinician interpretation of patient data, such as risk prediction tools, and Chapter 4 notes that most quality improvement interventions follow plan-do-study-act cycles based on data collection and quality measurement, similar to A&F [30], and that CP-FIT’s propositions are wide enough to be relevant to most attempts to change health professional behaviour. Finally, the interface recommendations for e-A&F systems derived from Chapters 5 and 6 have theoretical implications for how these systems could be designed. These recommendations are relatively specific, so are unlikely to generalise more widely unlike CP-FIT.

With regards to practical implications, this thesis has demonstrated how its theoretical concepts can be applied in practice. Chapters 5-7 used hypotheses from Chapters 3 and 4 to design, implement and evaluate PINGR, and in doing so refined and adjusted their arguments (Table 1). For example, Chapter 6 refined the e-A&F design guidelines initially derived from recommendations in Chapters 3 and 5, and emerging findings from Chapter 4, using findings from usability tests with GPs (PINGR version 2). Specifically, presentation information on one page, and prioritisation of this information was a key change to how e-A&F interfaces could be designed. Chapter 7 used CP-FIT to both design and evaluate PINGR (version 3) in a naturalistic field study. Importantly, this demonstrated that some of CP-FIT’s theoretical interpretations (e.g. suggesting actions to users) can work in clinical practice, and are acceptable to health professionals. Furthermore, it presented specific examples of how key concepts in CP-FIT play out in practice (e.g. compatibility with existing pay-for-performance schemes) and how they may inter-relate, in addition to new variables that fit with its propositions (i.e. complexity). Finally, these practical implications suggested further modifications to PINGR that are required for version 4.

Policy implications of this thesis may relate to the Healthcare Quality Improvement Partnership (HQIP) and the Quality and Outcomes Framework (QOF) – or its successor [31]. As outlined in Chapter 1, HQIP is an organisation that promotes and supports the use of A&F in the NHS. They publish guidance for service providers and also run National Clinical Audits (NCAs) [32]. Recommendations from CP-FIT may be useful for HQIP to incorporate into their best practice guidance for A&F [33], which on a practical note they could implement themselves in any number of NCAs. To that end, I am already a co-applicant on an NIHR HS&DR Project (Optimising the outputs of National Clinical Audits to support organisations to improve the quality of care and clinical outcomes), which will use recommendations from CP-FIT to explore their applicability to NCAs covering a range
of NCAs. With regard to QOF, as also highlighted in Chapter 1, although it is not strictly considered A&F, its use of performance measures (PMs) means there are overlaps. Therefore, recommendations from CP-FIT may be relevant – an interesting example may include QOF’s Function. CP-FIT recommends that ‘A&F should convince the recipient that its purpose is to support them improve care rather than punish them’ (Chapter 4). The current perception of QOF, particularly with changing targets and withdrawal of funding, does not meet this criterion [31]. Furthermore, findings regarding the complexity of patients highlighted by PINGR in Chapter 7 rings true with criticisms of QOF [11]. Finally, given that QOF is based on using primary care EHR data, many of the design recommendations for e-A&F systems in Chapters 5 and 6 may be relevant for systems that report QOF results – particularly regarding the presentation of recommended actions, as currently this is absent from QOF.

From a methodological point of view, this thesis has implications regarding the use of qualitative literature in complex intervention design for health professionals, and e-A&F system development. Complex interventions targeting health professionals are often developed from literature reviews, qualitative studies, or expert consensus [1,34]. Metasyntheses of qualitative evidence are rarely used, though have been a key step PINGR’s development and evaluation. In contrast, metasyntheses are used more regularly in the development of complex interventions targeting patients [35]. The advantages of performing metasyntheses are the degree of detail they provide regarding a phenomenon [36], and the potential to derive generalisable findings that go beyond the included studies [37] as demonstrated by Chapter 4. However, a major drawback is the amount of resource they require to conduct in a robust manner. As described above, the delay in finishing Chapter 4 impacted on the timelines of this PhD. Therefore before employing this technique, consideration should be given to the resources available, and the level of abstraction and rigour required for the intended aims. With regards to e-A&F system development specifically, I followed Borycki et al.’s evaluation framework [2], which recommends a sequential series of usability evaluations prior to testing in real clinical settings. As mentioned in Chapter 2, I chose to miss one of the prescribed steps of Clinical Simulation, which involves testing the system in a representative context but without ‘live’ patients [38]. The rationale was that this would be unnecessary given e-A&F systems are not generally used in patient-facing situations. The study described in Chapter 7 supports this assertion: interviews with participants often involved them performing tasks with PINGR under my observation. To some extent represents a modified Clinical Simulation study, suggesting that formal delineation between Naturalistic Testing and Clinical Simulation may be unnecessary in e-A&F development.

8.7 Future research
After making the necessary improvements to PINGR recommended in Chapter 7, my future research will have four main workstreams. The first will seek a more efficient way to generate relevant actions directly from EHR data. Clinical actions generated by information systems (including those from PINGR) are usually generated by translating clinical guidelines into computer-programming language. However, this is labour intensive, does not accommodate changes in evidence (such as local adaptations to guidelines or updated evidence), focuses on patient-level (rather than organisational) action, and does not prioritise their relative importance. My experience during this PhD suggests that systematically studying the differences in primary care EHR data between patients who receive ‘optimal’ care versus those who do not, and differences in GP practices who consistently provide ‘optimal’ care versus those who do not, may provide useful clinical actions to complement those derived from clinical guidelines. I propose to explore the use of machine learning to discover relationships between clinical events recorded in patients’ EHRs and their likelihood of receiving ‘optimal’ primary care. Risks associated with this approach include the ability to generate clinically useful actions, and issues regarding the ethical implications of suggesting clinical actions based on data. I will address these risks by assessing their face validity against existing clinical evidence and using a panel of clinicians and patients to review the actions. For example, case-control study could be conducted on patients with hypertension matched according to age, sex, and GP practice where the outcome of interest is whether they have controlled/uncontrolled blood pressure [39]. An association rule learning algorithm [40] could be applied to their entire EHR data to find specific Read codes associated with improved care (e.g. ‘Advised about diet’). These results could be checked for agreement with recommendations in clinical guidelines (e.g. [39]) and with a panel of GPs and patients before translation into a clinical action (e.g. ‘Provide advice on diet to patient’). This action could then be introduced into PINGR, and refined based on feedback from users and their interactions with it.

The second workstream will focus on efficient and scientifically robust ways to continually optimise the Feedback of Audit data to health professionals. CP-FIT and findings from Chapter 7 provide multiple hypotheses regarding the most effective ways to design e-A&F systems, such as whether Benchmarking and Prioritisation improves recipients’ Perception and Intention to take action. I propose to test these hypotheses by conducting randomised “A/B tests” in PINGR – a technique commonly used by e-commerce companies such as Amazon and Google, where visitors to their websites are randomised to different versions of the webpage, and their activity is tracked in order to decide which version is most ‘effective’ [17,41]. For each hypothesis, different versions of the software interface will be created and tested (e.g. presence or absence of Benchmarking). Users in will be randomised to one of the designs, and their behaviour monitored remotely. Depending on the specific design changes, data will be collected on how they interact with
the system (e.g. *Intention* = agree with an action), or implement the actions suggested by studying EHR data. Following completion of the A/B test, the most ‘effective’ design will be tested against another hypothesis-informed design. In this sense, the A/B tests will be sequential in order to continually optimise the presentation of actions to health professionals over time. Achieving adequate statistical power may be problematic, though I will explore the use of novel approaches to randomisation that require smaller sample sizes compared to traditional designs, such as micro-randomisation [42] where users are randomised multiple times throughout the study, or multi-arm multi-stage (MAMS) designs where multiple interface designs are compared simultaneously [43]. The use of ‘intermediate’ outcomes that are more common (e.g. accessing a particular page) rather than less common outcomes such as implementation of the action could be used. Furthermore, I will maximise the number of opportunities for randomisation by directly integrating PINGR into the primary care EHR so it is accessed more frequently – discussions with EHR providers have already begun to explore this possibility. Users may also become annoyed by the different designs or experience learning effects, though differences in designs should be subtle enough to avoid conscious recognition by users.

The third workstream will investigate whether and how e-A&F systems adapt to different geographical settings. I have already explored the possibility of testing PINGR outside its current location, with different Clinical Commissioning Groups (CCGs) in the UK, in addition to locations in Canada. In particular, with regard to the previous two workstreams, it will provide insights into how their outputs differ between locations, and why (i.e. clinical actions generated from EHR data and optimised interface designs from A/B tests). This larger scale implementation will also enable effects on patient outcomes to be estimated more robustly to address the next phase in the MRC development framework [1]. Though rather than jump straight to an RCT, given the ongoing development to PINGR and associated costs of an RCT, a quasi-experimental design such as an interrupted time series analysis may be more appropriate [44].

Finally, the fourth workstream will explore the feasibility and acceptability of communicating clinical actions, such as those from PINGR, directly to patients. During my PhD, I presented my emerging findings to a number of patient and public involvement in research (PPI) groups. Each time there was enthusiasm for PINGR to send messages to patients. Furthermore, health professionals supported the idea when it was explored in interviews (Chapter 7). I have already made arrangements with Evergreen Life, a leading NHS supplier of patient smartphone applications to collaborate on this topic. I propose to conduct a series of co-design workshops with patients and clinicians [45]. The first will explore the perceived needs of patient-facing messages based on existing literature and empirical experience from this PhD, and will use techniques such as paper-prototyping
and story-boarding to refine ideas [46]. Using these outputs, a prototype smartphone application will be developed and presented at the second workshop attended by the same participants where possible, before changes made and presented at a final workshop. The resulting prototype applications will then be field-tested for 20 patients across five GP practices. Mixed methods evaluation will include interviews with 10 clinicians and all 20 patients, and measures of health system usage, to explore reactions to the application and potential unintended consequences. Initial discussions with PPI groups highlighted that patients may feel anxious or worried after receiving a message. I will address this by highlighting the potential messages patients could receive, and explaining their significance in the informed consent process of the study. Furthermore, the interviews will seek to understand these reactions in more detail. This will also go some way to addressing the shortcomings of this thesis in terms of not exploring patients' reactions to the use of EHR data for quality improvement.

8.8 Patient and public involvement in this project

Prior to starting this thesis, patients had input into its design and planning. I initially sought input from a local PPI group (Primary Care Research in Manchester Engagement Resource; PRIMER). I presented my initial plans for the project, and invited feedback on its design. The proposal was well-received: one member commented "knowing how to improve important aspects of care is important to help patients receive the best care available"; another member said "[this project is] of direct benefit to the patient and public... anything which reduces the chances of avoidable long-term damage is vital." Further specific comments were made regarding: ongoing PPI involvement, such as incorporating the views of carers; and potential ethical problems when identifying missed aspects of care in the real-world that may have contributed to adverse outcomes. Based on these comments, I amended the study design to actively seek PPI from carers, and developed a robust mechanism of communicating potential quality issues that may have adversely affected patient outcomes back to GP practices in the study if they arose during PINGR’s naturalistic testing in Chapter 7 (which they did not).

The ongoing PPI strategy I chose to use was the “Research Partner” (RP) approach [47]. RPs are a relatively underused PPI methodology where members of the public are actively involved in the ‘doing’ of research, including: designing studies, collecting and analysing data, writing reports, and disseminating results. This approach was chosen after I attended a PPI workshop and consulting with experts in the approach from both the UK (Bec Hanley) and Australia (Anne McKenzie) in 2013. RPs were originally developed for studies of large quantitative datasets to provide patient-relevant interpretation of abstract data [47], therefore I felt this may be an appropriate approach to deal with and interpret the data presented in PINGR. I subsequently obtained an NIHR Research Design Service
Patient and PPI bursary and recruited an RP before the project began. The RP was a member of PRIMER who had experience as both a patient and carer. I used the funds to reimburse the RP’s time and travel expenses. We initially met four times over the first few months of the project. Unfortunately, they became ill shortly after and could no longer continue in the project. I attempted to recruit a further RP though was unsuccessful because of the time commitments it required.

I subsequently altered my PPI approach to one with less time commitment for members of the public. I capitalised on the existing infrastructures of the organisations in which I was based (the Health e-Research Centre, and the Greater Manchester NIHR Patient Safety Translational Research Centre). Both had existing PPI groups, which met regularly. During the project I presented my preliminary research findings to, and received feedback from, these groups four times over the remaining time of the project. One comment from the groups was that the research was (necessarily) clinician-facing, and that as patients many of them would also like to be informed of the information within PINGR. During one meeting I explored these ideas in more depth: most would only be interested in information relevant to them (e.g. what steps could be taken to improve their health, rather than the wider performance of their GP practice), and many believed that this could be done via a smartphone application developed in collaboration with patients. This led to the development of the research project and funding application described in section 8.6 above.

The limitations of the RP approach I chose for this thesis was that it relied on the deep involvement of one member of the public. If this person were suddenly indisposed, it put the approach at risk. Furthermore, it limits the range of views that can be garnered. Although PPI is not intended to be based on representative patient samples like research, it is accepted that a range of views are more robust [48]. The risk of the PPI approach I ultimately adopted was that it could have been a cursory exercise where members of the public were informed of the progress of the project, but had little influence or input into its future directions. Because I was aware of this limitation, I took specific steps to address it by: actively seeking opinions from all members of the PPI groups during the meetings; and acknowledging and responding to all their comments – either verbally or in writing following the meetings.

In terms of future plans, the fourth workstream described in section 8.6 above outlines how the suggestions made by PPI groups will be transformed into reality. This may be seen as a hybrid approach, where PPI is embedded as a core part of the research methodology (i.e. in the co-design workshops). In addition, a virtual panel of patients and
clinicians will support the other workstreams by: providing guidance on clinical areas on which to focus, and helping interpret machine learning and A/B test studies.

8.9 Chapter conclusions
This chapter has started with a summary of the findings of the five results chapters of this thesis. It then provided a critique of the strengths and weaknesses of its overall approach, discussing issues including its iterative methods, use of theory, potential for bias, narrow focus on care quality, and lack of patient-based research. It also compared its findings to the existing literature on e-A&F system development with particular regard to two recent systems of varying fortunes (DQIP and CARDSS Online). Next, it highlighted the implications these findings in terms of theory, practice, policy, and methodology. Finally, it discussed plans for future research in four specific areas: action plan derivation from EHR data, e-A&F interface optimisation using A/B tests, adaptation to different geographical settings, and direct communication with patients regarding actions they could take to improve their care. The final chapter concludes this thesis by examining how well it has achieved its aims and objectives.

8.10 Chapter references
(accessed July 1, 2015).


[47] A. Mckenzie, H. Haines, Consumer and Community Participation in Health and Medical Research: Fact Sheet Series, The University of Western Australia, Crawley, 2011.

Chapter 9

Conclusion

The quality of primary care can have important impacts on population health. At the same time it collects large amounts of detailed electronic health record (EHR) data that have the potential to help improve its quality. The overall aim of this thesis was to make progress towards better use of primary care EHR data for the purposes of quality improvement. It addressed this by using electronic Audit and Feedback (e-A&F) as a vehicle, and explored its ability to suggest actions for improvement to health professionals. In doing so, the thesis has presented the development of a theoretical model (Chapter 3 and 4) and set of recommendations (Chapters 4, 5, and 6) that can be used to guide EHR data analysis and its communication to health professionals for quality improvement purposes (Research Objective 1). It subsequently used these models and recommendations to develop an e-A&F intervention for UK primary care, the unique defining feature of which was that it suggests improvement actions to users based on EHR data analysis (Chapters 5-7; Research Objective 2). It then implemented and evaluated the intervention to test the models and recommendations, and derived generalisable knowledge about using EHR data for quality improvement (Chapters 5-7; Research Objective 3). These findings suggest it is both feasible and acceptable to health professionals for EHR data to be used by e-A&F systems to suggest actions for quality improvement, and may have important impacts on patient care. When designing their interfaces, attention should be paid to how they summarise clinical performance, and present patient lists and detailed patient-level information. Where possible, all relevant data should be displayed on one page, and prioritised in terms of performance. Implementation of e-A&F interventions is heavily influenced by the available resources of the organisations intended to use it, its compatibility with existing workflows, and ability to take action based on their feedback. Furthermore, unresolved tensions exist regarding how they may deal with patient complexity. Policymakers should consider the relevance of these findings for National Clinical Audits and pay-for-performance initiatives.
Appendix 1: Search terms (Chapter 4)

MEDLINE (Ovid) 1946 to January Week 3 2015 – searched 27th January 2015

<table>
<thead>
<tr>
<th>Search term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (audit* adj3 feedback).tw.</td>
<td>1676</td>
</tr>
<tr>
<td>2 Clinical Audit/</td>
<td>833</td>
</tr>
<tr>
<td>3 Medical Audit/</td>
<td>14940</td>
</tr>
<tr>
<td>4 Nursing Audit/</td>
<td>2953</td>
</tr>
<tr>
<td>5 Dental Audit/</td>
<td>370</td>
</tr>
<tr>
<td>6 Management Audit/</td>
<td>2394</td>
</tr>
<tr>
<td>7 Benchmarking/</td>
<td>10374</td>
</tr>
<tr>
<td>8 &quot;Commission on Professional and Hospital Activities&quot;/</td>
<td>229</td>
</tr>
<tr>
<td>9 Feedback/</td>
<td>26135</td>
</tr>
<tr>
<td>10 Feedback, Psychological/</td>
<td>2302</td>
</tr>
<tr>
<td>11 Utilization Review/</td>
<td>7145</td>
</tr>
<tr>
<td>12 Drug Utilization Review/</td>
<td>3078</td>
</tr>
<tr>
<td>13 Concurrent Review/</td>
<td>376</td>
</tr>
<tr>
<td>14 Peer Review, Health Care/</td>
<td>1306</td>
</tr>
<tr>
<td>15 (audit or audits or auditing).tw.</td>
<td>24805</td>
</tr>
<tr>
<td>16 feedback.tw.</td>
<td>75736</td>
</tr>
<tr>
<td>17 (review adj3 record?).tw.</td>
<td>9996</td>
</tr>
<tr>
<td>18 chart review.tw.</td>
<td>20512</td>
</tr>
<tr>
<td>19 (practice data or hospital* data).tw.</td>
<td>3244</td>
</tr>
<tr>
<td>20 benchmark*.tw.</td>
<td>15495</td>
</tr>
<tr>
<td>21 or/2-20</td>
<td>186397</td>
</tr>
<tr>
<td>22 exp Health Personnel/</td>
<td>376175</td>
</tr>
<tr>
<td>23 exp Hospitals/</td>
<td>206858</td>
</tr>
<tr>
<td>24 exp Professional Practice/</td>
<td>219267</td>
</tr>
<tr>
<td>25 Family Practice/</td>
<td>59981</td>
</tr>
<tr>
<td>26 Professional Competence/</td>
<td>20499</td>
</tr>
<tr>
<td>27 Clinical Competence/</td>
<td>66711</td>
</tr>
<tr>
<td>28 Physician's Practice Patterns/</td>
<td>42310</td>
</tr>
<tr>
<td>29 Nurse's Practice Patterns/</td>
<td>1068</td>
</tr>
<tr>
<td>30 Dentist's Practice Patterns/</td>
<td>1702</td>
</tr>
<tr>
<td>31 Quality Assurance, Health Care/</td>
<td>49537</td>
</tr>
<tr>
<td>32 Quality of Health Care/</td>
<td>58038</td>
</tr>
<tr>
<td>33 ((health* personnel or health care personnel or physician? or doctor? or clinician? or nurse? or provider? or practitioner? or resident? or professional? or nursing or clinical) adj3 (skill or skills or behavior or behavior or competence)).tw.</td>
<td>27615</td>
</tr>
<tr>
<td>34 ((clinical or medical or dental or private or general or family or</td>
<td>164564</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>35</td>
<td>(practice pattern? or pattern of practice).tw.</td>
</tr>
<tr>
<td>36</td>
<td>(quality adj (assurance or improvement or control)).tw.</td>
</tr>
<tr>
<td>37</td>
<td>(healthcare quality or healthcare quality or quality of healthcare or quality of health care or quality of care).tw.</td>
</tr>
<tr>
<td>38</td>
<td>Performance.tw.</td>
</tr>
<tr>
<td>39</td>
<td>((influence* or chang*) adj3 (behavior* or behavior*)).tw.</td>
</tr>
<tr>
<td>40</td>
<td>or/22-39</td>
</tr>
<tr>
<td>41</td>
<td>21 and 40</td>
</tr>
<tr>
<td>42</td>
<td>1 or 41</td>
</tr>
<tr>
<td>43</td>
<td>(((&quot;semi-structured&quot; or semistructured or unstructured or informal or &quot;in-depth&quot; or indepth or &quot;face-to-face&quot; or structured or guide) adj3 (interview* or discussion* or questionnaire*)) or (focus group* or qualitative or ethnograph* or fieldwork or &quot;field work&quot; or &quot;key informant&quot;)).ti,ab. or interviews as topic/ or focus groups/ or narration/ or qualitative research/</td>
</tr>
<tr>
<td>44</td>
<td>42 and 43</td>
</tr>
<tr>
<td>45</td>
<td>Animals/</td>
</tr>
<tr>
<td>46</td>
<td>Humans/</td>
</tr>
<tr>
<td>47</td>
<td>45 not (45 and 46)</td>
</tr>
<tr>
<td>48</td>
<td>44 not 47</td>
</tr>
</tbody>
</table>

EMBASE (Ovid) 1980 to 2015 Week 04 – searched 27th January 2015

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(audit* adj3 feedback).tw.</td>
<td>2130</td>
</tr>
<tr>
<td>2</td>
<td>Medical Audit/</td>
<td>38426</td>
</tr>
<tr>
<td>3</td>
<td>Feedback System/</td>
<td>59035</td>
</tr>
<tr>
<td>4</td>
<td>Negative Feedback/</td>
<td>11073</td>
</tr>
<tr>
<td>5</td>
<td>Positive Feedback/</td>
<td>6141</td>
</tr>
<tr>
<td>6</td>
<td>&quot;utilization review&quot;/</td>
<td>65402</td>
</tr>
<tr>
<td>7</td>
<td>&quot;medical record review&quot;/</td>
<td>60421</td>
</tr>
<tr>
<td>8</td>
<td>(audit or audits or auditing).tw.</td>
<td>46609</td>
</tr>
<tr>
<td>9</td>
<td>feedback.tw.</td>
<td>98828</td>
</tr>
<tr>
<td>10</td>
<td>(review adj3 record?).tw.</td>
<td>15290</td>
</tr>
<tr>
<td>11</td>
<td>chart review.tw.</td>
<td>38408</td>
</tr>
<tr>
<td>12</td>
<td>(practice data or hospital* data).tw.</td>
<td>5465</td>
</tr>
<tr>
<td>13</td>
<td>benchmark*.tw.</td>
<td>22442</td>
</tr>
<tr>
<td>14</td>
<td>or/2-13</td>
<td>352975</td>
</tr>
<tr>
<td>15</td>
<td>exp Health Care Personnel/</td>
<td>961050</td>
</tr>
<tr>
<td>16</td>
<td>exp Hospital/</td>
<td>719419</td>
</tr>
<tr>
<td>17</td>
<td>exp Professional Practice/</td>
<td>271288</td>
</tr>
<tr>
<td>18</td>
<td>Professional Competence/</td>
<td>24700</td>
</tr>
<tr>
<td>19</td>
<td>Nursing Competence/</td>
<td>544</td>
</tr>
</tbody>
</table>
Clinical Competence/
Health Care Quality/
Quality Control/

((health* personnel or health care personnel or physician? or doctor? or clinician? or nurse? or provider? or practitioner? or resident? or professional? or nursing or clinical) adj3 (skill or skills or behavior or behavior or competence)).tw.

((clinical or medical or dental or private or general or family or professional or hospital?) adj practice?).tw.

(practice pattern? or pattern of practice).tw.

(quality adj (assurance or improvement or control)).tw.

(healthcare quality or healthcare quality or quality of healthcare or quality of health care or quality of care).tw.

quality adj (assurance or improvement or control)).tw.

performance.tw.

((influenc* or chang*) adj3 (behavior* or behavior*).tw.

or/15-29

14 and 30

1 or 31

limit 32 to exclude medline journals

(interview: or qualitative).tw. or exp health care organization/

and

TI (audit* and feedback ) or AB (audit* and feedback )

S4 2 S40 AND S41 2,080

S4 1 ((MH "study design+" not MM "study design+") or MH "attitude" or (MH "interviews+" not MM "interviews+")) 801,586

S4 0 S39 (Limiters - Exclude MEDLINE records) 8,029

S3 9 S37 or S38 24,058

S3 8 S13 and S36 23,681

S3 7 TI (audit* and feedback) or AB (audit* and feedback) 882

S3 6 S14 or S15 or S16 or S17 or S18 or S19 or S20 or S21 or S22 or S23 or S24 or S25 or S26 or S27 or S28 or S29 or S30 or S31 or S32 or S33 or S34 or S35 763,238

S3 5 TI (influenc* N3 behavior* or influenc* N3 behavior* or chang* N3 behavior* or chang* N3 behavior* ) or AB (influenc* N3 behavior* or influenc* N3 behavior* orchang* N3 behavior* or chang* N3) 13,762

CINAHL Plus (Ebsco) 1937 to present – searched 27th January 2015
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S3 4</td>
<td>TI performance or AB performance</td>
<td>79,068</td>
</tr>
<tr>
<td>S3 3</td>
<td>TI (&quot;health care quality&quot; or &quot;healthcare quality&quot; or quality W1 healthcare or quality W2 care) or AB (&quot;health care quality&quot; or &quot;healthcare quality&quot; or quality W1 healthcare or quality W2 care)</td>
<td>23,252</td>
</tr>
<tr>
<td>S3 2</td>
<td>TI (quality W0 assurance or quality W0 improvement or quality W0 control) or AB (quality W0 assurance or quality W0 improvement or quality W0 control)</td>
<td>14,113</td>
</tr>
<tr>
<td>S3 1</td>
<td>TI practice N1 pattern* or AB practice N1 pattern*</td>
<td>1,665</td>
</tr>
<tr>
<td>S3 0</td>
<td>TI (clinical W0 practice* or medical W0 practice* or dental W0 practice* or private W0 practice* or general W0 practice* or family W0 practice* or professional W0 practice* or hospital* W0 practice*) or AB (clinical W0 practice* or medical W0 practice* or dental W0 practice* or private W0 practice* or general W0 practice* or family W0 practice* or professional W0 practice* or hospital* W0 practice*)</td>
<td>52,684</td>
</tr>
<tr>
<td>S2 9</td>
<td>TI (&quot;health personnel&quot; N3 competence or &quot;healthcare personnel&quot; N3 competence or &quot;health care personnel&quot; N3 competence or physician N3 competence or physicians N3 competence or doctor N3 competence or doctors N3 competence or clinician N3 competence or clinicians N3 competence or nurse N3 competence or nurses N3 competence or provider N3 competence or providers N3 competence or practitioner N3 competence or practitioners N3 competence or resident N3 competence or residents N3 competence or professional N3 competence or professionals N3 competence or nursing N3 competence or clinical N3 competence) or AB (&quot;health personnel&quot; N3 competence or &quot;healthcare personnel&quot; N3 competence or &quot;health care personnel&quot; N3 competence or physician N3 competence or physicians N3 competence or doctor N3 competence or doctors N3 competence or clinician N3 competence or clinicians N3 competence or nurses N3 competence or provider N3 competence or providers N3 competence or practitioner N3 competence or practitioners N3 competence or resident N3 competence or residents N3 competence or professional N3 competence or professionals N3 competence or nursing N3 competence or clinical N3 competence)</td>
<td>3,279</td>
</tr>
<tr>
<td>S2 8</td>
<td>TI (&quot;health personnel&quot; N3 behavior or &quot;healthcare personnel&quot; N3 behavior or &quot;health care personnel&quot; N3 behavior or physician N3 behavior or physicians N3 behavior or doctor N3 behavior or doctors N3 behavior or clinician N3 behavior or clinicians N3 behavior or nurse N3 behavior or nurses N3 behavior or provider N3 behavior or providers N3 behavior or practitioner N3 behavior or practitioners N3 behavior or resident N3 behavior or residents N3 behavior or professional N3 behavior or professionals N3 behavior or nursing N3 behavior or clinical N3 behavior) or AB (&quot;health personnel&quot; N3 behavior or &quot;healthcare personnel&quot; N3 behavior or &quot;health care personnel&quot; N3 behavior or physician N3 behavior or physicians N3 behavior or doctor N3 behavior or doctors N3 behavior or clinician N3 behavior or clinicians N3 behavior or nurse N3 behavior or nurses N3 behavior or provider N3 behavior or providers N3 behavior or practitioner N3 behavior or practitioners N3 behavior or resident N3 behavior or residents N3 behavior or professional N3 behavior or professionals N3 behavior or nursing N3 behavior or clinical N3 behavior)</td>
<td>2,703</td>
</tr>
<tr>
<td>S2 7</td>
<td>providers N3 behavior or practitioner N3 behavior or practitioners N3 behavior or resident N3 behavior or residents N3 behavior or professional N3 behavior or professionals N3 behavior or nursing N3 behavior or clinical N3 behavior )</td>
<td></td>
</tr>
<tr>
<td>S2 6</td>
<td>T1 ( “health personnel” N3 behavior or “healthcare personnel” N3 behavior or “health care personnel” N3 behavior or physician N3 behavior or doctors N3 behavior or clinician N3 behavior or clinicians N3 behavior or nurse N3 behavior or nurses N3 behavior or provider N3 behavior or practitioners N3 behavior or resident N3 behavior or residents N3 behavior or professional N3 behavior or professionals N3 behavior or nursing N3 behavior or clinical N3 behavior ) or AB ( “health personnel” N3 behavior or “healthcare personnel” N3 behavior or “health care personnel” N3 behavior or physician N3 behavior or doctors N3 behavior or clinicians N3 behavior or residents N3 behavior or professional N3 behavior or practitioners N3 behavior or residents N3 behavior or professional N3 behavior or professionals N3 behavior or nursing N3 behavior or clinical N3 behavior )</td>
<td></td>
</tr>
<tr>
<td>S2 5</td>
<td>T1 ( “health personnel” N3 skills or “healthcare personnel” N3 skills or “health care personnel” N3 skills or physician N3 skills or physicians N3 skills or doctor N3 skills or doctors N3 skills or clinician N3 skills or clinicians N3 skills or nurse N3 skills or nurses N3 skills or provider N3 skills or providers N3 skills or practitioner N3 skills or practitioners N3 skills or resident N3 skills or residents N3 skills or professional N3 skills or professionals N3 skills or nursing N3 skills or clinical N3 skills ) or AB ( “health personnel” N3 skills or “healthcare personnel” N3 skills or “health care personnel” N3 skills or physician N3 skills or doctors N3 skills or clinicians N3 skills or residents N3 skills or professional N3 skills or practitioners N3 skills or resident N3 skills or residents N3 skills or professional N3 skills or professionals N3 skills or nursing N3 skills or clinical N3 skills )</td>
<td></td>
</tr>
<tr>
<td>S2 4</td>
<td>(MH “Quality of Nursing Care”)</td>
<td></td>
</tr>
<tr>
<td>S2 3</td>
<td>(MH “Quality of Health Care”)</td>
<td></td>
</tr>
</tbody>
</table>

266
<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MH “Quality Assurance”</td>
<td>MH “Prescribing Patterns”</td>
<td>MH “Practice Patterns”</td>
<td>MH “Nursing Skills”</td>
<td>MH “Clinical Competence”</td>
<td>MH “Professional Competence”</td>
<td>MH “Professional Practice+”</td>
<td>MH “Hospitals+”</td>
<td>MH “Health Personnel+”</td>
<td>S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 or S9 or S10 or S11 or S12</td>
<td>TI benchmark* or AB benchmark*</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>16,265</td>
<td>2,899</td>
<td>9,901</td>
<td>2,838</td>
<td>25,035</td>
<td>10,651</td>
<td>185,938</td>
<td>77,109</td>
<td>372,615</td>
<td>53,381</td>
<td>4,333</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>933</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>362</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5,175</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3,245</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13,672</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25,013</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,657</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7,187</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5,172</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>853</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11,752</td>
</tr>
</tbody>
</table>
Appendix 2: Example data extraction form

<table>
<thead>
<tr>
<th>STUDY DETAILS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study ID</td>
</tr>
<tr>
<td>Date data extracted</td>
</tr>
<tr>
<td>Study Citation</td>
</tr>
<tr>
<td>Setting</td>
</tr>
<tr>
<td>Aim statement / research question of study (can be taken from the abstract)</td>
</tr>
<tr>
<td>Main findings / codes / framework / theory</td>
</tr>
<tr>
<td>Main changes to model (if any)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RESEARCH PARAMETERS [1]</th>
<th>#</th>
<th>COREQ reporting criteria</th>
<th>Describe if reported – otherwise “N”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>Interviewer/facilitator: Which author/s conducted the interview or focus group?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Credentials: What were the researcher’s credentials? E.g. PhD, MD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Occupation: What was their occupation at the time of the study?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Gender: Was the researcher male or female?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Experience and training: What experience or training did the researcher have?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Relationship established: Was a relationship established prior to study commencement?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Participant knowledge of the interviewer: What did the participants know about the researcher? e.g. personal goals, reasons for doing the research</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Interviewer characteristics: What characteristics were reported about the interviewer/facilitator? e.g. Bias, assumptions, reasons and interests in the research topic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Methodological orientation and Theory: What methodological orientation was stated to underpin the study? e.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis [NB – leave blank if do not explicitly mention]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Sampling: How were participants selected? e.g. purposive, convenience, consecutive, snowball</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Method of approach: How were participants approached? e.g. face-to-face, telephone, mail, email</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Sample size: How many participants were in the study?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Non-participation: How many people refused to participate or dropped out? Reasons?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Setting of data collection: Where was the data collected? e.g. home, clinic, workplace</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Presence of non-participants: Was anyone else present besides the participants and researchers?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Description of sample: What are the important characteristics of the sample? e.g. demographic data, date</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Interview guide: Were questions, prompts, guides provided by the authors? Was it pilot tested?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Repeat interviews: Were repeat interviews carried out? If yes, how many?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Audio/visual recording: Did the research use audio or visual recording to collect the data?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Field notes: Were field notes made during and/or after the interview or focus group?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Duration: What was the duration of the interviews or focus group?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Data saturation: Was data saturation discussed?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Transcripts returned: Were transcripts returned to participants for comment and/or correction?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Number of data coders: How many data coders coded the data?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Description of the coding tree: Did authors provide a description of the coding tree?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Derivation of themes: Were themes identified in advance or derived from the data?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Software: What software, if applicable, was used to manage the data?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Participant checking: Did participants provide feedback on the findings?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Quotations presented: Were participant quotations presented to illustrate the themes / findings? Was each quotation identified? e.g. participant number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data and findings consistent: Was there consistency between the data presented and the findings?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Clarity of major themes: Were major themes clearly presented in the findings?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Clarity of minor themes: Is there a description of diverse cases or discussion of minor themes?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**INTERVENTION**

| Has the intervention been judged a success? |   |
| How? |   |
| How was the intervention implemented / delivered? Use BCT Taxonomy where possible [2] |   |

<table>
<thead>
<tr>
<th>Modifiable elements [3] (*plus additional)</th>
<th>Describe (briefly) or Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was the feedback given to an individual, a group or both?</td>
<td></td>
</tr>
<tr>
<td>2. Was it given to the person in whom the practice change was desired (eg, healthcare provider vs hospital administrator)</td>
<td></td>
</tr>
<tr>
<td>3. Was there feedback about the processes of care (eg, rate of antibiotic prescription)</td>
<td></td>
</tr>
<tr>
<td>4. Was there feedback about patient outcomes</td>
<td></td>
</tr>
<tr>
<td>5. Was there feedback about something other than processes of care or patient outcomes (if yes, specified)</td>
<td></td>
</tr>
<tr>
<td>6. Was the feedback about individual provider performance</td>
<td></td>
</tr>
<tr>
<td>7. Was the feedback about the performance of the provider group</td>
<td></td>
</tr>
<tr>
<td>8. Was the feedback about individual patient cases</td>
<td></td>
</tr>
<tr>
<td>9. Was the feedback about an aggregate of patient cases</td>
<td></td>
</tr>
<tr>
<td>10. Did the feedback identify a specific behaviour(s) to be changed</td>
<td></td>
</tr>
<tr>
<td>11. What was the comparison provided in the feedback (specified)</td>
<td></td>
</tr>
<tr>
<td>12. Were graphical elements included in the feedback</td>
<td></td>
</tr>
<tr>
<td>13. What was the lag between the time of the</td>
<td></td>
</tr>
</tbody>
</table>
14. What rationale was given for using A&F (specified) [e.g. What are its theoretical underpinnings? (Either explicit references to theory, or implied mechanisms of change by authors)]

15. Was the feedback given face to face

16. Were providers explicitly asked to consider the implications the A&F had for their practice [e.g. action planning]

17. What was the total number of times the feedback was given (specified).

What was the task being studied?*

How was the audit performed?*

What data was used to perform the audit?*

### QUALITY APPRAISAL [4]

<table>
<thead>
<tr>
<th>Broad areas</th>
<th>Criteria</th>
<th>Each criterion (Y / N / N/A)</th>
<th>Overall judgment (Y / N / N/A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear statement of, and rationale for, research question/aims/purposes</td>
<td>Clarity of focus demonstrated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study thoroughly contextualised by existing literature</td>
<td>Evidence of systematic approach to literature review, location of literature to contextualise the findings, or both</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method/design apparent, and consistent with research intent</td>
<td>Rationale given for use of qualitative design</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussion of epistemological/ontological grounding</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rationale explored for specific qualitative method (e.g. ethnography, grounded theory, phenomenology)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussion of why particular method chosen is most appropriate/sensitive/relevant for research question/aims</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Setting appropriate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data collection strategy apparent and appropriate</td>
<td>Were data collection methods appropriate for type of data required and for specific qualitative method?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Were they likely to capture the complexity/diversity of experience and illuminate context in sufficient detail?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Was triangulation of data sources used if appropriate?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample and sampling method appropriate</td>
<td>Selection criteria detailed, and description of how sampling was undertaken</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Justification for sampling strategy given</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thickness of description likely to be achieved from sampling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Any disparity between planned and actual sample explained</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytic approach appropriate</td>
<td>Approach made explicit (e.g. Thematic distillation, constant comparative method, grounded theory)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Was it appropriate for the qualitative method chosen?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Was data managed by software package or by hand and why?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion of how coding systems/conceptual frameworks evolved</td>
<td>How was context of data retained during analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence that the subjective meanings of participants were portrayed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence of more than one researcher involved in stages if appropriate to epistemological/theoretical stance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Did research participants have any involvement in analysis (e.g. member checking)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence provided that data reached saturation or discussion/rationale if it did not</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence that deviant data was sought, or discussion/rationale if it was not</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context described and taken account of in interpretation</td>
<td>Description of social/physical and interpersonal contexts of data collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence that researcher spent time ‘dwelling with the data’, interrogating it for competing/alternative explanations of phenomena</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear audit trail given</td>
<td>Sufficient discussion of research processes such that others can follow ‘decision trail’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data used to support interpretation</td>
<td>Extensive use of field notes entries/verbatim interview quotes in discussion of findings</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clear exposition of how interpretation led to conclusions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Researcher reflexivity demonstrated</td>
<td>Discussion of relationship between researcher and participants during fieldwork</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Demonstration of researcher’s influence on stages of research process</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence of self-awareness/insight</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation of effects of the research on researcher</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence of how problems/complications met were dealt with</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demonstration of sensitivity to ethical concerns</td>
<td>Ethical committee approval granted</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clear commitment to integrity, honesty, transparency, equality and mutual respect in relationships with participants</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence of fair dealing with all research participants</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recording of dilemmas met and how resolved in relation to ethical issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation of how autonomy, consent, confidentiality, anonymity were managed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance and transferability evident</td>
<td>Sufficient evidence for typicality specificity to be assessed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Analysis interwoven with existing theories and other relevant explanatory literature drawn from similar settings and studies</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussion of how explanatory propositions/emergent theory may fit other contexts</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Limitations/weaknesses of study clearly outlined</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clearly resonates with other knowledge and experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Results/conclusions obviously supported by evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interpretation plausible and ‘makes sense’</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provides new insights and increases understanding</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Significance for current policy and practice outlined</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assessment of value/empowerment for</td>
<td></td>
<td></td>
</tr>
<tr>
<td>participants</td>
<td>Outlines further directions for investigation</td>
<td>Comment on whether aims/purposes of research were achieved</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### MAIN STRENGTHS/LIMITATIONS

<table>
<thead>
<tr>
<th>Main strengths of study (according to reviewer – i.e. you)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Main strengths of study (according to study author)</td>
<td></td>
</tr>
<tr>
<td>Main limitations of study (according to reviewer – i.e. you)</td>
<td></td>
</tr>
<tr>
<td>Main limitations of study (according to study author)</td>
<td></td>
</tr>
</tbody>
</table>

### SUPPLEMENTARY SEARCHES

<table>
<thead>
<tr>
<th>Papers in bibliography that may be relevant (out of how many?)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Out of how many?</td>
<td></td>
</tr>
<tr>
<td>- New papers to synthesise or for background discussion</td>
<td></td>
</tr>
<tr>
<td>Papers in citation searching that may be relevant (using WoS)</td>
<td></td>
</tr>
<tr>
<td>- Out of how many?</td>
<td></td>
</tr>
<tr>
<td>- New papers to synthesise or for background discussion</td>
<td></td>
</tr>
<tr>
<td>Papers in related article searching that may be relevant (Using WoS – limit to first 100 results)</td>
<td></td>
</tr>
<tr>
<td>- Out of how many?</td>
<td></td>
</tr>
<tr>
<td>- New papers to synthesise or for background discussion</td>
<td></td>
</tr>
</tbody>
</table>

### MEMOS

### References


### Appendix 3: Study details

<table>
<thead>
<tr>
<th>ID</th>
<th>Citation</th>
<th>Country</th>
<th>Setting</th>
<th>A&amp;F topic(s)</th>
<th>Target A&amp;F recipient(s)</th>
<th>Number of participants</th>
<th>Brief description of A&amp;F</th>
<th>Modifiable elements reported (out of 17)</th>
<th>Quality appraisal (out of 12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation</td>
<td>England</td>
<td>Primary care</td>
<td>Diabetes</td>
<td>Physicians Nurses</td>
<td>A&amp;F conducted by external group to primary care practice.</td>
<td>11</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>---------------</td>
<td>---------</td>
<td>-------------------</td>
<td>--------------------------------------------------------</td>
<td>-----</td>
<td>----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. Ivers, J. Barnsley, R. Upshur, K. Tu, B. Shah, J. Grimshaw, et al., Canada</td>
<td>Canada</td>
<td>Primary care</td>
<td>Chronic care</td>
<td>Physicians</td>
<td>12</td>
<td>Mailed feedback conducted by</td>
<td>16</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>S15</td>
<td>E. Nouwens, J. van Lieshout, M. Wensing, Determinants of impact of a practice accreditation program in primary care: a qualitative study., BMC Fam. Pract. 16 (2015) 78.</td>
<td>Netherlands</td>
<td>Primary care Chronic care, health care structure, Physicians Nurses</td>
<td>Netherlands</td>
<td>Written feedback provided by external team, which is discussed with a trained observer with the whole practice</td>
<td>14</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td>Study</td>
<td>Country</td>
<td>Setting</td>
<td>Feedback Type</td>
<td>Stakeholders</td>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>---------------</td>
<td>--------------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Page</td>
<td>Authors</td>
<td>Title</td>
<td>Country</td>
<td>Setting</td>
<td>Role</td>
<td>Strategy and Example</td>
<td>Year</td>
<td>Paywall</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>------</td>
<td>---------------------</td>
<td>------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>Study Title</td>
<td>Country</td>
<td>Setting</td>
<td>Professionals Involved</td>
<td>Findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------</td>
<td>-----------------</td>
<td>------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1271</td>
<td>M. Cameron, G. Penney, G. McLeer, M. Sharon, A. Walker, Impact on Maternity Professionals</td>
<td>Scotland</td>
<td>Hospital</td>
<td>Obstetrics, Physicians, Midwives</td>
<td>Three different interventions delivered by research team:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Authors</td>
<td>Title</td>
<td>Countries</td>
<td>Healthcare Provider Levels</td>
<td>Health Indicators</td>
<td>Feedback Type</td>
<td>Feedback Action</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>-------</td>
<td>------------</td>
<td>---------------------------</td>
<td>------------------</td>
<td>---------------</td>
<td>-----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>J.H.M. Veillard, M.L. Schiøtz, A.L. Guisset, A.D. Brown, N.S. Klazinga</td>
<td>The PATH project in eight European countries: An evaluation</td>
<td>Belgium, Estonia, France, Germany, Hungary, Poland, Slovakia and Slovenia</td>
<td>Obstetrics, Antimicrobial stewardship, Mortality, Admission rates, Staff experience, Health care structures, Patient experience</td>
<td>Physicians, nurses, managers</td>
<td>20</td>
<td>Multinational project. Feedback was calculated and results fed back to participating hospital managers using a hospital performance dashboard and individual indicators.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Printed feedback (anonymized service-level data); 2. + action planning letter; 3. 2 + Facilitated action planning face-to-face.
<table>
<thead>
<tr>
<th>ID</th>
<th>Authors</th>
<th>Title</th>
<th>Country</th>
<th>Setting</th>
<th>Intervention</th>
<th>Sample Size</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>281</td>
<td>O. Groene, N. Klazinga, V. Kazandjian, P. Lombrail, P. Bartels</td>
<td>The World Health Organization Performance Assessment Tool for Quality Improvement in Hospitals (PATH): an analysis of the pilot implementation in 37 hospitals.,</td>
<td>Belgium, Canada, Denmark, France, Slovakia, South Africa</td>
<td>Hospital</td>
<td>Obstetrics, Antimicrobial stewardship, Mortality, Admission rates, Staff experience, Health care structures, Patient experience</td>
<td>Physicians, nurses, managers</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Similar to Veillard et al. 2013. Hospitals collected data themselves then reported back to WHO.</td>
</tr>
<tr>
<td>2794</td>
<td>V. Grando, M. Rantz, M. Maas</td>
<td>Nursing home staff’s views on quality improvement interventions: a follow up study,</td>
<td>US</td>
<td>Nursing home</td>
<td>Nursing - general</td>
<td>Nurses, managers</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Two different interventions. Group 1: quarterly reports and educational workshop. Group 2: Same plus on-site / phone support from advanced practice nurses.</td>
</tr>
<tr>
<td>4222</td>
<td>H. Kristensen, L. Hounsgaard</td>
<td>Evaluating the impact of audits and feedback as methods for implementation of evidence in stroke rehabilitation.,</td>
<td>Denmark</td>
<td>Hospital</td>
<td>Stroke</td>
<td>Occupational therapists</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Quarterly feedback presented both orally and in writing as tables, which were handed out to the participants at</td>
</tr>
<tr>
<td>5235</td>
<td>M.M. Meijers, R.J.G. Halfens, D.M. Mijnarends, H. Mostert, J.M.G.A. Schols, A feedback system to improve the quality of nutritional care., Nutrition. 29 (2013) 1037–1041.</td>
<td>Netherlands</td>
<td>Nursing home</td>
<td>Nutritional care</td>
<td>Nurses, managers</td>
<td>30</td>
<td>Computerised feedback programme. Staff collect their own data using objective assessments e.g. patient questionnaires / assessments</td>
</tr>
<tr>
<td>581</td>
<td>G. Harvey, A. Kitson, Achieving improvement through quality: an evaluation of key factors in the implementation process, J. Adv. Nurs. 24 (1996) 185–195.</td>
<td>England</td>
<td>Hospital</td>
<td>Nursing - general</td>
<td>Nurses</td>
<td>14</td>
<td>Three interventions: 1) Monitor - Care processes and structure; Acute medicine and surgery; Peer assessment? Sample of patients; 250 criteria - y/n responses 2) QualPacs - 2 assessors - observe care and audit medical records; 68 criteria 3) DySSy - no pre-defined criteria; 3 stage process of defining quality, measuring, and taking action; continuous quality improvement; small groups of practitioners 4-6 work</td>
</tr>
<tr>
<td>------</td>
<td>-------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
<td>----------</td>
<td>---------------</td>
<td>--------</td>
<td>-----</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>7025</td>
<td>N. Shepherd, T.J. Meehan, F. Davidson, T. Stedman, An evaluation of a benchmarking initiative in extended treatment mental health services., Aust. Health Rev. 34 (2010) 328–333.</td>
<td>Australia</td>
<td>mental health - inpatient</td>
<td>Mental health</td>
<td>Nurses Managers</td>
<td>84</td>
<td>Written and face-to-face feedback provided by external team. A State-wide forum helped recipients review their findings, and discuss areas that had the potential for improvement.</td>
</tr>
<tr>
<td>7049</td>
<td>K. Siddiqi, J. Newell, What were the lessons learned from implementing clinical audit in Latin America?, Clin. Gov. An Int. J. (2009) 21–22. doi:10.1108/14777270910976157</td>
<td>Bolivia, Peru, Cuba</td>
<td>Primary care</td>
<td>TB diagnosis</td>
<td>Physicians Nurses</td>
<td>43</td>
<td>A&amp;F organised by a committee across different organisations. Results were fed back to the health professionals in each committee</td>
</tr>
<tr>
<td>8249</td>
<td>Z. Paskins, H. John, A. Hassell, I. Rowe, The perceived advantages and disadvantages of regional audit: a qualitative study, Clin. Gov. An Int. J. (2011).</td>
<td>England</td>
<td>Hospital outpatient</td>
<td>Rheumatology</td>
<td>Physicians, nurses, managers</td>
<td>7</td>
<td>Six different audits based on national clinical guidelines conducted locally coordinated by a central committee. Each audit was led by a consultant member of the committee coordinated by a group of trainees and had information technology support to design proformas and collate data from a clinical audit department within the region. Each unit received details of its individual performance with anonymised results from other units</td>
</tr>
<tr>
<td>516</td>
<td>M.B. Boyce, J.P. Browne, J. Greenhalgh, Surgeon’s</td>
<td>Ireland</td>
<td>Hospital</td>
<td>Orthopaed</td>
<td>Orthopaed</td>
<td>11</td>
<td>Patient-reported</td>
</tr>
<tr>
<td>ID</td>
<td>Authors</td>
<td>Title</td>
<td>Country</td>
<td>Hospital</td>
<td>Cancer Surgery</td>
<td>Occupation</td>
<td>Additional Info</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>---------</td>
<td>----------</td>
<td>----------------</td>
<td>-----------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>S71</td>
<td>M. Gort, M. Broekhuis, G. Regts</td>
<td>How teams use indicators for quality improvement - a multiple-case study on the use of multiple indicators in multidisciplinary breast cancer teams.</td>
<td>Netherlands</td>
<td>Hospital</td>
<td>Cancer surgery</td>
<td>Surgeons, nurses, radiologists, internal medicine specialists</td>
<td>External project facilitators to conduct A&amp;F based on national guidelines. Data for each indicator were registered in a standardized system. All participants were invited to share their experiences and project results during two working conferences. Between and after these conferences, the individual teams could organize additional team meetings to discuss results, progress and any other topic.</td>
</tr>
<tr>
<td>S104</td>
<td>V.L. Payne, S.J. Hysong</td>
<td>Model depicting aspects of audit and feedback that impact physicians' care</td>
<td>US</td>
<td>Primary care</td>
<td>Chronic care</td>
<td>Physicians</td>
<td>See Powell et al. 2014</td>
</tr>
<tr>
<td>ID</td>
<td>Title</td>
<td>Authors</td>
<td>Country</td>
<td>Setting</td>
<td>Stakeholders</td>
<td>Participants</td>
<td>Study Details</td>
</tr>
<tr>
<td>-----</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
<td>---------</td>
<td>------------------</td>
<td>-------------------------------------</td>
<td>----------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>

**Notes:**
- **US:** United States
- **Primary care:** Health care provided to individuals who are not hospitalized.
- **Chronic care:** Long-term care for conditions that are not expected to resolve or improve.
- **Physicians:** Medical doctors who diagnose and treat patients.
- **Nurses:** Health care professionals who provide care under the direction of a physician.
- **Managers:** Individuals responsible for overseeing and coordinating the operations of an organization.
- **Physicians, nurses, managers:** Stakeholders involved in providing and managing health care services.
- **Findings based on California intervention only:** Findings focus on the California intervention only.
- **Feedback on performance provided by physician organization:** Feedback on performance provided by the physician organization.
<table>
<thead>
<tr>
<th>Study ID</th>
<th>Authors</th>
<th>Title</th>
<th>Country</th>
<th>Setting</th>
<th>Data Source</th>
<th>Data Collection</th>
<th>Data Feedback</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>S112</td>
<td>R. Mcdonald, M. Roland</td>
<td>Pay for Performance in Primary Care in England and California: Comparison of Unintended Consequences</td>
<td>UK</td>
<td>Primary care</td>
<td>Chronic care, preventive medicine</td>
<td>Physicians</td>
<td>20</td>
<td>Linked to pay-for-performance. Derived from mainly paper medical records. Data collected by third parties.</td>
</tr>
<tr>
<td>S118</td>
<td>L.M. Chadwick, A. Macphail, J.E. Ibrahim, L. Mcauliffe, S. Koch, Y. Wells</td>
<td>Senior staff perspectives of a quality indicator program in public sector residential aged care services: A qualitative cross-sectional study in Victoria, Australia</td>
<td>Australia</td>
<td>Nursing home</td>
<td>Nursing - general</td>
<td>Nurses, Managers</td>
<td>56</td>
<td>Data collected by recipients and sent centrally. Fed back as written reports quarterly.</td>
</tr>
<tr>
<td>S14</td>
<td>Z. Landis-levi, R. Manjomo, O.J. Gadabu, M. Kam, B.N. Simwaka</td>
<td>Electronic medical record data automatically</td>
<td>Malawi</td>
<td>Hospital</td>
<td>HIV/AIDS</td>
<td>Physicians</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Authors</td>
<td>Year</td>
<td>Setting</td>
<td>Feedback Type</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>------</td>
<td>----------</td>
<td>---------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S120</td>
<td>S.L. Zickmund, et al.</td>
<td>2015</td>
<td>Malawi</td>
<td>Outpatient</td>
<td>Quarterly reports generated, also via supervisors face-to-face.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>Title</td>
<td>Authors</td>
<td>Country</td>
<td>Setting</td>
<td>Profession</td>
<td>Group Size</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>------------</td>
<td>------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>L. McLellan, T. Dornan, P. Newton, S.D. Williams, P. Lewis, D. Steinke, et al.</td>
<td>Pharmacist-led feedback workshops increase appropriate prescribing of antimicrobials</td>
<td>England</td>
<td>Hospital</td>
<td>Prescribing (general) - physicians - junior</td>
<td>10</td>
<td>Data collected and analysed by pharmacists from prescription cards. Appropriateness judged by an expert panel. Written feedback and face-to-face given. Workshops involving education and reflection also provided.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>L. Jeffs, D. Doran, L. Hayes, C.</td>
<td></td>
<td>Canada</td>
<td>Hospital</td>
<td>Nursing -</td>
<td>18</td>
<td>Data collected by</td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>Authors</td>
<td>Title</td>
<td>Summary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix 4: Theories, models, and frameworks considered during the synthesis process

<table>
<thead>
<tr>
<th>Theory, model or framework name</th>
<th>Found via</th>
<th>Used?</th>
<th>Reason if not used</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intervention description</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modifiable elements of A&amp;F [1]</td>
<td>Relevant article [1]</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Behaviour change technique taxonomy (BCTT) [2]</td>
<td>Theory-focused literature search</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Template for intervention description and replication (TIDieR) [3]</td>
<td>Relevant article [1]</td>
<td>No</td>
<td>Mismatched to observed findings.</td>
</tr>
<tr>
<td>Standards for Quality Improvement Reporting Excellence (SQUIRE) [4]</td>
<td>Relevant article [1]</td>
<td>No</td>
<td>Mismatched to observed findings.</td>
</tr>
<tr>
<td><strong>Feedback</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feedback intervention theory (FIT) [9–11]</td>
<td>Relevant articles [7,12] Included studies [13–15]</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Control theory (CT) [16]</td>
<td>Relevant article [12] Included study [17]</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Contemporary performance measurement systems [18]</td>
<td>Included study [14]</td>
<td>No</td>
<td>Mismatched to observed findings from health care.</td>
</tr>
<tr>
<td>Ilgen and Davis’ model of negative feedback [19]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Mismatched to observed findings from studies.</td>
</tr>
<tr>
<td>Regulatory focus theory [20,21]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Insufficient explanatory power for observed findings.</td>
</tr>
<tr>
<td>Feedforward theory [22,23]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Mismatched to observed findings.</td>
</tr>
<tr>
<td><strong>Goal setting and action planning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goal setting theory [24]</td>
<td>Relevant article [12] Included study [25]</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Guideline adherence</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cabana Guideline model [31]</td>
<td>Theory-focused literature search</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Guidelines interdependence model [32]</td>
<td>Contact with experts</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Pathman’s model of guideline adherence [33]</td>
<td>Included study [34]</td>
<td>No</td>
<td>Mismatched to observed findings.</td>
</tr>
<tr>
<td>Theory</td>
<td>Literature Search</td>
<td>Matched to Observed Findings</td>
<td>Reason for Mismatch</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-------------------</td>
<td>------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td><strong>Clinical decision support theories [35,36]</strong></td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Mismatched to observed findings.</td>
</tr>
<tr>
<td><strong>Behaviour change</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theoretical Domains Framework (TDF) [37,38]</td>
<td>Relevant article [12]</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Behaviour Change Wheel (BCW) [39]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Insufficient explanatory power for observed data.</td>
</tr>
<tr>
<td><strong>Psychological</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive dissonance [40]</td>
<td>Included study [25]</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Self Affirmation Theory [41]</td>
<td>Theory-focused literature search</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Persuasion theory [42]</td>
<td>Theory-focused literature search</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Pritchard and Ashwood’s Motivation Theory [43]</td>
<td>Included study [44]</td>
<td>No</td>
<td>Insufficient explanatory power for observed data and mismatch to observed data.</td>
</tr>
<tr>
<td>Theory of planned behaviour (TPB) [48]</td>
<td>Included studies [49,50]</td>
<td>No</td>
<td>Insufficient explanatory power for observed data.</td>
</tr>
<tr>
<td><strong>Educational theories</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formative feedback [51]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Insufficient explanatory power for observed data.</td>
</tr>
<tr>
<td>Sargeant’s self assessment models [52,53]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Mismatch to observed data.</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value chain of information [54]</td>
<td>Theory-focused literature search</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Delone and McLean [55]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Mismatch to observed data.</td>
</tr>
<tr>
<td>Fit between Individuals, Task and Technology [56]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Mismatch to observed data.</td>
</tr>
<tr>
<td>Strong Structuation Theory [57]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Mismatch to observed data.</td>
</tr>
<tr>
<td>Technology acceptance model [58]</td>
<td>Included study [8]</td>
<td>No</td>
<td>Mismatch to observed data.</td>
</tr>
<tr>
<td><strong>Context and implementation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diffusion of Innovations [59–61]</td>
<td>Included study [62]</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Consolidated Framework for Implementation Research (CFIR) [63]</td>
<td>Included studies [8,64,65]</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Multilevel quality improvement approach [66]</td>
<td>Included study [67]</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Normalisation Process Theory (NPT) [68]</td>
<td>Theory-focused literature search</td>
<td>No</td>
<td>Mismatch to observed data.</td>
</tr>
<tr>
<td>Knowledge-to-Action [69]</td>
<td>Relevant article [70]</td>
<td>No</td>
<td>Insufficient explanatory power for observed data.</td>
</tr>
<tr>
<td>Promoting Action on Research Implementation in Health Services (PARiHS) [71]</td>
<td>Included study [72,73]</td>
<td>No</td>
<td>Insufficient explanatory power for observed data.</td>
</tr>
<tr>
<td>Organisational Learning [74]</td>
<td>Included study</td>
<td>No</td>
<td>Insufficient explanatory power for observed data.</td>
</tr>
<tr>
<td>Sociology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social comparison [76]</td>
<td>Theory-focused</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference group behaviour [77]</td>
<td>literature search</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

**Models developed by synthesised studies**

| Locus of performance assessment [78]                                    | N/A              |
|                                                                       | No               |
| Ancillary benefits of performance measurement [79]                    | N/A              |
|                                                                       | No               |
| Unintended consequences of performance measurement [80]              | N/A              |
|                                                                       | No               |
| Audit cycle [81]                                                      | N/A              |
|                                                                       | No               |
| PROMs model [82]                                                      | N/A              |
|                                                                       | No               |
| Emotional model of feedback [15]                                      | N/A              |
|                                                                       | No               |

**References**


[17] N. Ivers, J. Barnsley, R. Upshur, K. Tu, B. Shah, J. Grimshaw, et al., My approach to this job is ... one person at a time, Can Fam Physician. 60 (2014) 258–266.


Facilitators and barriers to applying a national quality registry for quality

A.C. Eldh, M. Fredriksson, C. Halford, L. Wallin, T. Dahlström, S. Vengberg, et al.,

intervention to improve the management of patients with heart failure: the dynamic


to Spread Good Ideas: A systematic review of the literature on diffusion,

T. Greenhalgh, G. Robert, F. Macfarlane, P. Bate, O. Kyriakidou, Diffusion of


K.M. Cresswell, S. Sadler, S. Rodgers, A. Avery, J. Cantrill, S. a Murray, et al., An

L.J. Damschroder, D.C. Aron, R.E. Keith, S.R. Kirsh, J. a Alexander, J.C. Lowery,

A.C. Eldh, M. Fredriksson, S. Vengberg, C. Halford, L. Wallin, T. Dahlström, et al.,


M.J. Johnson, C.R. May, Promoting professional behaviour change in healthcare:


A.C. Eldh, M. Fredriksson, C. Halford, L. Wallin, T. Dahlström, S. Vengberg, et al., Facilitators and barriers to applying a national quality registry for quality


## Appendix 5: CP-FIT Constructs

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal setting</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a)</td>
<td>The clinical topic and its associated clinical behaviours or patient outcomes against which clinical performance will be measured are decided. This usually occurs as a one-off step at the beginning of the process.</td>
<td>Confidence: High&lt;br&gt;Papers (n=8): 7049, 7050, S154 (86), S175 (85), 2841, 4222, 8249, S109&lt;br&gt;Models: BCT (1.1 and 1.3), TDF, Locke and Latham (Goal)</td>
</tr>
<tr>
<td>(b)</td>
<td>Goal setting may be re-visited in the A&amp;F process cycle after Audit has taken place if, for example: if there are problems with collecting the correct clinical performance data (e.g. 2841, 7049, 8249), the goals are set too high to be achievable (e.g. 4222), or current achievement is already so high that further improvement is not possible (e.g. 7049).</td>
<td>Papers (n=5): 2841, 7049, 8249, 4222, 7049&lt;br&gt;Confidence: Moderate&lt;br&gt;Reason: Concerns regarding coherence, relevance, and methodological limitations of findings.&lt;br&gt;Models: BCT (1.1 and 1.3), TDF, Locke and Latham (Goal), FIT (Eliminating Feedback-Standard Gap Strategy – 2. Abandon / 3. Change standard)</td>
</tr>
<tr>
<td><strong>Audit</strong></td>
<td>Clinical performance data is collected on a defined population of patients and analysed in accordance with Goal setting.</td>
<td>Confidence: High&lt;br&gt;Papers (n=29): 1948, 2841, 5357, 627, 7025, 7301, 8249, S118, S132, S136, S175 (85), S154 (86), S62, S67, 1271, 5857, 627, 7049, 7050, 7816, 8167, 8249, S105, S109, S120, S127, S14, S15, S16, S175, S19, S25&lt;br&gt;Models: Nil</td>
</tr>
<tr>
<td><strong>Feedback</strong></td>
<td>A message regarding the measured clinical performance is generated and communicated to health professionals in whom improvement is desired.</td>
<td>Confidence: High&lt;br&gt;Papers (n=11): S154, 7049, 1948, 2023, 7025, S105, S17, S19, 2794, 5033, S52&lt;br&gt;Models: CT (Input), Ilgen (Complex feedback stimulus), FIT (Feedback intervention), BCTT (2.2. Feedback on behaviour; 2.7. Feedback on outcome(s) of behavior), Feedback (Locke and Latham)</td>
</tr>
<tr>
<td><strong>Interaction</strong></td>
<td>The health professionals receive the feedback message.</td>
<td>Confidence: High&lt;br&gt;Papers (n=45): (87), S175 (85), S154 (86), 1271, 1591, 1948, 2794, 2841, 3351, 512, 5357, 5857, 6087, 7025, 7049, 7816, S105, S109, S118, S127, S132, S136, S14, S15,</td>
</tr>
<tr>
<td>Category</td>
<td>Description</td>
<td>Confidence</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>Perception</strong></td>
<td>The health professionals interpret the feedback message. This does not have to be an accurate interpretation.</td>
<td>High</td>
</tr>
<tr>
<td><strong>Acceptance</strong></td>
<td>The health professionals believe the feedback message is an accurate portrayal of their performance. This belief does not necessarily have to be correct. They become aware of their measured level of performance.</td>
<td>High</td>
</tr>
<tr>
<td><strong>Verification</strong></td>
<td>Health professionals interrogate the feedback message or the underlying clinical performance data in an attempt to improve their Perception of the feedback message prior to its Acceptance.</td>
<td>High</td>
</tr>
<tr>
<td><strong>Intention</strong></td>
<td>The health professionals make a conscious decision to behave in a certain way in response to the feedback message.</td>
<td>High</td>
</tr>
<tr>
<td><strong>Behaviour</strong></td>
<td>The health professionals behave in a way consistent with their intentions. This may be an increase, decrease, change, or maintenance of behaviour.</td>
<td>High</td>
</tr>
</tbody>
</table>
### Potential unintended consequences

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Relevant evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gaming</strong></td>
<td>In response to A&amp;F, health professionals may unethically manipulate clinical data (e.g. record a care process has happened when it has not 1948) or change their patient population (e.g. a surgeon may choose not to operate on high-risk patients S127) in order to artificially improve their measured clinical performance.</td>
<td>Confidence: Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Models: Cognitive dissonance, self-affirmation</td>
</tr>
<tr>
<td><strong>Tunnel vision</strong></td>
<td>In response to A&amp;F, health professionals may become overly focused on the topic against which clinical performance is measured, to the detriment of other clinical areas. This may manifest during the care of individual patients (e.g. 6087) or during quality improvement activities (e.g. 2857).</td>
<td>Confidence: Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Papers (n=6): 1948, 2857, 512, 6087, S109, S117, S67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Models: Cognitive dissonance, self-affirmation</td>
</tr>
<tr>
<td><strong>Inappropriate care</strong></td>
<td>Patients may receive care that is either unnecessary or for which they have not provided consent in order to improve measured clinical performance.</td>
<td>Confidence: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Papers (n=2): 6087, S112</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Models: Cognitive dissonance, self affirmation</td>
</tr>
<tr>
<td><strong>Patient engagement</strong></td>
<td>Patients may become more engaged in their care.</td>
<td>Confidence: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Papers (n=2): S4, S62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Models: Nil</td>
</tr>
</tbody>
</table>

### Propositions

<table>
<thead>
<tr>
<th>Name</th>
<th>Description and effect</th>
<th>Relevant evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proposition 1</strong></td>
<td>A&amp;F interventions exert their effects by inducing patient-level behaviours</td>
<td>Confidence: High</td>
</tr>
<tr>
<td>Behavioural induction</td>
<td>in health professionals.</td>
<td></td>
</tr>
</tbody>
</table>

| **Proposition 2 – Capacity limitations** | Health care organisations have limited capacity to engage with and respond to the demands of A&F interventions. | Confidence: High |

| **Proposition 3 – Identity and culture** | Health care professionals and organisations have a strong set of beliefs and behaviours regarding how they provide patient care that influence how they interact with A&F interventions. | Confidence: High |

| **Mediating variables** | **Description and effect** | **Relevant evidence** |

| **Actionability** | Description: The feedback message’s ability to directly facilitate behaviours in A&F recipients. Effect: Greater Actionability tends to facilitate A&F processes. Supports proposition 1 – Behavioural induction: the more a feedback message can directly facilitate behaviours in A&F recipients the more effective it is. | Confidence: High Papers (n=25) Models: Locke and Latham (Self-efficacy); Ilgen (Locus of control); TDF (Beliefs about capabilities) |

| **Resource match** | Description: Whether the associated costs of the A&F intervention are matched by the available resource. Effect: Greater facilitates A&F processes. Supports propositions 2 – Capacity limitations: if a health care organisation’s resources match the costs of implementing an A&F intervention they are more likely to engage and respond to it. | Confidence: High Papers (n=31) Models: Diffusion of innovations (Dedicated time and resources; Cost); CFIR (Resources; Cost) |

| **Complexity** | Description: The difficulty of performing the A&F processes. Effect: Greater Complexity inhibits A&F processes. Supports proposition 2 – Capacity limitations: the simpler an A&F intervention is to engage with, the less resource it requires. | Confidence: High Papers (n=27): Models: Diffusion of innovations (Complexity); CFIR (Complexity), Locke and Latham (Task complexity) |

| **Relative advantage** | Description: Recipients' perceived benefits of the A&F intervention, often when compared to alternative existing or proposed ways of working. Aspects of an A&F intervention considered to have a relative advantage are (understandably) situation-specific, so its appearance as a mediating variable is inconsistent. Effect: Greater Relative advantage facilitates A&F processes. Supports proposition 2 – Capacity limitations: the more an A&F intervention has a perceived advantage over current ways of working, the | Confidence: High Papers (n=40) Models: Diffusion of innovations (Relative advantage); CFIR (Relative advantage) |
more likely it is to be adopted within the available resources.

| Compatibility | Description: The degree of 'fit' between the A&F intervention and characteristics of the recipient and their organisation e.g. beliefs, norms, values, culture, structures, processes, technical systems. Effect: Greater Compatibility facilitates A&F processes. Supports proposition 3 – Identity and culture: the more an A&F intervention can align with the beliefs, systems, and processes of an organisation and its staff, the greater its relevance, and the less disruption required for its implementation. |
| --- | |
| Credibility | Description: The perceived trustworthiness and reliability of the A&F intervention. Effect: Greater Credibility facilitates A&F processes. Supports proposition 3 – Identity and culture: the more trustworthy and reliable an A&F intervention, the more likely they are to believe it will help them improve patient care. |
| Social influence | Description: Interpersonal processes that cause A&F recipients to change their thoughts, feelings, or behaviours. Key aspects include: 1) Competition (between health professionals), 2) Social proof (the desire to behave in the same way as other health professionals), 3) Authority (to obey credible authority figures), 4) Liking (persuaded by people they like), and 6) Reference group behaviour (where health professionals feel part of a group and will change their behaviour if they believe their membership of that group is threatened). Effect: Greater Social influence facilitates A&F processes. Supports propositions 3 – Identity and culture: the more an A&F intervention can harness the social dynamics between health professionals, the more likely they are to be implemented. |

<table>
<thead>
<tr>
<th>Moderating variables</th>
<th>Relevant evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audit methods</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Description: The perceived accuracy of measured clinical performance. This may relate to: 1) the nature of source data, 2) the method of analysis, and/or 3) sample size.</td>
</tr>
</tbody>
</table>

Confidence: High
Papers (n=40)
Models: Nil
| Effect and mechanism (1): Greater Accuracy facilitates Acceptance, Intention, and Behaviour, by increasing **Credibility** (by more accurately reflecting recipients’ clinical performance), **Actionability** (by enabling recipients to focus on aspects of their clinical performance they believe requires improvement), Relative advantage (particularly if there is no existing A&F intervention against which to compare, or if an existing A&F intervention is less accurate), and **Resource match** (by only highlighting areas of clinical performance that they believe truly require improvement). | Confidence: High  
Papers (n=40)  
Models: Nil |
|---|---|
| Effect and mechanism (2): Lower **Accuracy** facilitates **Verification** because there is reduced **Credibility** of the feedback message and recipients are motivated to check it. | Confidence: Moderate  
Reasons: Concerns regarding the adequacy and coherence of findings.  
Papers (n=5): 1591, 5033, 7694, S38, S6 |
| **Exclusions** | **Conducted by recipients** |
| Description: Recipients can exclude patients they deem unsuitable to be included in the measurement of their clinical performance. For example if there is **Choice misalignment** (e.g. if the patient refuses care S112) or if it is **Clinically inappropriate** (e.g. due to existing comorbidities e.g. S6). | Description: Clinical performance data are collected and/or analysed by the recipients of the A&F intervention. |
| Confidence: High  
Papers (n=):  
Models: Nil | Papers (n=):  
Models: Nil |
| Effect and mechanism (1): Its absence inhibits **Acceptance** by decreasing **Credibility** (by less accurately reflecting recipients’ clinical performance), **Actionability** (by preventing recipients focusing on patients over which they can improve upon), **Relative advantage** (as most A&F interventions do not provide this feature), and **Compatibility** (by preventing health professionals from judiciously/autonomously applying scientific evidence, and providing patient-centred care, whilst still achieving high levels of clinical performance). | Confidence: Low  
Papers (n=3): S112, S117, S38 |
| Effect and mechanism (2): Its absence increases negative **Emotions** (frustration). | |
| Effect and mechanism (2): Improves (ripple effect) by engaging recipients in a common goal. | Confidence: Low.  
Reason: Substantial concerns regarding coherence, adequacy, and methodological limitations of findings.  
Papers (n=3): 5357, 7025, S118 |
|---|---|
| Effect and mechanism (3a): Increases audit Accuracy (ripple effect). | Confidence: Low.  
Reason: Substantial concerns regarding coherence, adequacy, and methodological limitations of findings.  
Papers (n=4): 7816, S132, S19, S67 |
| Effect and mechanism (3b): Decreases audit Accuracy (ripple effect). | Confidence: Low.  
Reason: Substantial concerns regarding coherence, adequacy, and methodological limitations of findings.  
Papers (n=4): 7816, S132, S19, S67 |
Reason: Substantial concerns regarding coherence, adequacy, and methodological limitations of findings.  
Papers (n=2): 627, S118 |

**Manual vs Automatic**  
**Description:** Whether clinical performance data are collected regarding individual patients by hand (either using paper e.g. S154 or electronic health records e.g. S109; Manual) rather than automatically (e.g. 1948; Automatic ).  
Papers (n=16): 1591, 1948, 5857, 7049, 7050, 7816, S105, S109, S132, S136, S14, S154, S16, S175, S6, S62  
Models: Nil  

**Effect and mechanism (1): Manual inhibits Audit by decreasing Resource match and increasing Complexity, whereas Automatic facilitates Audit by increasing Resource match and Relative advantage, and increasing Complexity.**  
Confidence: High  
Papers (n=11): 1948, 5857, 7049, 7050, 7816, S109, S136, S14, S16, S175, S62  
Models: Nil  

**Electronic health record**  
**Description:** Clinical performance data are collected from electronic health records.  
Papers (n=9): 2249, S1, S105, S112, S132, S136, S14, S38, S6  
Models: Nil  

**Primary effect and mechanism:** No clear effect. Likely confounder factor for Accuracy, and Manual vs Automatic.  
N/A  

**Number of data sources**  
**Description:** Number of sources of clinical data used to calculate clinical performance.  
Papers (n=4): 2841, 8249, S154, S175  
Models: Nil  

**Primary effect and mechanism:** Inhibits Audit by increasing Complexity (more data sources is more complex), and decreasing Compatibility (more data sources are less likely to ‘fit’ with existing data systems) and Resource match (more data sources require more resource).  
Confidence: Low.  
Papers (n=3): 8249, S154, S175  
Models: Nil  

**Patient-**  
**Description:** The data source is patient-reported e.g. patient-reported  
Papers (n=6): S136, S15, S16, S28, S30, S7
<table>
<thead>
<tr>
<th><strong>reported</strong></th>
<th>experience measures (PREM) or outcome measures (PROM). May correlate with Accuracy as some recipients may view patient-reported data as unreliable.</th>
<th>Models: Nil</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effect and mechanism (1a – model tension):</strong> Facilitates Acceptance and Intention by increasing Compatibility (with recipients’ motivations to provide patient-centred care).</td>
<td>Confidence: Moderate&lt;br&gt;Reason: Concerns regarding the coherence of findings&lt;br&gt;Papers (n=6): S136, S15, S16, S28, S30, S7</td>
<td></td>
</tr>
</tbody>
</table>

### Behaviour

| **Patient-level vs organisational-level** | Description: Behaviour may relate to individual clinicians caring for individual patients, or the organisations in which this care takes place. Patient-level behaviours can occur either during (e.g. 3351) or outside (e.g. 187) the point-of-care; and may be either retrospective/remedial (e.g. S104) or prospective/preventive (e.g. S1). Organisational-level behaviours aim to systematically changing the way care is delivered by a health care organisation, and may relate to starting, stopping, or modifying: services and protocols (e.g. 1948), or staff roles and training (e.g. 2857). Patient-level behaviours result from feedback acting as a reminder or new knowledge regarding how to behave in future, or specific instructions where Patient lists are provided. Organisational-level behaviours require more interpretation and resource to enact (time, skill, knowledge etc). | Papers (n=58): 1271, 1591, 187, 1948, 2023, 2794, 2841, 2857, 3351, 4222, 5033, 512, 5235, 5357, 5532, 5857, 627, 7025, 7049, 7050, 7194, 7301, 7694, 7783, 7816, 8167, 8249, S1, S104, S105, S112, S117, S118, S120, S127, S132, S136, S14, S15, S158, S159, S16, S17, S175, S19, S25, S28, S30, S32, S38, S4, S52, S6, S62, S67, S7, S71, S81. Models: Ilgen, CT, Multilevel quality improvement approach |
| **Effect and mechanism (1):** Organisational-level behaviour facilitates Performance improvement because it can facilitate multiple patient-level behaviours by augmenting the clinical environment in which they take place, whereas their absence can lead to limited effects only (Actionability). | Confidence: High<br>Papers (n=10): 1948, 2023, 2857, 512, 7816, S19, S4, S1, S32, S118 |
| **Type** | Description: The type of behaviour required to improve on suboptimal performance. This may be an increase, decrease, change or maintenance of behaviour. | Papers (n=3): 187, 8167, S30<br>Models: Nil |
| **Co-interventions** | **Problem solving**<br>Description: Analysis of reasons for sub-optimal clinical performance and formulation of solution(s) to address them. This may be performed as part | Papers (n=22)<br>Models: BCTT (1.2 Problem solving) |
of the A&F intervention, or the recipients may be required to do it themselves – either on their own (e.g. S1); or with support from an External change agent (e.g. S158), the feedback provider (e.g. S15), or peers (e.g. S109). May correlate with Action planning, Social support, Active delivery (when delivered face-to-face) and Delivery to a group (when delivered face-to-face) as they are often delivered concurrently.

| Effect and mechanism (1a): Facilitates Perception, Intention and Behaviour by increasing Actionability (by providing practical support on how to act effectively to the feedback message) and Resource match (by addressing health professionals’ general lack of knowledge and skills to perform these behaviours). | Confidence: High  
Papers (n=22): 1271, 5857, 7025, 7301, S1, S109, S118, S127, S132, S15, S150, S25, 1948, 4222, 7783, S117, S118, S158, S175, S62, S71, S81  
Models: Nil |
| Effect and mechanism (1b): Facilitates Organisational-level behaviour by increasing Actionability (by providing practical support on how to act effectively to the feedback message) and Resource match (by addressing health professionals’ general lack of knowledge and skills to perform these behaviours). | Confidence: Moderate  
Reasons: Concerns regarding adequacy of data.  
Papers: 1948, 7783, S175 |
| Effect and mechanism (2): Increases Teamwork (ripple effect) if Problem solving occurs through discussions between staff in the same organisation. | Confidence: Low  
Papers (n=4): 5857, 7025, S109, S71 |
| Effect and mechanism (3): Increases positive Emotions (ripple effect) if done in a supportive environment. | Confidence: Low  
Papers (n=3): S127, S150, S158 |
| Effect and mechanism (4): Increases Knowledge and skills – clinical (ripple effect). | Confidence: Low  
Papers (n=3): 4222, S118, S158. |

Action planning  
Description: Instructions for specific behaviours to improve clinical performance. They may be provided as part of the A&F intervention (e.g. 1271, S112) or generated by the recipients themselves (e.g. 7783, S1). Differs from Problem solving because suggested actions are not necessarily based on analysis of reasons for sub-optimal clinical performance; though may be correlated as they can co-exist. May also correlate with Active delivery (when delivered face-to-face), Delivery to a group (when delivered face-to-face) and Social support as they are often delivered concurrently.

| Effect and mechanism (1a): Facilitates Intention and Behaviour by increasing Actionability (providing practical support on how to respond | Confidence: High  
Papers (n= 21): 1271, 187, 1948, 5033, 5235, 5857, 7025, 5033, 5235, 5857, 7025, 1271, 187, 1948, 5033, 5235, 5857, 7025, 1271, 187, 1948, 5033, 5235, 5857, 7025, |
| Effect and mechanism (1b): Facilitates Organisational-level behaviour by increasing Actionability (by providing practical support on how to act effectively to the feedback message) and Resource match (by addressing health professionals' general lack of knowledge and skills to perform these behaviours). | Confidence: Moderate
Reasons: Concerns regarding adequacy of data.
Papers: 1948, 7783, S175 |
| Reminders Description: Point-of-care alerts (usually) integrated into a patient’s electronic health record (EHR) to prompt clinicians to recognise which clinical performance measures the patient has not achieved. These are generally only encountered when the EHR is opened by a clinician during a consultation with the patient. | Papers (n=4): S112, S117, S4, S62 Models: BCTT (1.4 Prompts/cues) |
| External change agent Description: The presence of individuals not affiliated with the recipients' own organisation whose intention is to facilitate A&F processes. This can be direct, for example by performing improvement behaviours themselves (e.g. 1591); or indirect by helping with action planning or problem solving (e.g. S62). | Papers (n=17):
Models: Diffusion of innovations (Role of change agency); CFIR (External Change Agents) |
| Effect and mechanism (1): Dependent on their specific role, External change agents can positively influence all A&F processes by providing an additional resource to increase Resource match. | Papers (n=17): |
| Effect and mechanism (2): Increases Knowledge and skills – Quality improvement (ripple effect) via Social influence (impacting knowledge to recipients) and Credibility. | Confidence: Moderate
Reason: Concerns regarding the adequacy of findings.
Papers (n=3): S81, S62, S120 |
| Financial reward Description: Payment received for either participating in, or improving clinical performance during an A&F intervention. | Papers (n=17):
Models: BCTT (10.2 Material reward [behaviour]; 10.10 Reward [outcome]; 14.2 Punishment; 14.3 and Remove reward) |
| Effect and mechanism (1a – model tension): Inhibits Acceptance by reducing Compatibility with recipients' professional role. | Confidence: Moderate
Reason: Concerns regarding adequacy and coherence of findings.
| Effect and mechanism (1a – model tension): Facilitates Interaction, Intention and Behaviour, possibly by increasing Resource match and Relative advantage. | 7783, 8167, S1, S112, S117, S120, S127, S15, S158, S175, S6, S62, S71, S81 |
| Effect and mechanism (2): Leads to **Gaming** (unintended consequence) by increasing **Resource match** (in an attempt to maintain resource, attempts to preserve financial income are preserved at any cost). | Confidence: Low  
Reason: Substantial concerns regarding adequacy of findings.  
Papers (n=2): S109, S112 |
|---|---|
| Effect and mechanism (4): Increases **Resource** (ripple effect). | Confidence: Low  
Reason: Substantial concerns regarding adequacy of findings.  
Papers (n=2): 627, S109 |
| **Non-financial reward**  
Description: A reward other than payment for either participating in, or improving clinical performance during an A&F intervention. For example, receiving a certificate or praise from a superior. | Papers (n=5): 1271, S104, S15, S30, S4  
Models: BCTT (10.3 Non-specific reward; 10.4 Social reward; 10.10 Reward [outcome]) |
| **Social support**  
Description: A&F recipients receive help from peers in discussing and reflecting on their clinical performance. This may be informal (e.g. 4222) or formal discussions (e.g. 4222) or different (e.g. 7025) organisations. May be correlated with Action planning, Problem solving, Active delivery (when delivered face-to-face) and Delivery to a group (when delivered face-to-face) as they are often delivered concurrently. | Papers (n=28):  
Models: BCTT (3.1 Social support [unspecified]); Diffusion of Innovations (Intentional spread strategies) |
| Effect and mechanism (1a): Facilitates **Intention**, **Behaviour**, and **Clinical performance** by increasing **Actionability** (by providing practical support on how to act effectively to the feedback message) and **Resource match** (by addressing health professionals’ general lack of knowledge and skills to perform these behaviours). | Confidence: High  
Reason: N/A.  
| Effect and mechanism (1b): Facilitates **Organisational-level** behaviour by increasing **Actionability** (by providing practical support on how to act effectively to the feedback message) and **Resource match** (by addressing health professionals’ general lack of knowledge and skills to perform these behaviours). | Confidence: Moderate  
Reasons: Concerns regarding adequacy of data.  
Papers: 1948, 2023, 7783, S175 |
| Effect and mechanism (2): Improves the **Teamwork** (ripple effect) by increasing **Social influence** (engaging the wider team towards the A&F intervention’s goal [Social proof, Liking, Reference group]). | Confidence: High  
Paper (n=7): 5357, 5857, 7025, 7783, S105, S15, S71 |
| Effect and mechanism (3): Facilitates Acceptance and Perception by reducing **Complexity** (making the feedback easier to understand). | Confidence: Moderate  
Reasons: Minor concerns regarding methodological |
| Effect and mechanism (4): Increases *Delivery – Active* (ripple effect) by ensuring that recipients receive the feedback message face-to-face. | Confidence: Moderate  
Reason: Minor concerns regarding methodological limitations and adequacy of findings.  
Papers (n=12): 1271, 2023, 4222, 5857, 7694, 7783, S127, S15, S158, S19, S33, S71 |
|---|---|
| Effect and mechanism (5): Improves the *A&F attitude* (ripple effect) by increasing *Social influence* (engaging the wider team in the A&F process [Social proof, Liking]). | Confidence: Moderate  
Reason: Minor concerns regarding methodological limitations and adequacy of findings.  
Papers (n=9): 1948, 2023, 7025, 7049, 7783, S120, S19, S30, S71 |
| Effect and mechanism (6): Improves the *Extra-organisational networks* (ripple effect) if social support provided by recipients from different organisations. | Confidence: Moderate  
Reason: Minor concerns regarding adequacy of findings.  
Paper (n=5): 4222, 7025, S15, S33 |
| Effect and mechanism (7): Improves the *Intra-organisational networks* (ripple effect) by engaging staff members within an organisation. | Confidence: Moderate  
Reason: Minor concerns regarding adequacy of findings.  
Paper (n=5): 5357, 5857, 7025, 7783, S105 |
| Effect and mechanism (8): Increases *Knowledge and skills – clinical* (ripple effect). | Confidence: Low  
Reason: Substantial concerns regarding adequacy of findings.  
Papers (n=3): 4222, S105, S158. |

**Clinical education**

Description: A&F recipients receive advice or demonstrations on how to perform clinical behaviours, or regarding the clinical outcomes, relevant to the clinical performance under measurement.

| Effect and mechanism (1): Increases *Knowledge and skills – clinical* (ripple effect). | Confidence: Low  
Reason: Substantial concerns regarding adequacy and methodological limitations of findings.  
Papers (n=4): 2023, 2794, 2857, S62 |
|---|---|
| Effect and mechanism (2): Increases Teamwork, especially if educational sessions delivered to a multidisciplinary group. | Confidence: Low  
Reason: Substantial concerns regarding methodological limitations, and coherence and adequacy of findings. |
### Feedback message

**Benchmarking**

<table>
<thead>
<tr>
<th>Description: Compares recipients' clinical performance with other health professionals. Others may include those internal or external to their organisation, and the comparison may be anonymous or identifiable. May correlate with Accuracy when comparisons between health professionals may be perceived as invalid e.g. without statistical adjustment in outcome measures.</th>
<th>Papers (n=34): Models: BCTT (6.2 Social comparison); FIT (Normative information)</th>
</tr>
</thead>
</table>
| **Effect and mechanism (1a):** Facilitates Intention and Behaviour via Social influence (by harnessing competition between recipients [social comparison theory], changing recipients' behaviour if they see others behaving differently [Social proof], and trying to maintain their status in a group of high performing clinicians – particularly if the benchmarking is identifiable [Reference group behaviour theory]). | Confidence: High  
Papers (n=12): 2857, 5033, 5857, 8249, S104, S118, S127, S132, S17, S19, S32, S7 |
| **Effect and mechanism (1b):** Facilitates Organisational-level behaviour by increasing Actionability (by providing insight into how other organisations' performance compares). | Confidence: Low  
Reasons: Substantial concerns regarding adequacy and relevance of data.  
Papers: 5033, S19 |
| **Effect and mechanism (2):** Facilitates Perception and Intention by reducing Complexity (enabling comparison of with others enables recipients to better understand how well they are performing, and which areas require improvement). | Confidence: Moderate  
Reasons: Minor concerns regarding coherence and adequacy of findings.  
Papers (n=11): 1948, 5857, 7783, S105, S17, S118, S120, S132, S175, S19, S25, S32, S6  
Models: FIT (Normative information), Social Comparison Theory |
| **Effect and mechanism (3):** Induces both positive and negative Emotions responses dependent on whether relative Performance level is high or low respectively (ripple effect) via Social influence (by increasing competition [social comparison theory]) and Compatibility with recipients' expectations. | Confidence: Moderate  
Reasons: Minor concerns regarding adequacy of findings.  
Papers (n=11): 512, 7025, S104, S158, S17, S4 |
| **Effect and mechanism (4):** Inhibits Acceptance (model tension) by decreasing Credibility (when comparisons between health professionals may be perceived as invalid e.g. without statistical adjustment in outcome measures). | Confidence: Moderate  
Reasons: Minor concerns regarding adequacy of findings.  
Papers (n=4): 7194, S127, S16, S17 |
| Effect and mechanism (5): Inhibits Acceptance (model tension) by decreasing Compatibility and Credibility (some A&F recipients believe comparing health professional’s performance is inappropriate and is a threat to their autonomy). | Confidence: Low  
Reasons: Substantial concerns regarding coherence and adequacy of findings.  
Papers (n=3): 1271, 2857, 7194. |
|---|---|
| Effect and mechanism (6): Facilitates Interaction by increasing Relative advantage (the ability to view other health professionals’ clinical performance information is not generally available) and Social influence (by increasing competition [social comparison theory]). | Confidence: Low  
Reasons: Substantial concerns regarding coherence and adequacy of findings.  
| Effect and mechanism (7): Increases Resource by enabling comparisons to other health organisations to persuade management to fund organisational change (Social influence). | Confidence: Low  
Reasons: Substantial concerns regarding adequacy and relevance of findings.  
Papers (n=3): 5033, S19, S25 |
| **Active delivery** | Description: How much the feedback message is actively 'pushed' to recipients. In general, feedback messages sent where users have to obtain the feedback themselves (e.g. web-based or computer application) are less active than those sent to them (e.g. via mail or e-mail), which are in turn less active than those delivered face-to-face. May correlate with Social support, Action planning and Problem solving when delivered face-to-face, as they are often delivered concurrently. | Papers (n=31):  
Models: A&F modifiable elements (Face to face). |
| Effect and mechanism (1a – model tension): Facilitates Interaction by reducing Complexity (making the feedback message simpler to receive). | Confidence: High  
Papers (n=22): 1271, 1591, 1948, 3351, 5033, 5857, 6087, 7049, 7194, 7694, 7783, S104, S118, S120, S132, S150, S17, S19, S4, S52, S6, S81 |
| Effect and mechanism (1b – model tension): Inhibits Interaction if solely requires formal face-to-face feedback sessions by decreasing Resource match (as they require significant time commitment from recipients). | Confidence: High  
Papers (n=7): 1271, 5857, 7694, 7783, S120, S52, S6 |
| **Delivery to a group** | Description: The feedback message is delivered to groups of health professionals rather than just individual health professionals. They will usually work together in the same organisation or team. May correlate with Social support, Action planning, Active delivery and Problem solving when delivered face-to-face, as they are often delivered concurrently. | Papers (n=31):  
Models: A&F modifiable elements (Groups of providers). |
<table>
<thead>
<tr>
<th>Effect and mechanism (4): Improves Teamwork (ripple effect) by increasing Social influence (engaging the wider team towards the A&amp;F intervention’s goal [Social proof, Liking, Reference group]).</th>
<th>Confidence: High Paper (n=10): 3351, 5357, 5532, 5857, 7025, 7783, S15, S52, S7, S71</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect and mechanism (2): Improves Extra-organisational networks (ripple effect) if feedback message is delivered to those recipients from different organisations.</td>
<td>Confidence: Moderate Reason: Minor concerns regarding adequacy of findings. Paper (n=5): 5857, 7025, 7783, 7816, 8249</td>
</tr>
<tr>
<td>Effect and mechanism (3): Improves Intra-organisational networks (ripple effect) by engaging staff members within an organisation.</td>
<td>Confidence: Moderate Reason: Minor concerns regarding adequacy of findings. Paper (n=6): 5357, 5532, 7025, 7783, S52, S7</td>
</tr>
<tr>
<td>Effect and mechanism (1): Improves the A&amp;F attitude (ripple effect) by engaging the wider team in the A&amp;F process.</td>
<td>Confidence: Low Reason: Substantial concerns regarding coherence and adequacy of findings. Paper (n=4): 7025, S15, S33, S7</td>
</tr>
<tr>
<td>Effect and mechanism (5): Causes negative Emotions (e.g. embarrassment) (ripple effect) if clinical performance is poor.</td>
<td>Confidence: Low Reason: Substantial concerns regarding coherence and adequacy of findings. Paper (n=4): 7025, S15, S33, S7</td>
</tr>
<tr>
<td>Effect and mechanism (6): Increases goal Controllability (ripple effect) by making recipients realise their role in the wider group.</td>
<td>Confidence: Low Reason: Substantial concerns regarding coherence and adequacy of findings. Paper (n=2): S7, 1271</td>
</tr>
</tbody>
</table>

**Framing**

<table>
<thead>
<tr>
<th>Framing Description: Whether recipients’ clinical performance is presented to emphasise their achievements (positive framing) or shortfalls (negative framing), independent of their actual level of performance (see Performance level). May correlate with Function and Non-financial reward, as they often co-exist.</th>
<th>Papers (n=6): 1271, S104, S117, S62, S67, S81 Models: Ilgen (Sign), FIT (Sign)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect and mechanism (1): Positive framing facilitates Acceptance by increasing Compatibility with recipients’ professional role, motivation, and expectations (as health professionals generally intend to and believe they provide a good standard of care).</td>
<td>Confidence: Low Reason: Substantial concerns regarding coherence and adequacy of findings. Paper (n=3): S104, S62, S81</td>
</tr>
<tr>
<td>Effect and mechanism (2): Positive framing causes positive Emotions by increasing Compatibility with recipients’ professional role, motivation, and expectations (as health professionals generally intend to and believe they provide a good standard of care).</td>
<td>Confidence: Low Reason: Substantial concerns regarding coherence and adequacy of findings. Paper (n=3): 1271, S104, S117</td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td><strong>Description:</strong> The frequency of feedback messages produced and delivered to recipients.</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Effect and mechanism (1a – model tension):</strong> Increased frequency inhibits <em>Perception</em> by increasing <em>Complexity</em> (too much feedback makes it more difficult to understand) and decreasing <em>Resource match</em> (feedback provided too often gives less time to act on it).</td>
<td><strong>Confidence:</strong> Moderate <strong>Reason:</strong> Concerns regarding adequacy of findings. <strong>Papers (n=4):</strong> 5532, 7194, S1, S120.</td>
</tr>
<tr>
<td><strong>Effect and mechanism (1b – model tension):</strong> Decreased frequency inhibits <em>Interaction</em> and <em>Intention</em> by increasing <em>Complexity</em> (making the feedback message more difficult to receive).</td>
<td><strong>Confidence:</strong> Low <strong>Reason:</strong> Substantial concerns regarding adequacy of findings. <strong>Papers (n=3):</strong> S17, S16, S14.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Function</strong></th>
<th><strong>Description:</strong> The recipients’ perception of whether the function of the A&amp;F intervention is to punish them for providing suboptimal care (punitive), or to support them improve care quality (supportive). Correlates with <em>Reporting, Ownership</em> and <em>Source – Location</em>, as they often co-exist.</th>
<th><strong>Papers (n=34):</strong> <strong>Models:</strong> Ilgen (Feedback function), Diffusion of Innovations (Meaning)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effect and mechanism (1):</strong> Supportive functionality facilitates <em>Acceptance</em> by increasing <em>Compatibility</em> with recipients’ professional role, and motivation (as health professionals generally want to improve care quality), whereas a punitive functionality decreases <em>Compatibility</em> with their sense of autonomy.</td>
<td><strong>Confidence:</strong> High <strong>Paper (n=18):</strong> 2249, 2857, 5357, 5857, 6087, 7194, S1, S104, S112, S118, S136, S19, S25, S28, S30, S6, S62, S7, S81</td>
<td></td>
</tr>
<tr>
<td><strong>Effect and mechanism (2):</strong> Punitive functionality increases negative <em>Emotions</em> (ripple effect) (e.g. fear and anxiety) by decreasing Compatibility with recipients’ professional role, autonomy, and motivation (as health professionals generally want to improve care quality).</td>
<td><strong>Confidence:</strong> High <strong>Paper (n=17):</strong> 1271, 1948, S104, S117, S127, S30, S38, S6, S67, S7</td>
<td></td>
</tr>
<tr>
<td><strong>Effect and mechanism (3a – model tension):</strong> Punitive functionality may facilitate <em>Intention, Behaviour,</em> and <em>Performance improvement</em> through <em>Social influence</em> (as health professionals want to maintain their status as a high performing clinician [Reference group behaviour]).</td>
<td><strong>Confidence:</strong> Moderate <strong>Reason:</strong> Concerns regarding the coherence of findings. <strong>Papers (n=10):</strong> 1948, 2249, 5033, S1, S104, S118, S19, S30, S33, S81</td>
<td></td>
</tr>
<tr>
<td><strong>Effect and mechanism (3b – model tension):</strong> Punitive functionality may inhibit <em>Intention, Behaviour,</em> and <em>Performance improvement</em> by decreasing <em>Compatibility</em> with recipients’ motivation (as they already want to improve care quality) and autonomy.</td>
<td><strong>Effect and mechanism (4):</strong> Punitive functionality facilitates <em>Tunnel vision</em> (unintended consequence) through <em>Social influence</em> (as health professionals generally have a sense of self-preservation, if not defined).</td>
<td><strong>Confidence:</strong> Low <strong>Reason:</strong> Substantial concerns regarding adequacy of findings. <strong>Papers (n=5):</strong> 1948, 2249, 5033, S104, S118, S19, S30, S33, S81</td>
</tr>
</tbody>
</table>
professionals tend to conform to authority (Authority) and via Compatibility (to correct their own expectations of themselves [self-affirmation, Cognitive dissonance]).

### Graphical elements

**Description:** Clinical performance is presented using graphical elements such as icons (e.g. 3351), use of colours (e.g. S117), dials, bar charts or line charts. May correlate with Trend, as they can co-exist.

### Number of metrics

**Description:** The number of quality indicators in the feedback message that summarise clinical performance.

### Number of patients

**Description:** The feedback message provides lists of patients included in the calculation of clinical performance. This may include those patients (not) receiving desired care processes (e.g. 1591, 2857), or those who have experienced a particular outcome of interest (e.g. 5532). Correlates with Detailed patient-level information, as they often co-exist.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effect and mechanism (2a – model tension): Increased number metrics may facilitate Tunnel vision by decreasing Resource match (leaving less time to address patients' concerns).</td>
<td>Confidence: Low</td>
<td>Reason: Substantial concerns regarding the coherence, adequacy, relevance, and methodological limitations of findings.</td>
<td>Papers (n=2): 6087, S109</td>
</tr>
<tr>
<td></td>
<td>Effect and mechanism (2b – model tension): Increased number metrics may reduce Tunnel vision by increasing Compatibility (ensuring the metrics comprehensively cover what is considered to be important aspects of clinical care).</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of metrics</th>
<th>Effect and mechanism (1): Increased number of metrics inhibits Perception and Intention by increasing Complexity (making the feedback message as a whole more difficult to understand and plan to address).</th>
<th>Confidence: Moderate</th>
<th>Reason: Minor concerns regarding adequacy of findings.</th>
<th>Papers (n=7): 1271, S1, S105, S112, S127, S154, S38</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effect and mechanism (2): Facilitates Perception, Intention and Behaviour (patient-level) by increasing Actionability (by providing the identity of patients who may have received poor care, so that it can be corrected)</td>
<td>Confidence: High</td>
<td>Papers (n=12): 2857, 7194, S1, S117, S38, 1591, 187, 3351, S112, S32, S38</td>
<td></td>
</tr>
</tbody>
</table>
and decreasing Complexity (enabling recipients to easily understand how their clinical performance may be suboptimal by reviewing the care of individual patients).

<table>
<thead>
<tr>
<th>Performance level</th>
<th>Description: The level of clinical performance communicated by the feedback message perceived by the recipient.</th>
<th>Papers (n=40):</th>
<th>Models: CT, Ilgen (Sign), FIT (Sign)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect and mechanism (1): Low performance level facilitates Verification by decreasing Compatibility (as health professionals often believe they provide good care, so they are motivated to check the data).</td>
<td>Confidence: High</td>
<td>Papers (n=5): S104, S118, S127, S30, S6</td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (2a – model tension): Low/high performance level facilitates/inhibits Intention, Behaviour and Performance improvement by increasing/decreasing Compatibility (as improving clinical performance aligns with their motivations [self-affirmation, cognitive dissonance, control theory, feedback intervention theory, goal theory]), and increasing/decreasing Actionability (low performance implies there is room for improvement).</td>
<td>Confidence: High</td>
<td>Papers (n=29): 1271, 1948, 2794, 2857, 3351, 5033, 5857, 7049, 7050, 7194, 7301, 8167, S104, S109, S132, S136, S150, S16, S17, S19, S127, S28, S30, S32, S38, S52, S6, S67, S71</td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (2b – model tension): Low performance level inhibits Acceptance by decreasing Compatibility with expectations (as health professionals often believe they provide good care [self-affirmation theory, cognitive dissonance theory, feedback intervention theory - Eliminating Feedback-Standard Gap strategies]) and/or Resource match (as they do not have the resources – time, skills, money – to spend on increasing performance).</td>
<td>Confidence: Moderate</td>
<td>Reason: Some concerns regarding coherence of findings – often non-acceptance may be explained by other variables.</td>
<td>Papers (n=17): 187, 2794, 4222, 5857, 7049, 7194, 7301, 8167, S1, S104, S132, S158, S16, S17, S19, S28, S30, S7</td>
</tr>
<tr>
<td>Effect and mechanism (3): Low/high performance level causes negative/positive Emotions (ripple effect) (e.g. disappointment) by decreasing/increasing Compatibility with expectations (as health professionals often believe they provide good care).</td>
<td>Confidence: High</td>
<td>Papers (n=18): 1271, 187, 2794, 512, 7025, S1, S104, S109, S117, S127, S132, S150, S158, S16, S17, S19, S28, S4</td>
<td></td>
</tr>
<tr>
<td>Prioritisation</td>
<td>Description: The feedback message includes design features to effectively summarises and communicates the relative importance of its contents e.g. patients most at risk S1, or highlighting areas of clinical performance that require most urgent attention S127.</td>
<td>Papers (n=11): 2857, 3351, 5033, 7194, S1, S105, S127, S136, S30, S38, S6</td>
<td>Models: Nil</td>
</tr>
<tr>
<td>Effect and mechanism (1): Its absence inhibits Interaction, Perception,</td>
<td>Confidence: High</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intention and Behaviour by increasing Complexity (making the feedback message easier to interpret and decide the most important aspects on which to focus), and decreasing Actionability (by making it unclear which aspects of clinical performance require urgent attention) and decreasing Relative advantage (as most existing health information systems do not prioritise information to users).</td>
<td>Papers (n=8): S1, S105, S127, S136, S30, S38, S6, 3351, 7194</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detailed patient-level information Description: The feedback message includes more detailed information than that provided by lists regarding the patients included in the calculation of clinical performance. For example, latest blood test results (e.g. S32). Correlates with Patient lists, as they often co-exist.</td>
<td>Papers (n=8): 1948, 2023, 2857, 7694, S127, S16, S32, S62 Models: Igen (Supporting data)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (1): Facilitates Verification and Acceptance by increasing Credibility (demonstrating to recipients how their clinical performance was calculated).</td>
<td>Confidence: Moderate Reason: Some concerns regarding adequacy of findings. Papers (n=3) S127, S32, 7694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (2): Facilitates Perception, Intention and Behaviour (patient-level) by increasing Actionability and decreasing Complexity (enabling recipients to easily understand how their clinical performance may be suboptimal, and which patients to target where appropriate).</td>
<td>Confidence: Moderate Reason: Some concerns regarding adequacy of findings. Papers (n=5): 1948, 2857, S127, S32, S62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (3): Facilitates Patient-level behaviour by increasing Actionability for individual patients.</td>
<td>Confidence: Moderate Reason: Some concerns regarding adequacy of findings. Papers (n=3): 1948, 2857, S62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative data Description: The feedback message includes qualitative comments about clinical performance. For example from individual patients (e.g. 512).</td>
<td>Papers (n=5): 1948, 512, 7694, S7, 5033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (1): Facilitates Intention and Behaviour by increasing Credibility (demonstrating how their clinical performance may impact individual patients) and Compatibility with professional role.</td>
<td>Confidence: Low Reason: Substantial concerns regarding adequacy of findings. Papers (n=3) 1948, 512, 5033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting Description: The feedback message is perceived by the recipient to also be communicated to additional entities. This may include external organisations such as governments (e.g. 2249) and insurance companies (e.g. S109), members of the public (e.g. S17), or those internal to the organisation such as managers (e.g. 1948). Correlates with Function, Ownership and Source – Location, as they often co-exist.</td>
<td>Papers (n=14): 1948, 2249, 5033, S104, S109, S127, S159, S17, S19, S28, S38, S67, S7 Models: Nil.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (1): Increases negative Emotions (ripple effect)</td>
<td>Confidence: High</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(e.g. anxiety, frustration) by decreasing *Relative advantage* and *Compatibility with professional role* and *motivation* (as health professionals feel this violates their autonomy and questions their motivation to improve care quality).

**Effect and mechanism (2):** Leads to *Gaming* (unintended consequence) because health professionals want to appear high performing to both themselves (*Compatibility* [self-affirmation, Cognitive dissonance]) and to others *Social influence* [reference group behaviour]).

Confidence: Low
Reason: Substantial concerns regarding adequacy of findings.

Papers (n=2): 1948, S67

| Source – knowledge and skill | Description: The perceived level of appropriate knowledge and skill of the person / organisation delivering the feedback message. This may relate to clinical knowledge and skills (e.g. 1591), or those relating to technical aspects of quality improvement (e.g. 2249). Correlates with *Source – Location*, as they often co-exist. | Papers (n=22): 1271, 1591, 2249, 5357, 5857, 7194, 7783, S1, S104, S112, S136, S19, S25, S28, S30, S38, S62, S67, S71, S81
Models: BCTT (9.1 Credible source) |
| --- | --- | --- |
| **Effect and mechanism (1):** Greater source knowledge and skill facilitates *Acceptance* and *Intention* by increasing *Credibility* and *Relative advantage* (as the opportunity to receive feedback from a credible source is valued by health professionals, and aligns with their sense of autonomy). | Confidence: High
| **Effect and mechanism (2):** Less source knowledge and skill increases negative *Emotions* (ripple effect) (e.g. frustration) by decreasing *Relative advantage* and *Compatibility with professional role* and *motivation* (as health professionals feel this violates their sense of autonomy). | Confidence: Moderate
Reason: Some concerns regarding adequacy of findings.
Papers (n=4): 2249, 5857, 7194, S28, S30 |
| **Source – Location** | Description: Whether the source is perceived to be internal or external to the health professionals’ organisation. Correlates with *Source – knowledge and skill*, as they often co-exist. | Papers (n=9): 1271, 2249, 5857, 7194, S28, S30, S38, S6, S67 |
| **Effect and mechanism (1):** External source inhibits *Acceptance* by reducing *Compatibility* with recipients’ sense of autonomy. | Confidence: Moderate
Reason: Some concerns regarding adequacy and coherence of findings.
Papers (n=5): 2249, 5857, 7194, S28, S30 |
| **Effect and mechanism (2):** External source increases negative *Emotions* (ripple effect) (e.g. frustration) by reducing *Compatibility with professional role* (as health professionals feel this violates their autonomy). | Confidence: Low
Reason: Substantial concerns regarding adequacy of findings.
Papers (n=4): S38, S6, S67 |
<table>
<thead>
<tr>
<th>Specificity</th>
<th>Description: The degree to which the feedback message presents the clinical performance of an individual clinician (e.g. 1948) versus their wider team (e.g. S7) or organisation (e.g. S19).</th>
<th>Papers (n=13): 1948, 3351, 5033, 512, 5532, S104, S127, S158, S19, S30, S4, S6, S7 Models: Ilgen (Feedback specificity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effect and mechanism (1): Increased specificity facilitates Interaction, Perception, Acceptance, Intention, Behaviour and Performance improvement by increasing Actionability (because the feedback message refers directly to the recipients' behaviour over which they have most control), and Relative advantage.</td>
<td>Confidence: High Papers (n=11): 1948, 3351, 512, 5532, S19, S104, S30, S158, S19, S6, S7</td>
</tr>
<tr>
<td></td>
<td>Effect and mechanism (2): Increased specificity facilitates Patient-level behaviour by increasing Actionability (because the feedback message refers directly to the recipients' behaviour over which they have most control) for individual patients the recipient cares for, rather than the wider organisation.</td>
<td>Confidence: Moderate Reason: Concerns regarding the adequacy of findings. Papers (n=4): 3351, 512, 5532, S7</td>
</tr>
<tr>
<td></td>
<td>Effect and mechanism (3): Improves Teamwork (ripple effect) by using Social influence (reference group behaviour) to enable staff members within an organisation to understand their specific role in the delivery of care.</td>
<td>Confidence: Low Reason: Substantial concerns regarding adequacy and coherence of findings. Paper (n=3): 3351, 5532, S7</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Description: How quickly the feedback message is sent relative to the time the clinical performance actually occurred. This could be near real-time (e.g. 1948) or years (e.g. 5033).</td>
<td>Papers (n=17): 1948, 2841, 2857, 3351, 5033, 5532, 5857, 7816, S1, S104, S105, S118, S120, S132, S154, S17, S25 Models: Ilgen (Timing), A&amp;F modifiable elements (Lag time)</td>
</tr>
<tr>
<td></td>
<td>Effect and mechanism (1): Increased timeliness facilitates Interaction, Perception, Acceptance, Intention, Behaviour and Performance improvement by increasing Actionability (because more timely data is easier to action), Credibility (because more timely data represents a more current picture of performance), and Relative advantage (because many A&amp;F interventions do not provide timely data).</td>
<td>Confidence: High Papers (n=15): 1948, 2841, 2857, 5857, 3351, S105, S118, 5033, 5532, S1, S120, S104, S132, S17, S25</td>
</tr>
<tr>
<td></td>
<td>Effect and mechanism (2): Improves Teamwork (ripple effect) by using Social influence (reference group behaviour) to enable staff members within an organisation to understand their specific role in the delivery of care.</td>
<td>Confidence: Low Reason: Concerns regarding the adequacy of findings. Paper (n=3): 3351, 5532, S7</td>
</tr>
<tr>
<td>Trend</td>
<td>Description: The feedback message provides information on the recipients' historical clinical performance, in addition to their most current. May correlate with Graphical, as often trend data is presented as line graphs.</td>
<td>Papers (n=18): 187, 5033, 512, 5357, 5532, 5857, 7049, 7050, S105, S118, S127, S132, S150, S17, S175, S4, S7, S71 Models: FIT (Velocity)</td>
</tr>
</tbody>
</table>
| Effect and mechanism (2): Can increase **Observability** (ripple effect) by demonstrating how health professionals' clinical performance has improved over time with the A&F intervention. | Confidence: Moderate  
Reason: Minor concerns about adequacy and coherence of findings.  
Papers (n=7): 5033, 5357, 5532, 5857, 7049, 7050, S7 |
|---|---|
| **Usability** | **Description:** The perceived user-friendliness and clarity of the feedback message.  
Models: Nil. |
| **Goal** | **Evidence base** | **Description:** The evidence base supporting the goal against which clinical performance is measured. This may be derived from clinical guidelines (e.g. S105), research studies (e.g. S120) or expert opinion (e.g. S38).  
Papers (n=20): 1271, 2023, 2249, 7049, 7050, 7816, 8167, S1, S105, S112, S117, S120, S132, S14, S30, S38, S4, S6, S7, S71  
Models: FIT (Norms) |
| **Effect and mechanism (1): Facilitates **Interaction, Perception, Acceptance, Intention, and Behaviour** by increasing **Relative advantage** (because existing clinical information systems may be less user friendly) and **Resource match** (by requiring less resource to interpret its findings), and decreasing **Complexity** (making the feedback message easier to understand and act upon). | Confidence: High  
| **Effect and mechanism (1): Facilitates **Acceptance** by increasing **Credibility** and **Compatibility** with recipients' motivation (ensuring they are providing evidence-based care). | Confidence: Moderate  
Reason: Minor concerns regarding the coherence of findings (some papers reported recipients not agreeing with evidence-based measures i.e. Compatibility 1271, 2249, S1).  
Papers (n=15): 1271, 2023, 2249, 7050, 7816, 8167, S1, S105, S112, S117, S120, S132, S14, S30, S38, S4, S6, S7, S71  
Models: FIT (Process-outcome), A&F modifiable elements (Process / Outcome) |
| **Process vs outcome** | **Description:** Whether clinical performance is measured regarding care processes (e.g. 5033) or patient outcomes (e.g. S127). Can correlate with **Accuracy** (and **Benchmarking**), as outcome measures are often perceived as inaccurate when comparing health professionals if not analysed with sufficient casemix adjustment, and **Controlability** as often process measures are perceived to be within health professionals' control.  
Papers (n=12): 1591, 5033, 5352, 5857, 8167, S120, S127, S16, S30, S32, S7, S71  
Models: FIT (Process-outcome), A&F modifiable elements (Process / Outcome) |
<p>| **Effect and mechanism (1): Process measures facilitate <strong>Perception</strong>, | Confidence: Moderate |</p>
<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
<th>Effect and mechanism (1)</th>
<th>Effect and mechanism (2)</th>
<th>Reason</th>
<th>Confidence</th>
<th>Papers</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance, Intention, Behaviour (patient-level), and Performance improvement</td>
<td>by decreasing Complexity (the consequence of process measures are easier to understand, whereas outcome measures often require further statistical analysis e.g. S120), and increasing Actionability (process measures are easier to influence) and Credibility (the attribution of process measures to a health professional's clinical performance are more believable).</td>
<td></td>
<td></td>
<td>Reason: Minor concerns regarding adequacy and consistency of findings. Papers (n=10): S120, S127, S16, S30, S32, S71, 5033, S7, 5532, 1591</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (2): Outcome measures lead to Gaming (unintended consequence) by encouraging health professionals to increase Actionability via other means.</td>
<td></td>
<td></td>
<td></td>
<td>Confidence: Low Reason: Substantial concerns regarding adequacy of findings. Papers (n=2): 8167, S127</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target</td>
<td>Description: An expected achievement level is set for clinical performance, generally according to expert opinion. This is different to Benchmarking, where others’ clinical performance is presented but without explicit judgment as to what levels of achievement are expected.</td>
<td></td>
<td></td>
<td>Papers (n=4): S109, S117, S154, 2794 Models: FIT (Normative information; Goal setting)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (1): Facilitates Goal setting, Perception and Intention by decreasing Complexity (making it easier for recipients to know what constitutes 'good performance' and therefore what requires a corrective response).</td>
<td></td>
<td></td>
<td></td>
<td>Confidence: Low Reason: Substantial concerns regarding adequacy and relevance of findings. Papers (n=2): S117, S154</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (2): Inhibits Acceptance by decreasing Credibility if targets are set too high.</td>
<td></td>
<td></td>
<td></td>
<td>Confidence: Low Reason: Substantial concerns regarding adequacy and relevance of findings. Papers (n=2): S117, 2794</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controllability</td>
<td>Description: The degree to which the topic of the feedback message, and any changes that need to be made in response to suboptimal performance, is perceived to be within the control of the recipient.</td>
<td></td>
<td></td>
<td>Papers (n=23): 1271, 187, 2249, 5532, 7025, 7301, 8167, S1, S104, S112, S117, S118, S120, S127, S132, S136, S154, S16, S25, S30, S52, S6, S71 Models: Nil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance</td>
<td>Description: The relevance of the feedback message topic to the recipients' job.</td>
<td></td>
<td></td>
<td>Papers (n=8): 7301, S120, S132, S136, S25, S52, S6, S71 Models: Diffusion of Innovations (Task relevance)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (1): Increased relevance facilitates Audit (when</td>
<td></td>
<td></td>
<td></td>
<td>Confidence: High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Importance</td>
<td>Implementation process</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description: The perceived clinical importance of the topic within the feedback message according to the recipient. An important topic will: 1) represent 'good' clinical care, 2) be comprehensive (i.e. measure all aspects of important clinical care), and 3) address a perceived quality problem (i.e. a clinical area in which performance is suboptimal).</td>
<td><strong>Importance</strong></td>
<td>Papers (n=16): 1271, 2841, 3351, 4222, 5357, 7049, 7050, 7194, S117, S118, S127, S132, S15, S62, S67, S81 Models: Diffusion of innovations (Re-invention), CFIR (Adaptability)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect and mechanism (1): Increased importance facilitates Interaction, Acceptance, Intention, Behaviour, and Performance improvement by increasing Credibility (because the recipient believes it is an important topic), Compatibility (with recipients' views on what is clinically significant) and Relative advantage (as they may not receive information about important topics from elsewhere).</td>
<td><strong>Effect and mechanism (1)</strong></td>
<td>Confidence: High Papers (n=21): 1591, 187, 2023, 2249, 2794, 2857, 4222, 512, 5857, 6087, 7049, 7050, 7194, 7694, 7816, 8167, S127, S16, S28, S30, S32, S6, S7 Models: Diffusion of innovations (Re-invention), Locke and Latham (Goal commitment)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description: Whether or not the A&amp;F intervention can be tailored to the context into which it is implemented to meet their perceived needs. This may relate to any aspect of the A&amp;F process e.g. design of the feedback message or how the data is collected during audit.</td>
<td><strong>Adaptability</strong></td>
<td>Papers (n=8): 7301, S120, S132, S136, S25, S52, S6, S71 Papers (n=28): 1591, 187, 1948, 2023, 2249, 2794, 2857, 4222, 5033, 512, 5857, 6087, 7049, 7050, 7194, 7694, 7816, 8167, 8249, S1, S127, S16, S28, S30, S32, S38, S6, S7 Models: Diffusion of innovations (Re-invention), CFIR (Adaptability)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description: The perceived cost of the intervention. This may relate to costs of time, human or financial resources. May correlate with Conducted by recipients and Active delivery (face-to-face), as this often equates to high costs.</td>
<td><strong>Effect and mechanism (1)</strong></td>
<td>Papers (n=22): 1271, 1591, 187, 1948, 5857, 6087, 627, 7816, S1, S105, S109, S112, S118, S120, S127, S132, S136, S14, S15, S154, S16, S25. Models: Diffusion of innovations (Re-invention), CFIR (Adaptability)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
and cultural backgrounds.

| Linkage at the development stage | Description: The feedback recipients have input into the design and implementation of the A&F intervention. | Papers (n=11): 2794, 7025, 7694, 8249, S109, S112, S117, S30, S38, S62, S81
Models: Diffusion of innovations (Linkage at the development stage) |
| --- | --- | --- |
| Effect and mechanism (1): Facilitates Goal setting, Audit, Interaction, Acceptance, Intention, and Performance by increasing Compatibility (with recipients' motivation and organisational goals and systems), Credibility, and Relative advantage. | Confidence: Moderate
Reason: Minor concerns regarding the coherence of findings.
Papers (n=9): 7025, 7694, 8249, S109, S112, S117, S30, S62, S81 |
| Observability | Description: The benefits of the A&F intervention are visible to feedback recipients. Correlates with Trend (ripple effect), as this is one way of demonstrating the benefits of involvement with an A&F intervention. | Papers (n=12): 3351, 5033, 5357, 5532, 5857, 7025, 7049, 7050, 7694, S132, S4, S7.
Models: Diffusion of innovations (Observability) |
| Effect and mechanism (1): Facilitates Goal setting, Audit (when Conducted collected by recipients), Interaction, Acceptance, Intention, Behaviour, and Performance by increasing Relative advantage (making its advantages more visible). | Confidence: High
Papers (n=12): 3351, 5033, 5357, 5532, 5857, 7025, 7049, 7050, 7694, S132, S4, S7 |
| Effect and mechanism (2): Increases A&F attitude (ripple effect) by increasing Relative advantage. | Confidence: Moderate
Reason: Substantial concerns regarding the adequacy of findings.
Papers (n=4): 7025, 7049, 7050, S132 |
| Effect and mechanism (3): Increases positive Emotion (ripple effect) by increasing Relative advantage. | Confidence: Low
Reason: Substantial concerns regarding the adequacy and coherence of findings.
Papers (n=3): 5532, S132, S4 |
| Ownership | Description: The degree of perceived ownership of the A&F intervention the recipients have, versus how much they feel it has been enforced upon | Papers (n=20): 1948, 2023, 2249, 5033, 5357, 6087, 7025, 7194, 7816, 8249, S1, S112, S117, S127, S136, S15, S19, |
| Effect and mechanism (1a – model tension): Greater ownership facilitates **Goal setting, Audit** (when **Conducted collected by recipients**), **Interaction, Acceptance, Intention, Behaviour, and Performance** by increasing **Compatibility** (with recipients’ motivation to provide high quality care, and their sense of autonomy), and therefore also its **Relative advantage**. | Confidence: High  
**Papers (n=11):** 2023, 2249, 5033, 5357, 6087, 7025, 7816, 8249, S118, S19, S81  
Models: **Diffusion of innovations** (Political directives), **CFIR** (External Policy & Incentives). |
|---|---|
| Effect and mechanism (1b – model tension): Less ownership and more forced implementation facilitates **Goal setting, Audit** (when **Conducted collected by recipients**), **Interaction, Acceptance, Intention, Behaviour, and Performance** via **Social influence** (with recipients responding to authority figures [Authority] and not wanting to appear as poor performers [Reference group behaviour]). | Confidence: Moderate  
**Reasons:** Concerns regarding the adequacy and methodological limitations of findings.  
**Papers (n=5):** 1948, S81, S136, S15, S4. |
| Effect and mechanism (2): Increases positive **Emotions** (ripple effect) by increasing **Compatibility** (with recipients’ autonomy and motivation to provide high quality care). | Confidence: Moderate  
**Reasons:** Minor concerns regarding the adequacy of findings.  
**Papers (n=4):** 1948, S38, S67, S81. |
| **Training and support**  
**Description:** Training and support is provided to A&F recipients regarding the intervention (not the clinical topic as in **Clinical education**). | **Papers (n=24):** 2794, 3351, 4222, 5033, 5235, 5857, 627, 7025, 7049, 7050, 7816, S118, S120, S127, S132, S15, S154, S175, S176, S30, S33, S38, S6, S62.  
Models: **Diffusion of innovations** (**Assessment of implications**). |
| Effect and mechanism (1): Increases **Knowledge and skills – quality improvement** (ripple effect). | Confidence: High  
**Papers (n=21):** 3351, 4222, 5033, 5235, 5857, 627, 7025, 7049, 7050, 7816, S118, S120, S127, S132, S15, S154, S175, S176, S33, S38, S62. |
| Effect and mechanism (2): Increases **Teamwork**, especially if educational sessions delivered to a multidisciplinary group. | Confidence: Low  
**Reasons:** Substantial concerns regarding methodological limitations, and coherence and adequacy of findings.  
**Papers (n=2):** 5357, S176 |

### Organisational context

**Champion**  
**Description:** Individuals within an organisation who dedicate themselves | **Papers (n=26):** 1591, 1948, 2023, 2841, 2857, 3351, 4222,
<table>
<thead>
<tr>
<th>Interventions</th>
<th>Effect and mechanism (1): Facilitates Goal setting, Audit and Feedback (when Conducted collected by recipients), Interaction, Intention, Behaviour, and Performance by increasing Resource match.</th>
<th>Competing priorities Description: Number and relative importance of other jobs or responsibilities relevant to feedback recipients in addition to the A&amp;F intervention, which are sufficiently different from the A&amp;F intervention. (NB competing priorities similar to the A&amp;F intervention may be described as Similar quality improvement interventions). Correlates with Resource, because competing priorities reduce resources.</th>
<th>Effect and mechanism (1): Inhibits Audit and Feedback (when Conducted collected by recipients), Interaction, Intention, Behaviour, and Performance by decreasing Resource match – especially if there is reduced Compatibility between the competing priorities and A&amp;F intervention.</th>
<th>Effect and mechanism (2): Leads to Tunnel vision (unintended consequence) by decreasing Resource match (less resource is available to focus on other aspects of care other than those measured by the A&amp;F intervention).</th>
<th>Effect and mechanism (3): Increases negative Emotions (ripple effect) by decreasing Resource match (making working life more stressful) and Compatibility (reducing the ability of the recipient to focus on aspects of clinical work they feel may be more important).</th>
<th>Similar quality improvement interventions Description: Presence of other quality improvement initiatives in addition to the A&amp;F intervention. They may focus on similar clinical topics. Correlates with Competing priorities, as similar quality improvement initiatives may be viewed as such.</th>
</tr>
</thead>
</table>
| Effect and mechanism (1a – model tension): Facilitates Goal setting, Audit and Feedback (when Conducted collected by recipients), Interaction, Perception, Acceptance, Intention, Behaviour, and Performance improvement by increasing Compatibility (with existing workflows and priorities), Resource match (as resources to engage in A&F may already exist from the other initiatives), and Relative advantage (if the A&F intervention has distinct advantages over the other interventions). | Confidence: Moderate  
Reason: Concerns regarding coherence of data.  
Papers (n=18): 187, 2841, 5357, 7049, 7050, 7816, 8167, S1, S112, S118, S127, S136, S15, S17, S175, S38, S62, S67 |
| Effect and mechanism (1b – model tension): Inhibits Goal setting, Audit and Feedback (when Conducted collected by recipients), Interaction, Perception, Acceptance, Intention, Behaviour, and Performance improvement by decreasing Resource match (if resources from engaging in the other initiatives are already used) and Relative advantage (if the A&F intervention does not have obvious advantages over the other interventions), and increasing Complexity (by making it unclear which intervention should be given priority). | Papers (n=9): 1591, 7025, 7049, 7050, 7816, S105, S132, S15, S38  
Models: Diffusion of Innovations (Extra-organisational networks), CFIR (Cosmopolitanism) |
| Extra-organisational networks  
Description: The effectiveness of A&F recipients' organisations' communications with other organisations. Correlates with Social support through a ripple effect. | Paper (n=9): 1591, 7025, 7049, 7050, 7816, S105, S132, S15, S38  
Models: Diffusion of Innovations (Extra-organisational networks), CFIR (Cosmopolitanism) |
| Intra-organisational networks  
Description: The effectiveness of communications with the A&F recipients' organisations. Correlates with Teamwork, as good intra-organisational networks are a feature of good teamwork. Correlates with Social support and Delivery to a group through ripple effects (moderator confidence). | Papers (n=22)  
Models: Diffusion of Innovations (Intra-organisational networks), CFIR (Networks & Communications) |
| Effect and mechanism (1): Facilitates Audit and Feedback (when Conducted by recipients), Interaction, Intention, Behaviour, and Performance improvement by increasing Actionability (providing practical support between colleagues on producing, communicating and responding effectively to feedback messages) and Compatibility (with existing communication channels). | Confidence: High  
Papers (n=19): 2857, 4222, 5033, 5532, 627, 7049, 7050, 8167, S105, S109, S117, S132, S15, S175, S32, S38, S6, S71, S81 |
| **Leadership support** | Description: Advocacy for the A&F intervention from members of top management within the A&F recipients’ organisation. | Papers (n=23): 1271, 1948, 2023, 4222, 512, 5357, 5532, 7025, 7050, 7816, 8167, S105, S117, S118, S132, S154, S16, S176, S19, S52, S62, S71, S81 Models: Diffusion of Innovations (Staff involvement and commitment); CFIR (Leadership Engagement) |
| **Effect and mechanism (1a):** Facilitates *Audit and Feedback* (when Conducted by recipients), Interaction, Perception, Intention, Behaviour, and Performance improvement by increasing Social influence (encouraging, permitting, and instructing staff to engage and respond [Influence theory – Authority]), Resource match (providing additional resources to engage with the A&F intervention as necessary), and Credibility (setting an example to health professionals to engage with the A&F intervention). | Confidence: High Papers (n=23): 1271, 1948, 2023, 4222, 512, 5357, 5532, 7025, 7050, 7816, 8167, S105, S117, S118, S132, S154, S16, S176, S19, S52, S62, S71, S81 |
| **Effect and mechanism (1b):** Facilitates Organisational-level behaviour by increasing Social influence (encouraging, permitting, and instructing staff to engage and respond [Influence theory – Authority]), Resource match (providing additional resources to engage with the A&F intervention as necessary), and Credibility (setting an example to health professionals to engage with the A&F intervention). | Confidence: High Papers (n=6): 1948, 512, 8167, S118, S81, S16 |
| **Opinion leaders** | Description: Advocacy for the A&F intervention from members of staff within the A&F recipients’ organisation who exert either formal or informal influence on the attitudes and beliefs of their colleagues through their authority, status, and credibility. Correlates with Champions and Leadership support because often opinion leaders are often all three. | Papers (n=10): 1591, 2023, 5357, 5857, 7816, 8167, S132, S15, S176, S62 Models: Diffusion of Innovations (Opinion leaders); CFIR (Opinion leaders) |
| **Effect and mechanism (1):** Facilitates Interaction, Intention, Behaviour, and Performance improvement by increasing Social influence (encouraging health professionals to engage with the A&F intervention [Influence theory – Authority and Liking]), and Credibility (setting an example to health professionals to engage with the A&F intervention). | Confidence: Moderate Reason: Concerns regarding the adequacy and coherence of findings. Papers (n=7): 2023, 5357, 5857, 8167, S132, S15, S176, S62 |
| **Resource** | Description: The amount of material and non-material resource available within the A&F recipients' organisation, including financial resource, human resource, time, space, and equipment. Correlates with Competing priorities, because competing priorities reduce resources. | Papers (n=49): 1271, 1591, 187, 1948, 2023, 2794, 2841, 2857, 3351, 4222, 5033, 512, 5235, 5357, 5532, 5857, 6087, 627, 7025, 7049, 7194, 7301, 7694, 7783, 7816, 8167, S1, S104, S105, S109, S112, S117, S118, S120, S127, S132, S14, S154, S16, S176, S19, S52, S62, S71, S81 Models: Diffusion of Innovations (Resource); CFIR (Resource match) |

329
<p>| <strong>Teamwork</strong> | Papers (n=25): 2023, 2857, 3351, 4222, 5357, 5532, 6087, 627, 7049, 7050, 7301, 7783, 8167, S105, S117, S118, S120, S127, S132, S14, S15, S16, S17, S19, S25, S30, S32, S38, S4, S52, S6, S62, S67, S71, S81  Models: Diffusion of Innovations (Dedicated time and resources); CFIR (Available resources); TDF (Environmental Context and Resources). |
| Effect and mechanism (1b): Facilitates Organisational-level behaviour by increasing Resource match. | Confidence: High  Papers (n=7): 2023, 2857, 6087, 8167, S118, S19, S4 |
| Effect and mechanism (2): Leads to Tunnel vision (unintended consequence) by decreasing Resource match (less resource is available to focus on other aspects of care other than those measured by the A&amp;F intervention). | Confidence: Low  Reasons: Substantial concerns regarding the coherence and adequacy of the findings.  Papers (n=3): 1948, 512, 6087 |
| <strong>Staff turnover</strong> | Description: The rate of staff at the A&amp;F recipients’ organisation replaced by new employees.  Papers (n=7): 2794, 7049, 7050, S118, S120, S127, S14  Models: Nil |
| Effect and mechanism (1): Inhibits Behaviour, and Performance improvement by decreasing Resource match (new staff require additional resource to train them in and make them aware of the A&amp;F intervention). | Confidence: Moderate  Reason: Minor concerns regarding the adequacy of findings.  Papers (n=7): 2794, 7049, 7050, S118, S120, S127, S14 |
| <strong>Teamwork</strong> | Description: The ability of the A&amp;F recipients’ organisation to work together effectively or cohesively towards a common goal. Correlates with Intra-organisational networks, as good intra-organisational networks are often a feature of good teamwork.  Papers (n=25): 2023, 2857, 3351, 4222, 5357, 5532, 6087, 627, 7049, 7050, 7301, 7783, 8167, S105, S117, S118, S132, S15, S175, S32, S38, S6, S67, S71, S81  Models: Nil |
| Effect and mechanism (1a): Facilitates Audit and Feedback (when Conducted by recipients), Interaction, Perception, Intention, Behaviour, and Performance improvement by increasing Actionability (providing practical support between colleagues on producing, communicating and responding effectively to feedback messages). | Confidence: High  Papers (n=22): 2023, 2857, 3351, 4222, 5352, 627, 7049, 7050, 7301, 7783, 8167, S105, S117, S118, S132, S15, S175, S32, S38, S6, S67, S71 |
| Effect and mechanism (1b): Facilitates Organisational-level behaviour by increasing Actionability (providing practical support between colleagues on responding effectively to feedback messages). | Confidence: Moderate  Reason: Concerns regarding the adequacy of findings.  Papers (n=3): 2857, 8167, S118 |</p>
<table>
<thead>
<tr>
<th><strong>Workflow fit</strong></th>
<th>Description: The degree of alignment of the A&amp;F intervention with the systems used and processes conducted by the recipient or their wider organisation. Understandably, this depends on the specific context into which the intervention is being implemented.</th>
<th>papers (n=18): 187, 2841, 2857, 3351, 7194, 7301, 7816, S1, S105, S112, S132, S14, S15, S176, S32, S6, S62, S67, S71 Models: Nil</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effect and mechanism (1):</strong> Facilitates <em>Audit</em> and <em>Feedback</em> (when <em>Conducted by recipients</em>), <em>Interaction</em>, <em>Perception</em>, <em>Acceptance</em>, <em>Intention</em>, <em>Behaviour</em>, and <em>Performance improvement</em> by increasing <em>Compatibility</em> (with existing workflows and systems), <em>Actionability</em> (ensuring they can engage with the A&amp;F intervention during their working lives) and reducing <em>Complexity</em> (by reducing the need to change their workflows to integrate the A&amp;F intervention).</td>
<td>Confidence: High</td>
<td>Papers (n=18): 187, 2841, 2857, 3351, 7194, 7816, S1, S105, S112, S132, S14, S15, S176, S32, S6, S62, S67, S71</td>
</tr>
<tr>
<td><strong>Patient population</strong></td>
<td><strong>Choice misalignment</strong></td>
<td>Description: Reasons for patients receiving suboptimal (measured) clinical care relating to their decisions or expectations. Relevant papers (n=16)</td>
</tr>
<tr>
<td><strong>Effect and mechanism (1):</strong> Inhibits <em>Acceptance</em>, <em>Intention</em>, <em>Behaviour</em>, and <em>Performance improvement</em> by decreasing <em>Actionability</em> (either the health professionals have little control over the care provided – leading to non-acceptance, or they cannot act upon it to improve) and <em>Compatibility</em> (with health professionals’ goals to provide patient-centred care).</td>
<td>Confidence: High</td>
<td>Papers (n=13): 187, 6087, 7049, 7050, 7194, 7301, S104, S127, S28, S30, S32, S38, S6</td>
</tr>
<tr>
<td><strong>Effect and mechanism (2):</strong> Leads to <em>Gaming</em> and <em>Inappropriate care</em> (unintended consequence) especially in the presence of <em>Financial reward</em> in an attempt to increase <em>Resource match</em> (to maintain resource, attempts to preserve financial income are preserved at any cost).</td>
<td>Confidence: Low</td>
<td>Reason: Substantial concerns regarding adequacy of findings. Papers (n=3): 6087, S112, S38</td>
</tr>
<tr>
<td><strong>Effect and mechanism (3):</strong> Increases negative <em>Emotions</em> (ripple effect) by decreasing <em>Compatibility</em> (between recipients’ motivations and expectations, because health professionals generally strive to provide high quality patient-centred care).</td>
<td>Confidence: Low</td>
<td>Reason: Substantial concerns regarding adequacy of findings. Papers (n=3): S1, S112, S30</td>
</tr>
<tr>
<td><strong>Clinically inappropriate</strong></td>
<td>Description: Explanations for patients receiving suboptimal (measured) clinical care relating to the their specific clinical characteristics. This may</td>
<td>Papers (n=16): 1591, 187, 2794, 4222, 6087, 8167, S1, S104, S112, S117, S127, S17, S30, S32, S67, S7</td>
</tr>
</tbody>
</table>
include where guideline care is contra-indicated due to existing conditions (e.g. 187), or their response to treatment is diminished (e.g. S117).

<table>
<thead>
<tr>
<th>Models: Cabana Guideline model, Guidelines interdependence model</th>
</tr>
</thead>
</table>

### Effect and mechanism (1):
Inhibits *Acceptance, Intention, Behaviour, and Performance improvement* by decreasing *Actionability* (either the health professionals have little control over the care provided – leading to non-acceptance, or they cannot act upon it to improve) and *Compatibility* (with health professionals’ goals to provide patient-centred care).

Confidence: High
Papers (n=12): 1591, 187, 2794, 4222, 6087, S1, S104, S17, S117, S30, S32, S6

### Effect and mechanism (3):
Leads to *Gaming* (unintended consequence) in attempt to increase *Compatibility* (between recipients’ motivations and expectations), because health professionals generally strive to provide high quality patient-centred care. *Credibility* (recipients will want to interrogate clinical performance data to look for reasons for suboptimal care).

Confidence: Moderate
Reason: Concerns regarding the adequacy of findings.
Papers (n=4): 8167, S112, S127, S67

### Effect and mechanism (2):
Leads to *Verification* by decreasing *Compatibility* (between recipients’ motivations and expectations), because health professionals generally strive to provide high quality patient-centred care and *Credibility* (recipients will want to interrogate clinical performance data to look for reasons for suboptimal care).

Confidence: Low
Reason: Substantial concerns regarding the adequacy of findings.
Papers (n=2): 1591, S104

### Effect and mechanism (4):
Increases negative *Emotions* (ripple effect) by decreasing *Compatibility* (between recipients’ motivations and expectations), because health professionals generally strive to provide high quality patient-centred care.

Confidence: Low
Reason: Substantial concerns regarding adequacy of findings.
Papers (n=3): S1, S117, S30

### Health professional

#### Description:
Degree of positivity towards A&F in general (e.g. potential effectiveness S6), and related quality improvement and technical concepts (e.g. information technology).

Papers (n=24): 1271, 2249, 3351, 5357, 5857, 7301, 8249, S1, S112, S136, S14, S15, S16, S17, S175, S176, S28, S30, S38, S52, S6, S67, S7, S81
Models: TDF (Beliefs about Consequences)

#### Emotions

Description: A reaction to receiving feedback. May be positive (e.g. happy) or negative (e.g. sad, anxious, angry).

| Knowledge and skills – clinical | Description: The recipients’ awareness and understanding of the theory and performance of tasks relevant to the clinical performance topic in the A&F intervention. May correlate with Clinical education via a ripple effect (low confidence). | Papers (n=2): S104, S4  
Models: TDF (Emotion), Locke and Latham (Satisfaction) | Effect and mechanism (1): Positive emotions facilitate Intention, Behaviour, and Performance improvement by increasing Actionability (making the recipient feel they can positively influence situations). | S30, S32, S38, S4, S6, S67, S7, S81  
Models: TDF (Emotion), Locke and Latham (Satisfaction) | Confidence: Low  
Reason: Substantial concerns regarding adequacy of findings. |
| Knowledge and skills – quality improvement | Description: The recipients’ awareness and understanding of theory and performance of tasks regarding relating to quality improvement (including the A&F intervention, and A&F in general). Correlates with Training and support via a ripple effect (high confidence). | Papers (n=21): 2841, 3351, 4222, 5033, 5235, 5357, 5857, 627, 7049, 7050, 7816, 8249, S1, S118, S120, S127, S132, S15, S16, S175, S52, S6, S62, S7, S71, S81  
Models: TDF (Knowledge; Skills) | Effect and mechanism (1): Increased quality improvement knowledge and skills facilitates Goal setting, Audit and Feedback (when Conducted by recipients), Interaction, Perception, Acceptance, Intention, Behaviour (both Patient-level – by providing skills on how to interpret feedback – and Organisational-level – by providing skills to plan and act improvement plans), and Performance improvement by increasing Actionability (providing the recipient with technical skills and knowledge to improve their performance), Resource match (increasing resource of health professionals with the requisite knowledge and skills) and Credibility (of the feedback message when the recipient can interpret it more effectively), and decreasing Complexity (making the A&F intervention | 2841, 3351, 5033, 5235, 5357, 5857, 627, 7049, 7050, S1, S120, S127, S132, S15, S16, S175, S6, S62, S7, S71 | Confidence: High  
Reason: Substantial concerns regarding adequacy of findings. |
Effect and mechanism (2): Increased quality improvement knowledge and skills facilitates *Organisation-level* behaviour by increasing *Actionability* (providing the recipient with technical skills and knowledge to improve their performance) by providing skills on how to interpret feedback, and to plan and act on improvement plans.

**General ripple effects from A&F processes**

<table>
<thead>
<tr>
<th>Description</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Behaviour</em> (both patient and organisational-level) leads to increased audit <em>Accuracy</em> by increasing the quality of recorded clinical data in response to suboptimal clinical performance caused by poor record keeping.</td>
<td>Papers (n=8): 187, 2023, 2794, 7049, 7816, S132, S32, S6 Confidence: High</td>
</tr>
<tr>
<td><em>Behaviour</em> (organisational-level) can lead to increased <em>Resource</em> e.g. by recruiting new staff (e.g. 8167), purchasing new equipment (e.g. 5033), or freeing up existing resource as a result of organisational improvements (e.g. S32), in an attempt to improve clinical performance.</td>
<td>Papers (n=8): 5033, 7025, 8167, 8249, S118, S19, S25, S32 Confidence: High</td>
</tr>
<tr>
<td>The entire process of conducting A&amp;F increases recipients’ <em>Knowledge and skill – clinical</em> by reminding (e.g. S52, S6) or informing (e.g. S62, S30) them of important aspects of clinical performance they did or did not already know respectively (via <em>Credibility</em> and <em>Compatibility</em>).</td>
<td>Papers (n=20): 1591, 187, 2023, 4222, 5532, 627, 7049, 7050, S1, S105, S118, S127, S132, S158, S25, S30, S4, S52, S6, S62 Confidence: High</td>
</tr>
</tbody>
</table>
Appendix 6: User tasks and Goal-Action structure provided to system evaluators

Task 1

AGREE WITH THE SUGGESTED ACTION PLAN FOR MONITORING BLOOD PRESSURE: "INTRODUCE A TEXT-MESSAGING SERVICE TO REMIND PATIENTS THEY NEED A BLOOD PRESSURE CHECK"

and then
SELECT BLOOD PRESSURE
and then
SELECT MONITORING
and then
CHECK THE AVAILABLE OPTION "AGREE" FOR THE ACTION "INTRODUCE A TEXT-MESSAGING SERVICE TO REMIND PATIENTS THEY NEED A BLOOD PRESSURE CHECK"

... and then
CHECK THE AVAILABLE OPTION TO MARK THIS ACTION AS COMPLETED

Task 2

DISAGREE WITH THE SUGGESTED ACTION PLAN FOR TREATING ASTHMA: "NOMINATE AN ASTHMA LEAD IN YOUR PRACTICE WHO CAN INITIATE SOME OF THESE CHANGES"

and then
SELECT ASTHMA
and then
SELECT TREATMENT
and then
CHECK THE AVAILABLE OPTION DISAGREE FOR THE ACTION "NOMINATE AN ASTHMA LEAD IN YOUR PRACTICE WHO CAN INITIATE SOME OF THESE CHANGES"

and then
TYPE IN THE AVAILABLE FORM THE RESPONSE "ALREADY DONE THIS"

Task 3

AGREE WITH THE SUGGESTED ACTION PLAN "ADD CODE ‘9H31’ (PATIENT IS UNSUITABLE) FOR PATIENT 5556051664 WITH HIGH BLOOD PRESSURE, CURRENTLY RECEIVING PALLIATIVE CARE, WHO MAY BENEFIT FROM BEING EXCLUDED FROM QUALITY STANDARDS."

and then
SELECT BLOOD PRESSURE
and then
SELECT EXCLUSIONS
and then
SELECT PALLIATIVE CARE
and then
SELECT PATIENT 5556051664 FROM THE LIST
and then
CHECK THE AVAILABLE OPTION AGREE FOR THE ACTION "ADD CODE ‘9H31’ (PATIENT IS UNSUITABLE)"

and then
COPY THE CODE ‘9H31’
and then
PASTE THE CODE ‘9H31’ INTO THE SEPARATE EHR APPLICATION
and then
CHECK THE AVAILABLE OPTION TO MARK THIS ACTION AS COMPLETED

Task 4

ASSESS WHETHER THE APPLICATION HAS IDENTIFIED THE CORRECT IMPROVEMENT OPPORTUNITIES AND QUALITY STANDARDS FOR PATIENT 6662563783 WHO'S ASTHMA IS BEING MONITORED

and then
SELECT ASTHMA
and then
SELECT MONITORING
and then
SELECT PATIENT 6662563783
and then
COPY THE PATIENT’S NUMBER INTO THE SEPARATE ‘EHR’ APPLICATION

DISAGREE WITH AND CORRECT THE IMPROVEMENT OPPORTUNITY "NON FACE-TO-FACE" THAT THE APPLICATION HAS CATEGORISED THE PATIENT 6662563783 WHO'S ASTHMA IS BEING MONITORED

and then
TYPE IN THE AVAILABLE FORM THE CORRECT ACTION "SHOULD BE IN THE NO OPPORTUNITIES GROUP"

DISAGREE WITH AND CORRECT THE QUALITY STANDARD 'ANNUAL REVIEW MISSED' THAT THE APPLICATION SUGGESTS THAT THE SAME PATIENT HAS MISSED

and then
TYPE IN THE AVAILABLE FORM THE CORRECT STATUS 'HAD REVIEW AT WORK'
Task 5
IDENTIFY SOME DESCRIPTIVE STATISTICS AT THE POPULATION LEVEL ABOUT THE CARE PROVIDED IN YOUR PRACTICE FOR PATIENTS WITH ASTHMA.
IDENTIFY HOW MANY PATIENTS HAVE BEEN EXCLUDED FROM AASTHMA QUALITY STANDARDS FOR ‘REASONS WE THINK’
SELECT ASTHMA
and then CHECK THE CORRESPONDING FIGURE TO IDENTIFY HOW MANY PATIENTS HAVE BEEN EXCLUDED FROM ASTHMA QUALITY STANDARDS FOR ‘REASONS WE THINK’
IDENTIFY THE PERCENTAGE OF PATIENTS THAT HAD MONITORED ASTHMA ON 1ST APRIL 2015
CHECK THE CORRESPONDING FIGURE TO IDENTIFY THE PERCENTAGE OF PATIENTS THAT HAD MONITORED ASTHMA ON 1ST APRIL 2015
IDENTIFY HOW MANY PATIENTS HAVE HAD FACE-TO-FACE OPPORTUNITIES TO HAVE THEIR ASTHMA MONITORED
SELECT MONITORING
and then CHECK THE CORRESPONDING FIGURE TO IDENTIFY HOW MANY PATIENTS HAVE HAD FACE-TO-FACE OPPORTUNITIES TO HAVE THEIR ASTHMA MONITORED

Task 6
IDENTIFY SPECIFIC INFORMATION ABOUT THE BLOOD PRESSURE TREATMENT PROVIDED TO PATIENT 5554632673. SPECIFICALLY YOU NEED TO FIND WHAT HIS/HER BLOOD PRESSURE READING WAS ON THE 12TH SEPTEMBER 2013 AND WHAT DATE WAS HIS/HER BLOOD PRESSURE MEDICATION INCREASED.
IDENTIFY WHAT THE BLOOD PRESSURE READING OF PATIENT 5554632673 WAS ON 12TH SEPTEMBER 2013
SELECT BLOOD PRESSURE
and then SELECT TREATMENT
and then SELECT PATIENT 5554632673
and then CHECK THE AVAILABLE FIGURE TO IDENTIFY HIS /HER BLOOD PRESSURE READING ON 12TH SEPTEMBER 2013
IDENTIFY WHAT WAS THE DATE WHEN THE BLOOD PRESSURE MEDICATION OF PATIENT 5554632673 WAS INCREASED
CHECK THE AVAILABLE FIGURE TO IDENTIFY THE DATE HIS/HER BLOOD PRESSURE MEDICATION WAS INCREASED

Task 7
ADD AND EDIT YOUR OWN TEAM/ORGANISATION ACTION PLANS FOR BLOOD PRESSURE MONITORING
SELECT BLOOD PRESSURE
and then SELECT MONITORING
and then ENTER YOUR OWN ACTION PLAN: “EMPLOY AN ADDITIONAL NURSE” AT THE TEAM/ORGANISATION TAB
and then CHECK THE AVAILABLE OPTION TO DENOTE THAT THIS ACTION HAS BEEN COMPLETED
and then ENTER ANOTHER ACTION PLAN : “ADD ADDITIONAL SURGERIES”
and then EDIT THIS NEW ACTION PLAN TO SAY: “ADD SATURDAY SURGERIES”
and then DELETE THE ACTION PLAN : “ADD SATURDAY SURGERIES”.

TASK 8 (the purpose of this task was to test some general features of the functionality of the PINGR system. These were not tested as part of a goal – action structure because they represented only atomic actions that in our case did not take place in the context of a broader goal).
SEARCH FOR PATIENT 5557989507
ORDER THE LIST OF PATIENTS BY THEIR LATEST SBP
SWITCH QUALITY STANDARDS
DOWNLOAD YOUR ACTION PLAN FOR PRINTING
Appendix 7: Example usability issue data collection form

Participant ID : XX
Date : XX/XX/XX
Sheet no. XX
Task version XX

Sample of the *data collection form* (v.02)

<table>
<thead>
<tr>
<th>Usability issue</th>
<th>Action No.</th>
<th>Heuristic category</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colour of selector buttons is faded giving an initial impression that they are disabled.</td>
<td>1.3</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

*we encourage the use of screenshots*
The text for ‘date medication increased’ doesn’t match with the graph [in terms of finding what the date is-needs hover over/ could add date to the text]

| 6.1 | 2   | 3   |
## Appendix 8: Sample of the usability issue evaluation form

<table>
<thead>
<tr>
<th>Usability heuristic</th>
<th>Task where issue occurred</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Visibility of system status</strong></td>
<td>Not clear the system status after a suggested action has been checked as completed by the user. For instance it is not clear whether the action is actually saved.</td>
<td>Action 1.1</td>
</tr>
<tr>
<td></td>
<td>When disagreeing with a suggested action, the follow up “Why” dialog box prompting the user to justify his decision has lost context with the previous action because it covers this area of the screen. Users should know what “why?” refers to (i.e. to which action it refers to) without having to refer back to the greyed out strikethrough action.</td>
<td>Action 2.1</td>
</tr>
<tr>
<td></td>
<td>The path used to note the current position in the system, e.g. “blood pressure &gt; monitoring” should be made more visible.</td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>It is not made immediately visible the new status of the system when a new patient has been selected from the list of patients.</td>
<td>Actions 3.1 and 4.1</td>
</tr>
<tr>
<td></td>
<td>When you select one of the ‘improvement opportunities’ it is not immediately apparent that anything has changed in the patient list unless the number of patients in the list changes substantially.</td>
<td>Action 3.3</td>
</tr>
<tr>
<td></td>
<td>When switching between NICE and QQF the new system status is not visible.</td>
<td>Action 8.2</td>
</tr>
<tr>
<td><strong>Match between the system and the real world</strong></td>
<td>The (X) glyph used to represent exclusions (excluded patients) does not seem relevant. Usually this type of X symbols are used to indicate a forbidden action or exit from the current status/action.</td>
<td>Action 3.2</td>
</tr>
<tr>
<td></td>
<td>Difficult to browse the patient list – Ordering options not helpful and visible.</td>
<td>Action 6.1</td>
</tr>
<tr>
<td></td>
<td>Use of vague terminology in the captions of graphs – For example, What do you mean by ‘reasons we think’ who are we</td>
<td>Action 6.3</td>
</tr>
</tbody>
</table>

1 For detailed description of tasks please refer to the attached “Heuristic_evaluation” document.
2 For a detailed description of the rating scale please refer to the attached “Rating” document.
and what reasons? There should be more information provided regarding the captions of the different variables presented in figures.

In the individual patient graph for blood pressure readings, the Date format has the year first. However, the format dd/mm/yy would have been the most obvious option.

<table>
<thead>
<tr>
<th></th>
<th>Action 6.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>and what reasons? There should</td>
<td></td>
</tr>
<tr>
<td>be more information provided</td>
<td></td>
</tr>
<tr>
<td>regarding the captions of the</td>
<td></td>
</tr>
<tr>
<td>different variables presented</td>
<td></td>
</tr>
<tr>
<td>in figures.</td>
<td></td>
</tr>
<tr>
<td>In the individual patient</td>
<td></td>
</tr>
<tr>
<td>graph for blood pressure</td>
<td></td>
</tr>
<tr>
<td>readings, the Date format has</td>
<td></td>
</tr>
<tr>
<td>the year first. However, the</td>
<td></td>
</tr>
<tr>
<td>format dd/mm/yy would have</td>
<td></td>
</tr>
<tr>
<td>been the most obvious option.</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 9: Screening Questionnaire

A. Personal characteristics

<table>
<thead>
<tr>
<th>Participant ID [to be completed by researcher]</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Today’s date [to be completed by researcher]</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>MALE / FEMALE</td>
</tr>
<tr>
<td>Age</td>
<td>15-24 / 25-34 / 35-44 / 45-54 / 55-64 / 65+</td>
</tr>
<tr>
<td>Current job</td>
<td></td>
</tr>
<tr>
<td>Year started current job</td>
<td></td>
</tr>
<tr>
<td>Years of experience undertaking audit / QI</td>
<td></td>
</tr>
</tbody>
</table>

B. Familiarity with Computers and the Web

How frequently do you use computers and the World Wide Web (WWW) during the week (please circle)

[1 = Less than an hour per week; 2 = One to four hours per week; 3 = Five to 10 hours per week; 4 = More than 10 hours per week]:

<table>
<thead>
<tr>
<th>Desktop computers or laptops:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g. MS Word office, Internet, other client applications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>World Wide Web:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g. Search engines, Social networking sites, other websites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## C. Use of health information systems and applications

**On the days in which you work in your current job**, how often do you use the following types of software (please circle):

<table>
<thead>
<tr>
<th>Software Description</th>
<th>Never</th>
<th>Half the days</th>
<th>Every day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Electronic Health Records</strong> E.g. EMIS, Vision, SystmOne</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Audit software</strong> E.g. QOF reporting tools, ‘Population reporting’ function in EMIS, BMJ Informatica Audit+, PRIMIS GRASP-AF, IMPAKT CKD tool</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>‘Pop-ups’, ‘reminders’ and ‘alerts’ within Electronic Health Records E.g. Medication alerts, QOF pop-ups “Pop-ups” and alerts within Electronic Health Records, templates within Electronic Health Records</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Information retrieval applications</strong> E.g. Websites like ‘Map of Medicine’ or ‘CKS’, Information-buttons like ‘Web Mentor’ within EMIS.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Clinical risk calculators</strong> E.g. iPhone apps for Well’s scores or QRISK website</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Documentation templates</strong> E.g. Standardised templates within Electronic Health Records</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Decision support systems</strong> E.g. Diagnostic support tools like ‘Isabel’</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Appendix 10: Tasks
All participants were provided with the same contextual background information about a fictional primary care practice each time they were given a task. They were asked to use this information to inform their judgments during the tasks.

“You are a GP partner at Grove Medical Practice, located in a small industrial town in North West England. You have just over 10000 patients registered who are mainly elderly with high rates of multimorbidity. There are 6 full-time partners (including you) and one part-time salaried doctor. You employ 3 nurses, 1 healthcare assistant, 1 practice manager, 1 deputy practice manager, and 10 other administrative and reception staff. Historically you have been a high-achieving QOF practice, but recent changes to your contract mean your income has dropped significantly in the last 2 years. You have therefore agreed with your partners that only changes to the practice that are cost-neutral, cost-saving or significantly improve patient care should be implemented.”

Task 1:
You want to make improvements at a practice-level for patients with atrial fibrillation at your practice.
Please use the software find relevant suggested improvement actions, and indicate whether or not you agree with them given the information presented.

Task 2:
You want to improve the management of a patient with uncontrolled hypertension according to NICE targets.
Choose the patient you think it would be most important to address first according to your own judgment (you are free to use whatever criteria you wish – there is no right answer). Then find the suggested improvement actions for this patient and indicate whether or not you agree with them given the information in the software and the patient’s medical record (below).

Task 3:
You are about to see patient 123 in surgery. Please find whether there are any suggested improvement actions that could be used to improve their care, and if yes, indicate whether or not you agree with them given the information in the software and the patient’s medical record (below).
Task 4:
You have an idea to improve how your practice cares for patients with asthma. Your idea is to invite a local consultant to do a talk. Please add this to the software where you judge is most appropriate.

Task 5:
Please find the patient with the most improvement opportunities available at your practice across all conditions.

Task 6:
You want to implement all the improvement actions you have agreed with in the software that are awaiting implementation (both at practice-level and patient-level). Please download them so you can share them with your colleagues at your practice and ask for their help to implement them.

Task 7:
You have now implemented all of the improvement actions you downloaded in the previous activity. Please choose one (whichever you like, there is no right answer) and indicate in the software that it has been implemented.
Appendix 11: Post-test questionnaires

System Usability Scale (SUS) [1]

<table>
<thead>
<tr>
<th></th>
<th>Strongly disagree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think that I would like to use this system frequently</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I found the system unnecessarily complex</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I thought the system was easy to use</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I think that I would need the support of a technical person to be able to use this system</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I found the various functions in this system were well integrated</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I thought there was too much inconsistency in this system</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I would imagine that most people would learn to use this system very quickly</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I found the system very cumbersome to use</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I felt very confident using the system</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>I needed to learn a lot of things before I could get going with this system</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

References
### Action Interface questionnaire

<table>
<thead>
<tr>
<th>Actions</th>
<th>Difficult</th>
<th>Easy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Agreeing and disagreeing with practice-level actions for atrial fibrillation was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2. Agreeing and disagreeing with actions for a patient with uncontrolled hypertension was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3. Agreeing and disagreeing with actions for patient 123 was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4. Adding an action to invite a consultant to do a talk was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5. Finding the patient with the most improvement opportunities was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6. Downloading your plan was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7. Indicating an action had been implemented was…</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

### Object Interface questionnaire

<table>
<thead>
<tr>
<th></th>
<th>Unclear</th>
<th>Clear</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. The presentation of practice-level performance data was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9. The presentation of patient lists was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10. The presentation of patient-level data was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11. The presentation of action plans was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>12. The use of charts was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>13. The language used in the application was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>14. The use of colour in the application was…</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>15. The font used was…</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

---

³ A full description of each task is detailed in Table 1 of Chapter 6
Appendix 12: Interview schedule

Questions (in red) are organised according to constructs in Normalisation Process Theory (www.normalizationprocess.org/). Their tense was adapted depending on whether the interview was a baseline or second interview.

<table>
<thead>
<tr>
<th>Coherence</th>
<th>Cognitive participation</th>
<th>Collective action</th>
<th>Reflexive monitoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sense making work</td>
<td>Relationship work</td>
<td>Enacting work</td>
<td>Appraising work</td>
</tr>
</tbody>
</table>

Coherence is the sense-making work that people do individually and collectively when they are faced with the problem of operationalizing some set of practices.

Cognitive Participation is the relational work that people do to build and sustain a community of practice around a new technology or complex intervention.

Collective Action is the operational work that people do to enact a set of practices, whether these represent a new technology or complex healthcare intervention. Like all NPT constructs, it has four components.

Reflexive Monitoring is the appraisal work that people do to assess and understand the ways that a new set of practices affect them and others around them.

1.1 Differentiation: An important element of sense-making work is to understand how a set of practices and their objects are different from each other.

How do you currently use audit tools? How does PINGR fit with that?

What are the barriers/facilitators to using audit tools? How may PINGR help with that?

What do you understand as the purpose of PINGR? Can you describe what PINGR does?

How do you think you should use PINGR? As a) an individual user, and b) an organisation?

How have you used PINGR? As a) an individual user, and b) an organisation?

How is PINGR different from what you usually do? In terms of: a) interacting with audit data; b) developing action plans; c) changing clinical practice.

2.1 Initiation: When a set of practices is new or modified, a core problem is whether or not key participants are working to drive them forward.

How useful is PINGR?

What is the point of PINGR?

What have you done/not done (do/not do) with PINGR?

3.1 Interactional Workability: This refers to the interactional work that people do with each other, with artefacts, and with other elements of a set of practices, when they seek to operationalize them in everyday settings.

Can you show me what you do with PINGR?

In general, and in terms of a) use of the data, b) developing action plans, c) changing clinical practice

How has PINGR helped/hindered your everyday work?

What effect has it had on a) use of audit data, b) developing action plans, c) clinical practice?

4.1 Systematization: participants in any set of practices may seek to determine how effective and useful it is for them and for others, and this involves the work of collecting information in a variety of ways.

How do you know whether PINGR is or isn’t helpful? In general, and in terms of a) use of audit data, b) developing action plans, c) changing clinical practice?

How do you know whether PINGR is or isn’t helpful? In general, and in terms of a) use of audit data, b) developing action plans, c) changing clinical practice?
<table>
<thead>
<tr>
<th>Tools you use?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.2 Communal specification</strong>: Sense-making relies on people working together to build a shared understanding of the aims, objectives, and expected benefits of a set of practices. What do your colleagues understand about PINGR? How have your colleagues used PINGR? What benefits does PINGR bring and to whom? How valuable are these benefits? What are PINGR’s (relative) advantages/disadvantages? In general and in terms of: a) interacting with audit data; b) developing action plans; c) changing clinical practice.</td>
</tr>
</tbody>
</table>

| 2.2 Enrolment: Participants may need to organize or reorganize themselves and others in order to collectively contribute to the work involved in new practices. This is complex work that may involve rethinking individual and group relationships between people and things. Who is involved in: a) interacting with audit data; b) developing action plans; c) changing clinical practice. Who has been involved with using PINGR? How has PINGR changed how you work? a) individually and b) as a group/practice team? Who has been willing/able to invest time and energy into using PINGR? Who hasn’t? |

| 3.2 Relational Integration: This refers to the knowledge work that people do to build accountability and maintain confidence in a set of practices and in each other as they use them. How has PINGR affected your relationships with other colleagues in the practice; with patients; and with others (e.g., health care managers)? How has it affected your confidence in carrying out your role? |

| 4.2 Communal appraisal: Participants work together - sometimes in formal collaboratives, sometimes in informal groups to evaluate the worth of a set of practices. They may use many different means to do this drawing on a variety of experiential and systematized information. How have you discussed your experiences of PINGR with others the practice? What space is/has there been space to discuss PINGR in the practice? |

| **1.3 Individual specification**: Sense-making has an individual component too. Here participants in coherence work need to do things that will help them understand their specific tasks and responsibilities around a set of practices. What could/has help(ed)/hinder(ed) you use PINGR? How has PINGR helped you: a) interact with / understand audit data; b) developing action plans; c) changing clinical practice. |

| 2.3 Legitimation: An important component of relational work around participation is the work of ensuring that other participants believe it is right for them to be involved, and that they can make a valid contribution to it. Who has been/should be involved in a) interacting with audit data; b) developing action plans; c) changing clinical practice. Who has been/should be involved in using PINGR? |

| 3.3 Skill set Workability: This refers to the allocation work that underpins the division of labour that is built up around a set of practices as they are operationalized in the real world. Who gets to do the work is an important element of any set of practices. Who does what? In relation to a) interacting/responding to the data; b) developing action plans; c) changes to direct clinical work? In consultations and in the practice? Who has taken on the role of using PINGR? Who hasn’t been involved? How much training/learning did you/do you need to do before using PINGR? |

| 4.3 Individual appraisal: Participants in a new set of practices also work experientially as individuals to appraise its effects on them and the contexts in which they are set. From this work stem actions through which individuals express their personal relationships to new technologies or complex interventions. How have you reflected your/on any experiences of using PINGR? Are you clear on the effects of PINGR? On a) your clinical practice, b) your organisation, c) patient outcomes |
### 1.4 Internalization
Finally, sense-making involves people in work that is about understanding the value, benefits and importance of a set of practices. How has PINGR changed what you prioritise? In terms of a) responding to data, b) developing action plans, c) changing clinical practice

How does PINGR fit with the overall goals and activities of the practice? And CCG?

### 2.4 Activation
Once it is underway, participants need to collectively define the actions and procedures needed to sustain a practice and to stay involved.

What has/will help/hinder making PINGR become a routine part of practice/your activity? In this practice and beyond?

What has/will help/hinder making PINGR become a routine part of practice/your activity? In this practice and beyond?

### 3.4 Contextual Integration
This refers to the resource work - managing a set of practices through the allocation of different kinds of resources and the execution of protocols, policies and procedures.

How does using PINGR fit/not fit with other work? Done by a) you, and b) the practice? E.g. clinical work, quality improvement work, non-clinical work?

What effect has/will PINGR had/have on sharing work/resources/power/responsibility between staff at the practice?

### 4.4 Reconfiguration
Appraisal work by individuals or groups may lead to attempts to redefine procedures or modify practices - and even to change the shape of a new technology itself.

Based on your experiences, have you made any changes to your practice from using PINGR?

How do you think PINGR should be changed?

What do you think would be the barriers / facilitators to using pingr? How could it be changed to help you use it?