Re-engineering healthcare systems to use evidence from practice

Citation for published version (APA):

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
## Contents

Abstract 7

Declaration 9

Copyright 12

Eligibility of the Candidate 14

List of Publications 16

1 Introduction 18

2 Chronology and Organisation 23

3 Feasibility and Recruitment System for Improving Trial Efficiency: FARSITE 28
   3.1 Context 28
   3.2 Motivation 29
   3.3 Appraisal 30

4 Community Health eLabs: Bringing Together People, Data and Methods for the Benefit of the Population 33
   4.1 Context 34
4.2 Motivation ................................................. 35
   4.2.1 Maximising information utility for the community .... 35
   4.2.2 A consistent and coherent ethical, legal and governance
         framework ............................................ 37
   4.2.3 Removing barriers to reuse .......................... 39
   4.2.4 Federations of eLabs ............................... 40
4.3 Appraisal ................................................ 40
   4.3.1 Implementations .................................... 40
   4.3.2 Influence ......................................... 42
   4.3.3 Metrics ............................................ 43

5 Exchange and Reuse of Knowledge in Healthcare Systems 44
   5.1 Context .............................................. 45
   5.2 Motivation ........................................... 46
      5.2.1 Sharing Knowledge ............................... 47
      5.2.2 Replication of findings ......................... 47
      5.2.3 Reproducibility of Science ...................... 48
      5.2.4 A Health Data Science Commons ................. 48
   5.3 Appraisal ............................................ 51
      5.3.1 Implementation .................................. 51
      5.3.2 Influence ...................................... 53
      5.3.3 Metrics ....................................... 53

6 Closed Loop Healthcare Systems 54
   6.1 Context ............................................. 55
   6.2 Motivation .......................................... 56
      6.2.1 Policy Simulation ............................... 56
      6.2.2 Care Pathway Analysis ......................... 57
6.2.3 Closing the loop ............................................. 58
6.3 Appraisal ......................................................... 59
  6.3.1 Policy Simulation ......................................... 59
  6.3.2 Care Pathway Analysis .................................... 62
  6.3.3 Closing the loop ......................................... 63
  6.3.4 Metrics ...................................................... 64

7 Combining Health Data Uses to Ignite Health System Learning 65
  7.1 Context ......................................................... 65
  7.2 Motivation ...................................................... 66
  7.3 Appraisal ......................................................... 67
    7.3.1 Metrics ...................................................... 67

8 Discussion ......................................................... 68

9 Conclusion ......................................................... 73

Bibliography .......................................................... 74

Appendix A - Published Works .................................. 87

Appendix B - Health Informatics Design Patterns .............. 96
List of Tables

2.1 The research objectives addressed by each of the publications . . .  26
# List of Figures

2.1 The relationship between the publications submitted in support of the work. .............................................. 27

4.1 Extending ethical oversight and adding information governance by employing the community-centered eLab model for research or service-development with health records. ......................... 36

5.1 The commons functional architecture.............................................. 50

6.1 Closing the loop with COCPIT and IMPACT................................. 60

6.2 Conceptual care pathway analysis model.................................... 63

8.1 Learning Health System Cycle..................................................... 69
Abstract

RE-ENGINEERING HEALTHCARE SYSTEMS TO USE EVIDENCE FROM PRACTICE
John David Ainsworth
A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy (PhD), 2015

Health and care services need to be optimised to meet the future demand that will be placed on them. This will require a closer coupling of service and research, whereby innovations in services can be rapidly tested and evaluated, and feed back into a continual process of service optimisation. The timely delivery of information from services into research is critical to this cycle. However, there are serious problems with the evidence on which healthcare and public health practice is based: it is expensive to produce; it takes a long time to produce; it takes a long time to influence professional practice; it is crude, relating to the average participant and simple treatment definitions under ideal conditions. In other words, it gives a low-resolution picture of how a patient might respond to treatment or a how a sub-group of the community might respond to a public health intervention. This work is an exploration of informatics solutions to these problems with an aim to reengineer health care systems to make best use of the
evidence available.
Declaration

Candidate Name  John Ainsworth

Faculty of Medical and Human Sciences

Thesis Title: Re-engineering healthcare systems to use evidence from practice

The nature and extent of the candidate’s own contribution and the contribution of co-authors and other collaborators to each of the publications presented is as follows:

Publication 1. The initial FARSITE concept arose from a team discussion. The research and development to translate the concept to implementation was undertaken by the Ainsworth, the results of which are presented in this paper. Ainsworth also led the implementation team. This paper was written by the Ainsworth and Buchan reviewed the manuscript.

Publication 2. The concepts of eLab and Work Objects presented in this paper were original ideas and research of Ainsworth. This paper was written by Ainsworth and Buchan reviewed the manuscript.

Publication 3. This paper was the collective work of the eLab Technical Architecture Group, which was chaired by Ainsworth. Ainsworth’s
work on Work Objects and eLabs led him to establish the eLab Technical Architecture Group, to which he contributed the background research and provided intellectual leadership.

Publication 4. This paper was the collective work of a team, comprising of people skilled in many disciplines. Ainsworth designed the architecture of the IMPACT system, led the implementation team and managed the project. Ainsworth was the principal author of the paper, with contributions provided by the other authors.

Publication 5. This paper describes the application of the eLab and Research Objects concepts to a range of scenarios, based on the original ideas of the Ainsworth. Ainsworth designed the system architecture and led the implementation team. This paper was written by the candidate. The other authors reviewed the manuscript.

Publication 6. Ainsworth conceived of COCPIT, undertook the research and designed the system architecture. Ainsworth led the implementation team. This paper was written by the Ainsworth and Buchan reviewed the manuscript.

Publication 7. Ainsworth conceived of the string matching methodology, initiated and directed the research reported in this paper and is the senior author. This paper was written by Williams and the other authors reviewed the manuscript.

Publication 8. The ideas and research are a result of an intellectual debate and discussion between Ainsworth and Buchan over a number of years. Ainsworth was the principal author of the paper, with contributions provided by Buchan.

ii All of the work presented has been completed whilst the candidate has been
a member of staff of the University of Manchester.

iii No portion of the work referred to in this 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.

I confirm that this is a true statement and that, subject to any comments above, the submission is my own original work.

Signed:

Date:
Copyright

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, trade marks 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=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Eligibility of the Candidate

Degrees

1992 - 1993: MSc Cognitive Science, University of Manchester

1988 - 1991: BSc (Hons) Physics, University of Manchester.

Employment History

2010 - present: Senior Research Fellow, University of Manchester

2011 - present: Deputy Director of Centre for Health Informatics, University of Manchester

2012 - present: Deputy Director of Health e-Research Centre, University of Manchester

2008 - 2010: CTO North-West e-Health (NWeH), University of Manchester

2005 - 2008: Project Manager, NIBHI, University of Manchester

2004 - 2005: Research Associate, School of Computer Science, University of Manchester

2001 - 2004: Principal Consultant, PA Consulting
1998 - 2001: Senior Software Engineer, Cisco Systems

1994 - 1998: Software Engineer, Nortel Networks

Research Profile

Principal Investigator on four research grants with a combined value of £400,157 in the period upto 2015.

Co-Investigator on twenty research grants with a combined value in excess of £20m in the period upto 2015.

Co-supervisor for two PhD students.

Thirty research papers published in peer review journals upto 2015.

Nineteen refereed conference publications upto 2015.

Twenty-one invited talks upto 2015.

Two patent applications.
List of Publications


Chapter 1

Introduction

Healthcare provision is a part of all organised societies [1]. It may take on many forms, with differences that depend on a number of factors including: the scientific and technological sophistication of the society; the political organisation; the economic development; and the resources available. It is only relatively recently that five thousand years of tradition based medicine has given way to a norm of evidence based practice. This movement can be traced back to the 1948 streptomycin trial [2], the establishment of the MRC and the tireless campaigning of Archie Cochrane [3]. An evidence based system is essential for managing the quality of healthcare, measured in terms safety, effectiveness, efficiency, acceptability and equity [4]. As recently as the 1990s the concept of evidence based healthcare was still being explained to healthcare professionals [5]. Health systems around the world are still on a journey to more evidence based practice and many common treatments have no evidence base, they are the result of tradition and reasoning. Consider the example of surgery to take out a torn meniscus from the knee joint: the practice is based on flawed reasoning that the meniscal cartilage fragment acts like a loose ball bearing damaging the surrounding articular cartilage and putting the patient at risk of osteoarthritis. It was only when the
economic pressures of the NHS in the early 1990’s drove up waiting lists that a
natural experiment revealed the futility of menisectomy: a third of patients took
themselves off the waiting list in a year, and there are now randomised trials of
conservative vs. surgical management to confirm this finding [6].

Healthcare in the western world is organised to treat diseases, particularly in
their late stages toward the end of life. This leads to three negative consequences.
First, healthcare organisations treat disease-specific problems rather than whole
patients in the contexts of their lives that affect those diseases. Second, health
science produces research that is narrow and deep, encouraging the discovery
of more about less, which compounds the first problem and leaves major gaps
in medical knowledge such as the understanding of multi-morbidity. The sum
total of medical knowledge predicts less than a third of what will happen to the
average patient in response to the average treatment [7], [8]: healthcare systems
thus operate at low resolution. Third, the focus on late-stage disease treatment
is at the expenses of: disease prevention, health promotion and public health. To
quote William Farr [9]:

Diseases are more easily prevented than cured and the first step to
their prevention is the discovery of their exciting causes.

This remains as true today as it was in 1838, and is arguably of greater
relevance now than it was then. The progress toward evidence based practice
over the preceding seventy years is to be welcomed. However, to realise the
potential of evidence based healthcare requires a bigger ‘lab’ to increase the reach
and resolution of the evidence: in other words, we should think of and develop
healthcare systems as big cohort studies with nested trials.

The gold standard method for generating evidence is the randomised con-
trolled trial (RCT), ideally double-blind. Clinical RCTs enable us to identify
the most effective treatments, therapies and services. However, due to practical-
ities of execution, they are performed on narrow subsections of the population, 
for example excluding women of child-bearing age or patients with multiple dis-
eases. RCTs are also carried out in somewhat artificial settings, where a lot of 
observations are made and the attention to detail of care may be greater than 
usual, consequently RCT evidence can fail to generalise [10]. Epidemiological 
studies provide evidence on the evidence of the distribution and determinants 
of disease in populations. The natural histories of diseases are usually revealed 
by a network of insights across multiple epidemiological studies. However their 
power is constrained by the population size they cover and need careful design to 
identify and control for potential confounding factors. The canonical example of 
an epidemiological study is Framingham [11], a study that has provided most of 
what is known about the life course of cardiovascular disease. Birth cohorts pro-
vide the most powerful method for understanding the natural history of disease, 
but produce evidence slowly and are resource intensive. A limitation of these 
methodologies is that the data collected are determined at the start of the study 
and therefore unlikely to reveal something that wasn’t planned or anticipated.

There is another source of healthcare evidence, that which is produced by the 
process of actually delivering healthcare. This evidence is often referred to as real 
world evidence or routine healthcare data. It is generated in clinical encounters, 
pathology labs, and hospitals - it is the medical history of the patients accessing 
healthcare. Over the past two decades the transition from paper based records 
to electronic records has made the reuse of this data as a source of evidence at 
whole population level a tractable proposition. Although there are issues with the 
data relating to its quality, such a missingness and inconsistency in terminology, 
Electronic Health Records (EHR) provide a complete longitudinal data for each 
individual in the population.
Throughout society, in most aspects of our lives, over the course of the past twenty years there has been an information revolution. Ever increasing computing power and global network connectivity provided by the Internet has reached into nearly every aspect of our lives and changed them beyond recognition. From education to shopping to friendship to entertainment - our daily lives have been transformed. Industry and commerce has not been slow to adapt or adopt, to both the new opportunities afforded and the chance to reinvent existing business models. One of the most striking developments has been the rise of data analytics, that has been enabled by the digital footprints we leave as we traverse the new electronic landscape. Data analytics is characterised by (i) data collection as a by product of another activity; (ii) using machine learning techniques to discover hitherto unknown structure in the data that can then be used to model individual behaviours and make predictions based on those models.

There is one sector that has not benefited from the information revolution - healthcare. Healthcare does not routinely optimise itself by analysing the information it produces nor does it routinely use the data to look for hidden structure and discover new knowledge. The aim of the work presented in this thesis is to examine how we can re-engineer our healthcare systems to make the best use of evidence from practice to deliver care that is safer, more effective, more efficient, more acceptable and distributed with greater equity. It is postulated that this will require tighter coupling between services and research to establish a reinforcing feedback loop.

This work had five objectives:

i. Develop informatics systems to enable data flows between service and research.

iii. Develop informatics systems that enable users to collaborate.
iii. Develop informatics systems that enable knowledge to be exchanged and reused.

iv. Enable the reuse of health care data for the benefit of the communities that produce it.

v. Examine the informatics required for healthcare systems that can learn from their own data and human insights, collectively.
Chapter 2

Chronology and Organisation

The first paper [Publication 1] presented in the thesis was written and published in 2009 and describes the FARSITE system for clinical trial protocol feasibility analysis and recruitment. It is the starting point for the research undertaken into closing the loop between service and research. It focused on a narrow application area, but developed important ideas that were further developed and applied in the research that followed. The key finding from this paper is the demonstration of how informatics systems can be used to solve a problem at the intersection of service and research. The initial FARSITE concept arose from a team discussion. The research and development to translate the concept to implementation was undertaken by the Ainsworth, the results of which are presented in this paper. Ainsworth also led the implementation team.

The second paper [Publication 2] presented for this thesis was written and published in 2009 and introduces the two key ideas of eLab and Work Objects as a mechanism to close the loop between service and research and so enable truly evidence based healthcare. This paper presents a generalisation of the ideas developed in [Publication 1]. This paper was part of the UK eScience activity of the time, and it represents application of the eScience paradigm to healthcare.
The eLab is presented as a trustworthy environment for the safe use of EHR data and for collaboration amongst healthcare professionals, service managers, policy makers and researchers. Work Objects are developed as the currency of sharing of data and methods in the eLab and federations of eLabs. There are a number of themes in this paper that reoccur throughout the six other papers in the thesis, namely federation, local ownership and local control, public benefit of data, do once and share. The concepts of eLab and Work Objects presented in this paper were original ideas and research of Ainsworth.

The third paper [Publication 3] provides a more comprehensive treatment of Research Objects (science-specific Work Objects). An eLab technical architecture group was establish bringing together researchers from many domains (music, chemistry, bioinformatics) to specify the fundamentals of ROs. The concepts of permanence, graceful degradation, repeatability, and reuse from [Publication 2] are fully developed, and the paradigm is extended. This paper was written in 2010 and published in 2011. Ainsworth’s work on Work Objects and eLabs led him to establish the eLab Technical Architecture Group, to which he contributed the background research and provided intellectual leadership. This paper was the collective work of the eLab Technical Architecture Group, which was chaired by Ainsworth.

The fourth paper [Publication 4], written in 2010 and published in 2011, concerns the development of a generic policy simulation for commissioning decision support and its application to Coronary Heart Disease. The predictive modelling framework introduce in this paper enables the impact of service policy changes to be simulated prior to implementation. This paper was the collective work of a team, comprising of people skilled in many disciplines. Ainsworth designed the architecture of the IMPACT system, led the implementation team and managed the project.
The fifth paper [Publication 5] reports on the development and deployment of NHS eLabs and Research Objects in a healthcare setting for the purpose of reusing EHR. Written and published in 2012 this paper reports our findings when we have tried to apply these concepts in a real world healthcare setting. Four case studies are presented - primary care, bariatric surgery, public health and longterm conditions. This work is based on the original ideas of Ainsworth. Ainsworth designed the eLab system architecture and led the implementation team.

The sixth paper [Publication 6], written and published in 2012, presents a tool for analysing Integrated Care Pathways and the journeys that patients take through the healthcare system. The tool enables the variance between expected and observed care to be computed. This has a wide range of potential applications including data quality analysis, clinical audit, missed opportunity detection and care pathway redesign. In conjunction with the IMPACT system presented in [Publication 4] these two tools provide the foundation for adaptive healthcare systems that learn from the evidence that is produced. Ainsworth conceived of COCPIT, undertook the research and designed the system architecture. Ainsworth led the implementation team.

The seventh paper [Publication 7] written and published in 2014 presents the development and evaluation of novel string matching methodology for care pathway variance analysis and follows directly from [Publication 6]. Ainsworth conceived of the string matching methodology, then initiated and directed the research reported in this paper.

The eighth paper [Publication 8] was written in the period 2012 to 2105 and published in 2015. It distils the work of the preceding publications into a conceptual informatics framework for re-engineering health systems that are adaptive, the concept of a Learning Health System [12]. It explores why the information revolution in other industries has not translated into healthcare systems, and
introduces the concept of bidirectional evidence pipelines, that can be used for multiple purposes. We introduce the Data Action Latency (DAL) metric for measuring learning health system maturity. The ideas and research are a result of an intellectual debate and discussion between Ainsworth and Buchan over a number of years.

The chronological relationship between the papers is shown visually in Figure 2.1. Table 2.1 maps the research objectives to each of the publications.

<table>
<thead>
<tr>
<th>Publication</th>
<th>Objective Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Preserving consent-for-consent with feasibility-assessment and recruitment in clinical studies: FARSITE architecture</td>
<td>(i), (ii), (iii), (iv)</td>
</tr>
<tr>
<td>2. e-labs and work objects: towards digital health economies</td>
<td>(i), (ii), (iii), (iv), (v)</td>
</tr>
<tr>
<td>3. Why linked data is not enough for scientists</td>
<td>(iii)</td>
</tr>
<tr>
<td>4. IMPACT: a generic tool for modelling and simulating public health policy</td>
<td>(ii), (iii), (v)</td>
</tr>
<tr>
<td>5. elab: Bringing together people, data and methods to enhance knowledge discovery in healthcare settings</td>
<td>(i), (ii), (iii), (iv)</td>
</tr>
<tr>
<td>6. COCPIT: a tool for integrated care pathway variance analysis</td>
<td>(ii), (iii), (v)</td>
</tr>
<tr>
<td>7. Using string metrics to identify patient journeys through care pathways</td>
<td>(iii), (iv)</td>
</tr>
<tr>
<td>8. Combining Health Data Uses to Ignite Health System Learning</td>
<td>(iv), (v)</td>
</tr>
</tbody>
</table>

**Research Objectives**

(i) Develop informatics systems to enable data flows between service and research.

(iii) Develop informatics systems that enable users to collaborate.

(iii) Develop informatics systems that enable knowledge to be exchanged and reused.

(iv) Enable the reuse of health care data for the benefit of the communities that produce it.

(v) Examine the informatics required for healthcare systems that can learn from their own data and human insights, collectively.

Table 2.1: The research objectives addressed by each of the publications
This thesis is not organised entirely chronologically. It is organised to reflect the progression of ideas and discovery. Chapter 3 starts as the beginning with the first paper, FARSITE (Publication 1) from which three themes of research emerge. It then proceeds thematically, with a chapter for each of the themes. Chapter 4 describes Community Health eLabs (Publications 2, 5, 6). Chapter 5 is about the Exchange and Reuse of Knowledge in Health Care (Publications 2, 3, 5). Chapter 6 considers Closing the Loop between Service and Research (Publications 4, 6, 7). Chapter 7 presents the final paper, which takes the findings from the research themes to answer the question about how we can re-engineer healthcare system to adapt and learn. The closing chapter in thesis summarises and discusses the findings and concludes the thesis.
Chapter 3

Feasibility and Recruitment System for Improving Trial Efficiency: FARSITE


This is Publication 1 in the List of Publications.

3.1 Context

The background to this research was the finding that 30% of UK clinical trials fail to recruit the expected number of participants and of those that succeeded only 45% manage to recruit in the planned time-frame [13]. This adds extra cost and delay to the introduction of new medicines where clinical trials are needed. At
the same time clinical trials were becoming an internationally competitive business, with the pharmaceutical industry looking beyond the traditional developed western countries. The withdrawal from research activity in the UK by Pfizer [14] was symptomatic of this trend. The NIHR Clinical Research Networks, which have the remit to service industry, were under pressure to deliver a solution to the problem. Ad-hoc estimates based upon best guess knowledge of clinical staff were no longer acceptable. There was an imperative to improve clinical trial feasibility and recruitment.

3.2 Motivation

The motivation for this work was to develop a system that would be used to assess clinical trial protocol feasibility and to aid clinical trial recruitment whilst maintaining patient level privacy and confidentiality. A long standing problem for clinical studies was that of consent-for-consent. Simply stated, a patient’s consent is required to view their medical history to determine if they would be eligible to participate in a trial, regardless of the eventual outcome, where those records disclose the individuals identity. So in effect, in order to be able to recruit from a whole population group, that population would need to have given their consent. This is impractical, costly to do and unlikely to yield a response from enough of the population to deliver any benefit. This is a very good example of where the disconnect between service and research causes problems. There is no such restriction on the patient’s clinician assessing their suitability for a clinical trial, in fact they have a duty to provide the best treatment (including experimental treatments where available) for their patients. However, only a small fraction of clinicians are research active, and so the ability to scale to whole population through clinicians is not there. The FARSITE system was conceived of to provide
an informatics bridge between service and research in support of clinical studies for the development of new therapies.

### 3.3 Appraisal

The FARSITE tool was developed as an application to improve the feasibility analysis of studies, supporting their planning and subsequent recruitment of participants. It provides an informatics solution to the problem of consent-for-consent. The key innovation/insight in this work was (a) the use of two databases - one containing identifiable data (the ‘service’ database), the other containing the same data but anonymised (the ‘research’ database) (b) the exchange database queries across the service-research boundary (c) the re-writing of queries to return patient identity when the authority exists to do so. Patent applications for this invention were filed [15] in the United States of America. There are a number of shortcomings with FARSITE or areas where the tool could be improved. Firstly, FARSITE system lacks a formal model of a clinical trial protocol, for example CDISC ODM [16] and as such is unable to share or interoperate with other systems at a level beyond individual disease definitions, although these may be very complex statements. There is no support for free text (non-coded data) and it has never been used beyond primary care data. It does not have a federated query architecture, though given it has a service oriented architecture it should be relatively easy to extend in this way. The FARSITE architecture has proven to be well founded. It has been influential in the development of both EHR4CR [17] and CPRD [18]. There are also a number of technical design patterns first implemented in FARSITE that were used successfully in platforms that were to follow such as STELAR [19], IMPACT [20], COCPIT [21], eLab [22]. These are catalogued in Appendix B. FARSITE is not the only clinical trial protocol
feasibility tool to have been developed, for example EPCRN [23], STRIDE [24], DebugIT [25], i2b2/SHRINE [26]. However, it is unique in that FARSITE is available at the time of writing as a commercial product of the University of Manchester and Salford NHS spin-out North West eHealth (www.nweh.org.uk).

It has been deployed widely across the NIHR Comprehensive Research Networks in England, and also by the Farr Institute across the North of England, such that it can now reach a population of more than 1 million. FARSITE was influential in the development of the Salford Lung Study [27]. The FARSITE solution to the consent-for-consent problem has also proved to useful in areas beyond clinical trials, finding new applications in risk stratification, case finding and health improvement.

Perhaps even more importantly, especially in the context of the work presented in this thesis, FARSITE contained three ideas that were to become major research themes. The first of these was the notion of web-based platform for collaboration around Electronic Health Record data. A generalisation of this idea led to the research into Community Health eLabs (Chapter 4). The second idea was the uses of query objects as packages of knowledge about clinical trial eligibility. A generalisation of this idea led to the development of Research Objects (Chapter 5). The third and final idea was the establishment of a data flow and an interface for interaction between service and research, termed “crossing the clinical-research boundary” in the original paper. A generalisation of this idea led to our work on Closing the Loop (Chapter 6).

**Metrics**

Publication 1 was published in the series Studies in Health Technology and Informatics which is prominent in the field of Health Informatics. It has been cited nine times. It was also awarded the best paper prize at HealthGrid 2009 in Berlin. The key indicators of the value of this work is to be found in the number
of systems and projects that it has influenced.
Chapter 4

Community Health eLabs: Bringing Together People, Data and Methods for the Benefit of the Population

This theme draws on three publications:


This is Publication 2 in the List of Publications.


This is Publication 5 in the List of Publications.

*This is Publication 6 in the List of Publications.*

### 4.1 Context

Consider the question: does the interaction of tamoxifen with antidepressants lead to recurrent breast cancer in a substantial number of women? The interaction is biologically plausible [28], but its pharmaco-epidemiology is not well characterised. So the question might be asked for audit or research purposes and there are a number of ways to approach this. The investigator might ask a regional cancer registry but find: incomplete data on recurrent cancers; little data on adjuvant therapies such as tamoxifen; and no data on co-prescriptions such as antidepressants. The latter might be prescribed for a number of reasons, such as to counter the hot-fush side effect of tamoxifen, or to treat depression. Alternatively, a more complete, but less quality-assured, source of data could be used. In England this could be one of the anonymised extracts of primary care information systems. These include, CPRD (www.cprd.com) from practices which are paid to participate and collect some additional data, and Q-Research (www.qresearch.org), from practices which are not paid to participate but have agreed to share routine care data for research. An investigator wishing to address the tamoxifen-antidepressants interaction question via CPRD or Q-Research would use the ethics and research governance procedures organised around those databases.
4.2 Motivation

Existing approaches to “secondary use” healthcare data repositories tend to create specialised subsets that are fragmented, or centralised data warehouses that are detached from the community they describe. The fragmentation and detachment of health records from the community that creates and owns them is counter-productive and results in ethical, governance and data quality problems. The alternative is a community-driven model that enables population-based uses of health records while preserving individual privacy, that can be federated together by local communities in terms of ethics, governance and technology.

4.2.1 Maximising information utility for the community

Secondary use [29] clinical data repositories are often fragmented and/or detached. Fragmented repositories contain a subset of care records, for example a disease specific register, such as a cancer registry. Fragmented repositories may have been created by extracting information from care records and/or by explicit data entry. The data contained in the repository is a partial view of the patient’s care and so its utility is limited. Detached repositories are created from electronic health records, but they are managed and governed by an agency not directly involved in the care of the patient/citizen. The removal of the data from the community creates a number of problems. The data is now under a different ethical and governance framework, therefore its uses are beyond the day-to-day control or sight of the contributing community. The meta-data, typically contained within people, is lost: for example the knowledge of the clinical biochemist that his laboratory changed the assay for creatinine at a certain time point, after which a different formula for estimating kidney function needs to be used [30].
Once a detached data repository is created, there is no way for a patient to withdraw their consent and have their data removed. Finally, there is no feedback loop between the analysis of the data and the provision of health services in the communities from which the data is derived.

We proposed the development of Community Health eLab model as the solution. The work is presented in Publications 2, 5 and 6. The eLab model is community-centered rather than database-centered, as shown in Figure 4.1.

Under the eLab model, communities may integrate records from primary care, secondary care, specialist care sources such as cancer registries, and administrative sources such deaths and demographic data: this would be natural for the formation of an integrated health record system for that health community. The eLab database would be an extract of this integrated health record, with personal identifiers removed. Most importantly, the record linkage takes place...
within that community for the purposes of informing individual care and running local care-services. Keeping within the secure network of the healthcare agencies and contractual obligations with information governance, the anonymised linked database would then be searched. In the tamoxifen example the anonymised database would be searched for records of women diagnosed with a first occurrence of breast cancer who have also taken tamoxifen and certain anti-depressants within a given period. Their records are then also checked for recurrence of the breast cancer. If the number of full-term pregnancies were also sought for each woman, there would be a remote risk of a woman with an exceptionally large multiple-pregnancy being identified deductively as having possibly been given antidepressants. In the eLab model this disclosure takes place within clinical information governance; therefore, it can be stopped from leaving the secure healthcare environment of that community. Under a research database scenario, in contrast, the disclosure would take place in a research environment. Therefore, we conclude that the eLab model would enhance the ethical oversight and governance of this kind of observational research using routine healthcare records.

4.2.2 A consistent and coherent ethical, legal and governance framework

An eLab brings together data, methods and expertise to maximise the benefit to a community from the uses of the anonymised records of its members. This concept maps conveniently to local health economies, which oversee local healthcare services for geographical populations of around 300,000. Given a natural community context, research ethics committees may find it easier than in a more loosely defined context to assess potential eLab uses with respect to the principles of autonomy, beneficence, non-maleficence and justice [31]. If several communities
wish to share expertise to make best use of eLabs, they can work in a federation while each maintains the ownership and governance of its local eLab. Ethical oversight for eLab uses might run efficiently and effectively at the level of the federation, especially where each member is in the same family of governance, e.g. a country. Furthermore, the federation principle could be taken up to a global level via nested structures of trust, benefit and governance, to improve the interoperability of global health intelligence. The eLab model was designed to make the uses of personal data more secure, transparent and auditable, and thereby more governable. The autonomy of individual citizens is preserved through their right to opt out of the eLab; the public health services may be well placed to ensure that balanced information about the benefits and risks of participation are discussed with the local population. The public health services may also be well placed to ensure that the eLab is used for maximum public health benefit, for example through equity audit of healthcare services influencing commissioning, or through healthcare safety surveillance and research improving the quality of care, thereby promoting beneficence and justice. Non-maleficence is ensured by keeping the information management of the eLab within the agency responsible for safeguarding the most sensitive personal information, for example the English National Health Service (NHS) and its clinical information governance. CCGs are well placed to act as the custodian for integrated health records, and in turn, eLabs of anonymised health records, upholding the law, and the highest standards of ethics and governance, while seeking maximum health-gain for their community.
4.2.3 Removing barriers to reuse

The motivation for the eLab is to ensure direct benefit to the community from which the data was generated. However, there are other barriers that have prevented the efficient and effective reuse of health records by analysts and researchers serving the needs of that community. Firstly, the data need to be made available, but for the data custodians the theoretical risk of identity disclosure, leads to over caution and reluctance to make data available. Where data has been made available, a second problem emerges. Researchers who wish to use the data lack the knowledge of the policies in place for data access and can face inconsistent interpretation and application of the resulting requirements. This again has recently been highlighted in the second Caldicott review [32]. A third problem is that procedures required by data custodians to ensure effective information governance through the whole life-cycle of research (from data acquisition to archiving) are often opaque, and incomplete. Consequently the researchers lack the procedural knowledge to efficiently and effectively use the data.

The eLab provides a secure environment for managing, exploring and analysing anonymised data from health records, accessible through a web browser. The eLab environment is secured through both software and operational governance procedures. The eLab addresses these three barriers to reuse of evidence from practice. By establishing each eLab as a NRES research databank, a local governance board enables consistent application of policy and considers ethics in the context of the community it serves. The eLab then supports this by binding the information governance to the data, by explicitly embedding it in the researchers workflow. Within the eLab environment the researcher can identify the variables required for their research questions, submit an application to the governance board, receive the decision and finally access the data in one place. For the
data custodians, a complete audit trail is automatically generated. The eLab model provides an integrated environment for health record reuse that benefits the researchers, the data custodians, and the population that create the data.

4.2.4 Federations of eLabs

The motivation behind the eLab is to provide local ownership and local control of data for the community. A further extension to the idea of eLabs is to create a federation of eLabs. That is a network of eLabs that agree to collaborate in the interest of the communities they serve. A Federated eLab, provides users with access to data integrated from multiple Health Community eLabs. The Federated eLab performs distributed queries across the data repositories of all Health Community eLabs within the eLab Federation and performs real-time data integration. This distributed approach eliminates the need for a large central data warehouse and allows the members of the Federation to retain complete control of their data and the data that is exposed to the federation. For example, the management of the withdrawal of patient consent is done easily within the community. For research applications a federation of eLabs can be used for rapid replication of findings in heterogeneous populations.

4.3 Appraisal

4.3.1 Implementations

The three publications that describe the eLab reflect three iterations of the concept and their corresponding implementation. The original concept for the eLab, as described in Publication 2, was first implemented at the University of Manchester and deployed in the Salford health economy with the Salford Integrated
Record as the data source. In this iteration, a very generic and flexible platform was developed providing the eLab capabilities as originally intended. It was not successful, as it was rarely used by the NHS analysts or the research community as planned. The reasons for failure were due in a large part to the expectation that if the eLab was built then the users would come to it. With hindsight this was naive, especially given that we were deploying into the NHS, knowing the slow pace of adoption and innovation in the NHS. The eLab would have been a giant leap into the unknown, when actually what was required was a series of incremental improvements from the starting point. The fact that it was so generic and flexible meant the learning curve for NHS analysts was too steep to climb. The benefits of investing time and effort in learning the eLab were not communicated effectively. This indicates a second reason for failure - adoption will only occur if it is actively made to happen. Technology alone cannot effect cultural change. These findings informed the second and third iterations. In second generation eLab, as described in Publication 5, application specific eLabs were created. Building on the first generation platform, bespoke application eLabs were created for defined communities. As these eLabs met a specific need previously articulated by the end users they were, not surprisingly perhaps, more successful. However, they were inflexible and were not general purpose tools, the pendulum had swung too far in the opposite direction. In the third generation of eLabs we attempted to redress the balance with the COCPIT, as described in Publication 6. COCPIT provides a collaboration environment, a unique analysis capability for care pathway variance, and can be easily adapted to most longitudinal healthcare datasets. The iteration of the eLab seems to have found the right balance between generality and specificity. The COCPIT tool is used extensively for Missed Opportunities Modelling [33] and at the time of writing is available as a commercial product of the University of Manchester and Salford NHS spinout
venture North West eHealth (www.nweh.org.uk).

4.3.2 Influence

The ideas developed in the eLab series of papers can also be found in other independently developed solutions such as i2b2 [34], SAIL [35] or SHIP [36]. An objective of the Transform project was to develop an extensible privacy and confidentiality framework for secondary uses of health record data. As part of this work the “Zone Model” was developed [37], that defined three zones, the care zone the non-care zone and the research zone. The first two zones are data source zones, the third is a data sink zone. Broadly speaking the care zone has identifiable data, whilst the non-care zone and the research zone have either pseudonymous or anonymous data The model provides a way to examine the risk gradient as data flows between zones. Our Community Health eLab approach defines a new way of organising these data flows, and we have argued for eliminating the distinction between zones, where data does not flow between zones as it stays under local ownership and control of the population that owns it. A fourth generation of eLab is now being developed for the research consortia. This generation of eLab is focused on specialist research datasets such as the asthma birth cohorts managed by STELAR [19] or the food allergy datasets of the iFAAM consortium. These eLabs focus on collaboration and data management and do not provide analysis capabilities as any effort to do so would be unlikely to displace the highly functional tools the community already uses. The eLab model can also be seen in the European Medical Information Framework (EMIF). The ideas based upon federation were developed into a proof of concept [38]. The failure to deploy NHS eLabs at scale meant the ideas on federation remain at the proof of concept stage. However, the national eInfrastructure planned by the Farr Institute will provide
a future testbed for the ideas of federation, particularly the rapid replication of
findings across multiple communities. The Greater Manchester Academic Health
Science Network are also embarking on the DataWell project to federate primary
and secondary care data sources across the NHS in Greater Manchester. The
second Caldicott review of information governance published in 2013 calls for the
NHS and adult social care services to pledge to “anonymise the data collected
during the course of your care and treatment and use it to support research and
improve care for others and to inform you of research studies in which you may
be eligible to participate”’ [32]. Dame Caldicott also called for the establishment
of safe havens for making use of healthcare data. The eLab series of papers are
amongst those that have influenced this direction of travel. However, the Health
and Social Care Act of 2012 [39] gave the Health and Social Care Information
Centre (HSCIC) statutory powers to collect and link healthcare data. HSCIC
immediately initiated the Care.Data programme [40] with the aim of creating a
single data warehouse linking primary and secondary care data for all citizens
in England. Eighteen months later Care.Data’s future was in the balance as the
programme had been seriously damaged by the loss of public trust [41]. The
impact of the devolved health budget to Greater Manchester [42] may further
escalate the tension between centralisation and localisation.

4.3.3 Metrics

Publication 2 was published in the series Lecture Notes in Computer Science
which is widely regarded in Computer Science. It has been cited eleven times.
Publication 5 and Publication 6 were published in the series Studies in Health
Technology and Informatics which is prominent in the field of Health Informatics.
The former has not been cited, the later has been cited six times.
Chapter 5

Exchange and Reuse of Knowledge in Healthcare Systems

This theme draws on three publications:


*This is Publication 2 in the List of Publications.*


*This is Publication 3 in the List of Publications.*
5.1 Context

The availability of data about health and care is ever increasing. A relative "data deluge" sits alongside a shortage of health data analysts. At the same time, models that try to make sense of the data continue to be produced by academia, providing a bewildering choice for those analysing health systems in the wild. The investment in people has failed to keep pace with the corresponding investments in data and models. The NHS is suffering from a drought of skilled analysts that can make sense of the deluge of data and models. In recognition of this fact, in 2011 the MRC led a consortium of funders in an initiative to establish eHealth Informatics Research Centres (eHIIRCs) with a remit to train the next generation of health informatics workforce, which became a key activity of the Farr Institute [43]. A 2011 report by McKinsey [44] examine the needs for data scientists across all industry sectors and predicted a shortfall of 140,000 skilled people in the US alone. Given the under capacity in the workforce, the question then becomes how can we ensure that we don’t duplicate effort. Can we use informatics to underpin the “do once and share” ethos?

Consider how scientific findings are reported and disseminated. A paper is published which describes the hypothesis, the methods and the findings, but lacks the detail that would be required for another researcher to actually reproduce the findings. There is a ‘missing middle’ of assumptions and decisions that
are made and yet rarely reported. This leads to a lack of transparency in science. Data science is an end-to-end digital activity. Every action, decision or assumption made by a data scientist can be captured electronically. The knowledge captured in these digital assets may be the result of intensive human activity. However, because they are digital, given the appropriate infrastructure, the costs for dissemination and reuse are negligible in comparison. Data science can be transparent and therefore it is imperative for its practitioners to work towards full transparency, reporting the ‘missing middle’.

5.2 Motivation

The ability to package and publish knowledge as a digital asset in a such a way that can it subsequently be discovered and exchanged, was the organising principle behind the work in this theme. The capability to do this would have far reaching implications not just within healthcare but to any area where digital knowledge assets are produced and their exchange is of value. In the first paper (Publication 2) we used the nomenclature of Work Objects to reflect the role we envisaged for them in health systems analysis. In the following papers (Publication 3 and 5) we generalised the concept and renamed them as Research Objects to reflect their resonance with allied work in the digital publication field. Within the Community Health eLab Research Objects are the currency of collaboration, and within a federation of eLabs they are the currency of knowledge exchange.

Research Objects are semantically rich aggregations of (potentially distributed) resources that provide a layer of structure on top of information. A Research Object provides a container for a principled aggregation of resources, produced and consumed by common services and shareable within and across organisational boundaries. A Research Object bundles together essential information relating
to experiments and investigations. This includes not only the data used, and methods employed to produce and analyse that data, but also the people involved in the investigation.

5.2.1 Sharing Knowledge

Research Objects can be used to overcome the shortfall of analytical capacity in the NHS. Research Objects created within the eLab, can be used to capture and package common analytical tasks e.g. the prevalence of diabetes across a region or Accident and Emergency waiting times. Through a federation of eLabs these Research Objects can be distributed and reused to other localities. This enables a community of analysts to be formed across the NHS, borrowing strength from each other. Research Objects can also be subjected to quality control either through formal validation or via crowd sourced reviews or recommendations. The use of Research Objects in this way also provides two other benefits. The first is that Research Objects become educational tools. An analyst can inspect the knowledge content embedded in the Research Objects and learn from it. The second is that Research Objects embedded within an eLab provide an organisational memory. They are persistent, self-contained, coherent and complete, and so even if the original author leaves the organisation, a usable record of their work remains.

5.2.2 Replication of findings

Research Objects can be thought of a packages of evidence generation that can be transferred from one population to another. Within health data science we have the opportunity to test findings from one population to another population by reproducing the analysis with data from the new locale. Research Objects enable
the entire analysis to be captured such that it can be transferred to another new population dataset within a community eLab, semantically mapped to the new population dataset and rerun. This would enable data scientists to harness the heterogeneity of populations in their research and provide a mechanism for rapid replication of emerging signals.

5.2.3 Reproducibility of Science

Research Objects provide the mechanism to capture the hitherto hidden assumptions researchers make. Consider for example, the process of extracting quantified drug exposure in prescriptions for oral steroids. The researcher has a series of decisions to make about how to treat ambiguous situations such as overlapping prescriptions, or missing end dates. The different possible permutations of these decisions will each produce a different dataset, which in turn will potentially produce different results. Therefore, for reproducibility and transparency of science, it is vitally important that they are captured and recorded for publication. Without publication, there is not the possibility for the assumptions to be challenged. For the first time it exposes the ‘missing middle’ of data science for review and critique by their peers. This is true scientific transparency.

5.2.4 A Health Data Science Commons

Research Objects lead naturally into the concept of a health data science commons (‘the Commons’), an online shared space for the aggregation and exchange of the digital assets created by the researchers and data scientists. It would act as a bridge for the exchange of knowledge between the NHS and academia. A health data science commons is a virtual space into which digital assets of scientific research can be published, indexed, discovered and consumed. A health
data science commons can provide a step change in data science for research by enabling the exchange and reuse of digital assets and this would fundamentally change the way scientists perform their work. It requires cultural change within the community. Currently, there is no mechanism or incentive for reusable Research Objects nor the capability to reproduce findings by the exchange of digital objects. For most if not all scientists, the thought of exposing their data and working practices is a daunting prospect. However the goal of the commons is to make sure that the ultimate beneficiaries of open science will be the scientists themselves who will be able to scale and accelerate discovery through the use of the Commons. This will lead to less duplication of effort and so enable funders to leverage and multiply their past investments in future funded activities. The Commons will provide a persistent unique identifier for each digital asset. The identifier will enable digital assets to be found, shared and attributed enabling them to be cited by other scholarly works. Each digital asset will have associated (i) provenance, which minimally, identifies the creator(s) of the asset, those that have subsequently modified it, and how it was modified, and (ii) descriptive metadata, which is required for each asset to facilitate indexing, discovery and reuse. These three basic rules are sufficient to create a functioning data science commons that will be of use to Data Scientists.

We define the rules of the commons as follows:

1. Each unique digital asset placed into the Commons must have a unique identifier.

2. That unique identifier must allow the digital asset to be found, shared and attributed.

3. Attribution requires associated provenance that, minimally, identifies the
creator(s) of the unique research object, those that have subsequently modified it, and how it was modified.

The core functional components of a commons are shown in the Figure 5.1. Every component in the figure can be independently developed and instantiated by any organisation that wishes to do so. This is not a monolithic service or software stack. The commons is defined by its rules and the standard interfaces between the components in the figure. Digital assets in the commons will use the Research Objects framework (www.researchobject.org) which specifies existing standards for archiving, provenance and annotation. Digital Object Identifiers (www.doi.org) will be used to identify Research Objects.

The Commons will achieve critical mass when it becomes a necessary and routine part of the data scientist’s activities. Until that point is reached, the community around the Commons would need to be created and actively promoted. This must include mining existing sites for content and raising awareness
of the existence and purpose of the commons and the nature of the reward environment.

The ability to exchange these valuable pieces of knowledge within a commons will accelerate science by making the associated assets more readily available for reuse. For example, by providing a mechanism that enables clinical codes to be reviewed and rated by a community of experts, this will lead to convergence on an accepted definition when there are competing alternatives, leading to standardisation and interoperability.

The commons is a further generalisation of the ideas of federation which were articulated in the publications. The Commons differs from a federation in that no agreement needs to be enacted to be part of it. It is much more similar to the Internet model in that participation is open, requiring just conformance with standards and conventions, and the endpoint has complete control over access and the publication of content.

5.3 Appraisal

5.3.1 Implementation

Publications 2 and Publication 5 described Research Object implementations, that were part of the corresponding eLabs. Each analysis that is run within the eLab is saved in a Research Object (RO). Every RO contains the research or audit protocol, the data queries and the anonymised data extracts. Where necessary, they may also contain information such as the statistical analysis scripts, annotated statistical result/log files, manuscripts or slides. ROs are managed within the security of the NHS network. The success of these Research Object implementations were fate-shared with the success of the eLabs themselves, which is
discussed in the previous chapter. To reiterate the point the principal barrier to adoption and use were not technical but cultural. The failure to establish a federation of eLabs meant that the opportunity to test the exchange and reuse of knowledge via Research Objects did not materialise. However, this did lead to the development of the ideas behind the Commons, that does not require the formal agreements and associations of a federation, and would be a first step towards a federation sharing Research Objects. Publication 3 which sets out a generalisation of Research Objects led to additional work on the formalisation of Research Objects. The work of the eLab Technical Architecture Group is now continuing with HeRC as part of the CHIP-SET theme. New concepts such as check-lists for minimum information models have been added [45]. There is also work under way to develop Research Objects that carry with them the information required to enable semantic mapping of data sources to methods embedded in the Research Object. There are now implementations of Research Objects in the STELAR eLab [19], clinicalcode.org [46] and EHR4CR [17]. The Farr Institute is developing a data science commons which is accessible at www.farrcommons.org. The goal is to develop further the ideas of rapid replication using Research Objects between the Farr Institute centres.

There are still a number of areas where further research is needed. The first is Identity and versioning of research objects. The second concerns how to quality control Research Objects, possibly by crowd sourcing, and then how this would affect the ranking of Research Object search results. The third is to develop incentives for scientist, tool builders and scientific publishers to use Research Objects.
5.3.2 Influence

The ideas developed in Publication 3 have had widespread influence beyond health informatics reaching a broad range of disciplines. Other systems of knowledge exchange and reuse that have had a parallel development and influence are Scientific Knowledge Objects [47], and nano-publications [48]. Within the bioinformatics field myExperiment [49], and Workflow4ever [50] implement Research Objects. In systems biology there is Sysmo [51]; in music archivation [52]. The Big Data 2 Knowledge (BD2K) [53] programme in the US includes elements of Research Objects, especially the BioCaddie work which is directly influenced. It is also developing a data science commons. The emergence of electronic phenotyping, of which clinicalcodes.org [46] is an example, has at its core the notion that electronic phenotypes can be exchanged and reused (see for example phekb.org [54]). A full list of initiatives is maintained at www.researchobject.org/initiative.

5.3.3 Metrics

Publication 2 was published in the series Lecture Notes in Computer Science which is widely regarded in Computer Science. It has been cited eleven times. Publication 3 was published in the journal Future Generation Computer Systems. It has been cited by 154 other publications. Publication 5 was published in the series Studies in Health Technology and Informatics which is prominent in the field of Health Informatics. It has no citations to date.
Chapter 6

Closed Loop Healthcare Systems

This theme draws on three publications:


*This is Publication 4 in the List of Publications.*


*This is Publication 6 in the List of Publications.*


*This is Publication 7 in the List of Publications.*
6.1 Context

Long-term conditions, particularly vascular diseases, consume the largest proportion of healthcare budgets and are a major target for public health initiatives. Moving interventions up-stream to earlier stages of the natural histories of diseases would delay or prevent subsequent events, thereby reducing the amount of suffering over the average lifetime, and saving money [55]. Health policy-makers and those planning and managing local health services are poorly served by oversimplified estimates of the potential public health impacts of making changes to the pathways of care or taking preventive public health measures [56]. These estimates are often unreliable because the models do not adequately represent the complexity of the disease, population or care over time.

Population health impact estimation is usually done by a small group of analysts synthesising evidence and producing a report for a decision-making team [56]. For example, how should the balance be struck between investments in statins vs. smoking cessation vs. physical activity promotion in respect of health impact for a defined population over five years? There are three problems with this approach: (i) there are not enough analysts to support current decision-making needs, yet the available data and literature to consider is increasing - it is also unlikely that health systems could afford to employ more analysts, and furthermore they are in short supply; (ii) a static report does not enable ‘what if’ scenario planning, so the options that are appraised are inflexible; (iii) most healthcare commissioning groups do not have the skills or time to build realistically complex models which take all reasonable factors into consideration, so decisions may be biased by where a narrowly defined model focuses - this may reflect the interests of service providers more than the needs of the population served.
Policy changes are reflected in service provision, but services are rarely titrated to the needs of local populations because sufficiently timely and detailed information about those needs is not available [56]. Relevant data is, however, available. It is valid to ask the question what is happening in my locality? Does it differ significantly from the national guidance? If so what are the causes of the variance? Should this entail a local customisation of the pathway? Is this customisation actually relevant at the national level?

6.2 Motivation

6.2.1 Policy Simulation

Populations are under-served by local health policies and management of resources, partly because of a lack of realistically-complex models for appraising a sufficiently wide range of intervention-options. A similar lack of usefully-complex models impedes clinical audit. Rising computing power coupled with advances in machine learning and healthcare information enables relevant models to be constructed and executed. However, such models are generally not accessible to health professionals because the workforce does not have the requisite technical knowledge or skills. This was the motivation behind the work of Publication 4 and the creation of an accessible system for executing and analysing the results of simulated public health and healthcare policy interventions. The system was designed to be accessible and usable by modellers and policy-makers alike. By employing a unified modelling strategy, based on probabilistic graphical models, we expected that the same information system will be used to explore the population health impacts of potential changes to care pathways.

The motivation was to build an informatics system that would bring together
an interdisciplinary team to organise and analyse typical data, and to develop tools that can be used by non-specialists to run interactive, ‘what if’, scenario planning activities, which are both methodologically robust and locally relevant. This would draw commissioners and providers of services together over new views of their local NHS data, in a way that is amenable to patient and public involvement.

It is theoretically possible to construct probabilistic graphical models of disease and healthcare pathways, and to use the resulting probabilistic networks to simulate outcomes for populations. Such a simulation system would enable the user to compare different intervention scenarios, with the ability to modify both clinical and public health interventions, and measure the effectiveness based on both clinical outcomes and costs. Such a system would bring together public health professionals, clinicians and service commissioners in interactive scenario planning activities to inform policy decisions. The ideal system would enable users to construct and share models around ‘what if’ scenarios easily; to execute individual simulations quickly; and to interpret simulation results collectively. Large-scale simulations like this would provide greater accuracy but consume more computational resources. The construction of the underlying models requires collaboration between health economists, epidemiologists, biostatisticians, informaticians and typical decision-makers/leaders (public health professionals, healthcare managers, and clinicians).

6.2.2 Care Pathway Analysis

The need for usefully-complex policy models described above applies equally to care pathways. Instead of the general population, the population of interest might be restricted to the patients under the care of the service being modelled. The
statistical and computational challenges, however, are very similar to the policy modelling. Some models cross the two contexts of clinical audit and service-commissioning. For example, to address inequalities in healthcare it is necessary to consider both general population and patient population models, illuminating inequalities by gender or socio-economic status in healthcare access, utilisation and outcomes. There is a twenty year history of information systems for sharing clinical guidelines, and pathways of care in general [57], [58]. What now needs to be developed is an open, unified modelling framework [59] for building information systems that can illuminate the differences between observed and expected care pathways, and simulate ‘what if’ scenarios for developing care services. The motivation was to build an informatics system that healthcare professionals would use to describe the expected care pathways for services in the health system. The system would then ingest actual care journeys of patients from electronic health records and the produce an analyse the variance of observed from expected care. By segmenting the population into groups based on demographics factors, inequalities in care provision could be identified.

6.2.3 Closing the loop

The vision behind policy simulation and care pathway variance analysis was to develop the information systems needed to improve healthcare planning, by monitoring inconsistencies and inequalities in healthcare provision for people with chronic vascular disease, to assess the likely impacts of interventions intended to improve patient care and public health. In effect closing the loop to ensure the evidence form practice is fed into the policy making process which in itself changes practice.

The general method is as follows and is also shown graphically in Figure 6.1.
1. Extract, summarise and analyse care pathways from clinical information systems.

2. Use the information to help refine care pathway analysis tools with each implementation theme.

3. Use the information to measure patient outcomes.

4. Use the information to measure inequalities in health and healthcare provision.

5. Simulate the effects for local populations of modifying care pathways, including: changes in referral thresholds; changes in treatment thresholds; public health interventions; and targeting under-served groups.

6. Use the analyses and simulations to support decisions about commissioning and quality improvement in local services.

7. Return to step 1.

6.3 Appraisal

As both IMPACT and COCPIT were designed to work as standalone tools as well as in a closed loop system, the appraisal firstly evaluates each paper individually.

6.3.1 Policy Simulation

Any simulation is an approximation and simplification of reality. The challenge is to create a simulation that is realistic enough to make valid predictions, but is not so complex as to be impenetrable to understanding or computationally intractable. From a technology point of view IMPACT was a success in that it made
Figure 6.1: Closing the loop with COCPIT and IMPACT.
a usefully complex model available through a web browser, and made dynamic
simulations and choice experiments simple and easy to use for non-technical spe-
cialist. It provided a generic modelling framework with which usefully complex
models of diseases could be created and large populations could be simulated in
near real-time. This performance was attributable to the use of GPU cards for
parallel execution of code [60].

Another objective that was achieved with IMPACT was to provide complete
transparency to the entire modelling chain. Many simulation tools and packages
are black-box, meaning their internal algorithms and assumptions are not open to
review and critique. IMPACT provided through its website details of exactly how
the simulator worked; how the models were constructed and the evidence base
used; details of model validation and the datasets used in the validation process;
even the source coded was available via the website. IMPACT also implemented
Research Objects (see Chapter 5 and it was possible to export simulations as
Research Objects.

Developing a model, such as the chronic heart diseases (CHD) model that
was developed initially for the IMPACT simulator is a complex, resource inten-
sive process. The evidence required to build such a model is scattered through
the literature, coming from trials and epidemiological studies. Synthesising the
evidence to create the model requires a person who has broad and deep domain
knowledge (most likely therefore a clinician) and a significant amount of their
time. Even if this barrier can be overcome we are then faced with the problem of
finding data against which the model can be validated. In the case of CHD, and
this will be the case for all long-term conditions, this requires an epidemiological
study of a large population over a long enough period for the outcome to be
known.
IMPACT was not widely adopted amongst the target community of commissioners and policy makers. One of the reasons can be attributed to the availability of models for the reasons outlined above. However, the main reason was a failure to engage beyond the research community and work with the NHS. Without this level of co-development with the target end user community and a commitment to meet their needs and expectations, informatics tools will be seen and treated as nothing more than an interesting curiosity.

6.3.2 Care Pathway Analysis

In contrast to IMPACT, COCPIT was developed in conjunction with the end users of the tool, and this is a key factor in its success. It has multiple applications in safety, audit and care quality and it can also be used for case finding and risk stratification. It has also been the inspiration for the Missed Opportunities Modelling theme in HeRC [33], [61]. The care pathway variance analysis methodology was a novel way to examine the differences between expected and actual care. Further exploration of a method based on shortest distance string matching as described in Publication 7 shows much promise. However, the methodology has unresolved problems. A conceptual model of a two node pathway is shown in Figure 6.2, and without additional data it is impossible to distinguish the reasons for patients leaving or joining care pathways at nodes that are neither the first or an endpoint. For example, from the electronic health record data a patient moving away from the area served by the healthcare provider is indistinguishable from a patient that has become disengaged from their healthcare. Perer’s Care Pathway Explorer [62] replicates many of the findings and features of COCPIT, whilst adding novel flow visualisations. COCPIT is available at the time of writing as a commercial product of the University of Manchester and Salford NHS
Figure 6.2: Conceptual care pathway analysis model.

spin-out, North West eHealth (www.nweh.org.uk).

6.3.3 Closing the loop

IMPACT and COCPIT are the building blocks on which a closed loop healthcare system can be built, but it remains a theoretical proposition as we have not yet tested it in practice. The reasons for this are consistent with the emerging theme of this research, which is that the informatics solutions are capable, but it requires a health system that is willing and able to embrace the concept of closed loop healthcare. There is one important consideration that needs to be taken into account when planning such a change, which is the time-scale. The question that needs be asked is how long will it take for the evidence of effectiveness to emerge in the electronic records? Decades would too long, less than six months is unrealistic. The appetite and patience for change of the health system managers
is a key determinant of success. The next chapter was the result of consideration of these factors.

### 6.3.4 Metrics

Publication 4 was presented at MedInfo 2010 and was selected as one of the best papers for publication in the journal Methods of Information in Medicine which is the journal of the International Medical Informatics Association (IMIA). It has been cited seven times. Publication 6 was published in the series Studies in Health Technology and Informatics which is prominent in the field of Health Informatics. It has been cited by six other publications. Publication 7 was presented at the American Medical Informatics Association Conference 2014 and was published in the proceedings. It has not yet been cited by any other publications.
Chapter 7

Combining Health Data Uses to Ignite Health System Learning


*This is Publication 8 in the List of Publications.*

7.1 Context

Health systems are paradoxical. They aim to treat diseases to enable people to live healthier, longer lives. By enabling people to live longer they increase the burden on the healthcare system because a greater need for healthcare is an inevitable part of ageing. Therefore by meeting their aim health systems make it harder to meet that aim in the future. These systems contain an inherent positive feedback loop. This can be easily seen from the increase in life expectancy across the developed world [63]. The success at treating some diseases indirectly increases the prevalence of other disease later in life, such as dementia [64], simply
because people reach an age where these diseases occur. The success of increasing life expectancy has another consequence. The social care burden also increases with age; the elderly require more help to live. At the same time as the increased demands from an ageing population there have been increasing healthcare demands caused by changes in lifestyle across the developed world. The increase in the prevalence of obesity and diabetes has been predicted to be such that it could consume the entire NHS budget by 2020 [65]. The increasing demands on resources [66] comes at a time where in the wake of the global financial collapse the state of the public finances and austerity policies require healthcare budgets to be cut. There is a real need to do more with less. This requires healthcare systems to optimise the usage of resources to provide the maximum health gain for the population they serve.

7.2 Motivation

Health systems have a tendency to organise in a compartmentalised way. There are silos of commissioning, audit, management of finances, research, public health etc. Each silo will have its own data analysis pipe-line. This duplicates effort and dilutes the resources available for analysis, impeding system-wide knowledge exchange. The concept of Learning Health System [12] refers to a health system that is organised to adapt its services, processes and structure to optimise its performance based upon the evidence of past performance and through experimentation [67]. The motivation behind this work was to develop the concept of a learning health system, identifying the re-engineering of information systems and data flows that would be necessary, particularly in the context of the NHS. This work weaves together the three research themes of eLabs, Research Objects and Closed Loop Healthcare to create the blueprint for a learning health system
organised around the optimisation of care pathways. Creating a learning health system requires cultural change as well as re-engineering of the information systems and data flows. Learning health systems are by their very definition about evidence based incremental change. We posit Data Action Latency as the metric that can measure this change and enable evaluation of the maturity of a learning health system.

7.3 Appraisal

At the time of writing this work had only just been published and so it is not possible to objectively assess its impact. However, the key ideas developed in the work were developed into a proposal to the UK government in 2015 to create four Connected Health Cities across the North of England. Each of these will pilot the implementation of a learning health system organised around care pathway optimisation. The impact of this work is to be found in the UK governments commitment to Health North [68] with £20m committed between 2015-18.

7.3.1 Metrics

Publication 8 was published in the journal Methods of Information in Medicine which is the journal of the International Medical Informatics Association (IMIA). It has not yet been cited by any other publications.
Chapter 8

Discussion

The aim of the work presented in this thesis was to examine how we can re-engineer our healthcare systems to make the best use of evidence from practice to deliver care that is safer, more effective, more efficient, more acceptable and distributed with greater equity. It was postulated that this will require tighter coupling between services and research to establish a reinforcing feedback loop. The concept of a closed loop between service and research is the foundation on which learning health systems are built. It has its roots in the systematic clinical audit in the NHS in the 1990s [69]. Friedman developed a six step model to describe the loop in a learning health system [70] as shown in Figure 8.1. Friedman describes the characteristics of the information systems that enable each step in the cycle. Friedman’s model is domain independent and scale-free, and so there should be a high degree of correspondence between the characteristics of the information systems presented in this thesis and those described in Friedman’s model. The first two steps required information systems for “managing communities of interest” and “policies for governing access to data” that together, is a concise way to describe the functionality of the eLab. The third step requires “technology for aggregating and analysing data”, which the COCPIT tool provides for care
pathway analysis. Step four calls for “technology and policy for making knowledge persistent and shareable” for which Research Objects are a proven solution. Step five needs “mechanisms for tailoring messages to decision makers” and step six “mechanism for capturing changed practice” both of which were objectives for the development of the IMPACT tool. There is excellent correspondence between them for every step of the model, which since they have been developed independently, mutually reinforces face validity of both.

However, there are limitations of the informatics platforms presented in this thesis. Data Action Latency (DAL) was proposed as the key metric to measure the performance of a closed loop healthcare system and it would require fully integrated systems exchanging detailed provenance information about decisions,
actions and outcomes to achieve this. Whilst Research Objects satisfy the functional requirements for the transactional currency of learning health systems, this is yet to be tested in practice, as the systems presented were not integrated to the degree that would be required to implemented a closed loop system. We should also consider whether the tools have general applicability and can be reused in a wide range of scenarios. The eLab and Research Objects are generic, reusable technologies that have abstraction and practical utility in balance. The same cannot be said for IMPACT and COCPIT. The cost of developing models for the IMPACT tool renders it impractical as a tool for learning health system. COCPIT however is easily used and applied, but it is limited to producing evidence based upon before and after comparisons. This is suitable for optimisation but is not applicable to scenarios where there is a choice between options. The randomised control trials provide the methodology for making such choices but how this could be realised within a closed loop system has not been explored and also represents a challenge to the learning health systems community as a whole. Through the course of this work we have built a number of large software systems for the health informatics domain that have two things in common. The first is that the systems are designed for the same end user group, namely non-technical healthcare professionals. The second is that the systems are designed to work with data produced from routine care. Looking back at these systems - FAR-SITE, eLab, IMPACT, COCPIT - we see similar problems arising and similar solutions being applied. At a system design level there are a number of health informatics design patterns, analogous to software design patterns [71], that have emerged. These health information system design patterns are catalogued in Appendix B.

Beyond common platforms, design patterns and the learning health system model, is there an intermediate level such as a blueprint or recipe for introducing
closed loop health systems? The intention being to share know-how and thus foster adoption. For example, what triggers an organisation to initiate health system learning and how does it identify where to start? It would seem reasonable to assume that such a blueprint would need to differentiate between scale, as implementation at national level is surely different to implementation within a single hospital. An extension of this idea would be the development of a learning health systems maturity model, against which progress on implementation could be assessed. This would also act as a guide for further development.

Future closed loop health systems must also begin to incorporate new data sources, beyond health care records. Over the course of the last six years, the period over which this work was completed, two technology driven ideas have come to prominence that have the potential to radically change the nature of healthcare and health systems. These are connected digital health and precision medicine. This work has examined how to best make use of the sources of data to optimise health systems, focusing on practice and research. The advent of pervasive connectivity and ubiquitous computing in modern telecommunications networks and devices introduces a new source of data. This data is generated by the individual, and is about the individual and their interaction with their environment [72]. This is revolutionary for healthcare, as it will provide for the first time high-resolution longitudinal data about an individual and their health. This truly will be a new age of high resolution healthcare. Research is required to understand how this data can be incorporated into a learning health system. We have argued that one of the failings of health systems is the treatment of diseases and not individuals. Pragmatic trials, where patients are randomised at the point of care, enable experiments to be performed to test the effectiveness of therapies in the real world of routine healthcare [73]. The extension to this ideas is the n-of-1 trial where a series of therapies may be tried on the patient to discover
the most effective. This idea becomes more powerful when genome information is incorporated. The cost of sequencing a human genome has fallen by three orders of magnitude over a 20 year period [74] and consequently having the patients genome sequence available at the point of care. Advances in machine learning and increases in computing power are providing insights into the endotypes of diseases. However, whilst a positive development this is still not treating the patient as a gestalt. It is a move to differentiate subclasses of disease and a recognition that response to therapy is not uniform.

Finally, creating a closed loop health system is not simply just a matter of engineering; success requires the existence of a receptive environment in which it can be developed. The environment needs to be open and willing to change practice and operational processes. It is arguable that organisational and cultural change is actually the most difficult aspect of developing a learning health system and the change required is not solely within the health care providers. The boundary between service and research, that causes the delay between knowledge discovery being reflected in clinical practice [12], must be removed. This is in turn requires a new civic partnership and the development of new social contract.
Chapter 9

Conclusion

The aim of the work presented in this thesis was to examine how healthcare systems could be re-engineered to make the best use of evidence from practice in order to deliver better care. We postulated that this will require tighter coupling between services and research to establish a reinforcing feedback loop. Three themes of research were pursued, namely, community health eLabs, knowledge exchange and reuse, and closed loop health systems, which collectively span the five objectives for the work.

The first objective was to develop informatics systems to enable data flows between service and research. The FARSITE model [75] demonstrated how bi-directional information flows can be established across the service and research boundary. This idea was fully developed in the most recent work [76] where bi-directional pipelines of evidence were introduced as the solution for eliminating duplication across secondary uses of clinical audit, service management, commissioning, public health and research. The second objective, to develop informatics systems that enable users to collaborate, was fully developed in the eLab [77], [20], [22], [21] model and continues to refined and developed in further research [19]. The third objective, to develop informatics systems that enable knowledge
to be exchanged and reused, was also fully realised through the development of Research Objects [77], [78], [22]. The fourth objective, to enable the reuse of health care data for the benefit of the communities that produce it, was partially met. The development of the Community Health eLab model will enable this, but requires socio-technical change throughout the health system for it to be implemented and adopted at scale. The fifth and final objective, to examine the informatics required for healthcare systems that can learn from their own data and human insights, was met through foundational work in [20], [21] and concluded in the final publication [76].

Three significant findings emerged from the work as a whole. The first is that informatics methods and systems required to re-engineer healthcare systems for system-wide learning are a viable proposition. The importance of co-design with the end-users can not be overstated; it is critical to success. The second is that health systems are difficult to change, and technology without a cultural change programme to actively make it happen is doomed to failure. “Build it and they will come” is a fallacy. There is also an important finding here about the nature of change. Large scale step change does not work; only incremental change where benefits are accrued with each iteration has a chance of success. The third is the principle of localism - that data generated by a community should benefit the community. The approach of ‘local ownership and local control’ works in practice because the information governance, ethical use and connection to the community needed to interpret the data remains. The final publication set out the blueprint for the informatics for a Learning Health System, based on the findings of the preceding work, and will be the starting point for the next phase of this research.

Looking back through the totality of the work that spans a six year period one thing stands out. There is no clear boundary between service and research, and future gains will be made by further dismantling this artificial construct.
Bibliography


Health’s Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data”. In: *Journal of the American Medical Informatics Association* 21.6 (2014), pp. 957–958.


[65] N Hex, C Bartlett, D Wright, M Taylor, and D Varley. “Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs”. In: *Diabetic Medicine* 29.7 (2012), pp. 855–862.


Appendix A - Published Works
Publication 1: Preserving consent-for-consent with feasibility-assessment and recruitment in clinical studies: FARSITE architecture.
Preserving consent-for-consent with feasibility-assessment and recruitment in clinical studies: FARSITE architecture

John AINSWORTH a,1 and Iain BUCHAN a

a School of Community Based Medicine, University of Manchester, UK.

Abstract. Best practice guidance for clinical studies asks investigators to employ the highest possible standards in privacy and consent. When considering the feasibility of a clinical study, issues of privacy extend not only to actual but also to potential study participants. The consent required to access records to determine whether or not an individual might be eligible to participate in a study is sometimes referred to as consent-for-consent. Some initiatives to enhance the efficiency of study-recruitment could compromise consent-for-consent, for example by inviting a patient to take part in a study without the knowledge of their attending clinician. Through iterative working with experts and examination of protocols we explored a range of scenarios for assessing the feasibility of clinical trials and observational studies, and recruiting participants. The main requirement we identified was to speed up feasibility-assessment and recruitment while preserving the patient-clinician trust relationship that is central to consent-for-consent. We present an appropriate information system architecture, FARSITE (Feasibility Assessment and Recruitment System for Improving Trial Efficiency), and show in principle that faster recruitment into clinical studies need not compromise best practice in privacy or consent. We show that FARSITE is a specific instance of an ‘e-Lab’ architecture for assembling data, methods and expertise around study protocols and defined populations.

Keywords. clinical trial, clinical study, design, feasibility, protocol, consent, recruitment, privacy, ethics, law, e-Lab

1. Introduction

Medical research has explicit governance in most nations, which has been guided internationally by the Declaration of Helsinki and its revisions over the past 45 years [1]. Some nations and agencies have more time-consuming requirements than others for the administration of medical research, and participation in clinical studies can be difficult to achieve. In a review of UK-supported clinical trials [2] more than half of the investigators asked the funding agency for an extension and a third did not hit their recruitment targets. A recent point of debate in the UK about inefficiency in clinical studies has centred on consent. Specifically, consent-for-consent, which means the consent required to search an individual’s health record to determine whether or not they should be invited to participate in a clinical study [3]. Traditionally this has

1 John Ainsworth, Health Methodology Research Group, School of Community Based Medicine, University of Manchester, Oxford Road, Manchester, UK. M13 9PL. john.ainsworth@manchester.ac.uk
involved clinicians and researchers making judgements based on the study protocol and local circumstances. More recently, however, research ethics committees have started to move away from this opt-out approach to consent to an opt-in system whereby the patient alone, or the patient with the advice of their attending clinician, must first declare their wish to be approached to participate in a study. This move has been accompanied by a fall in participation rates, and concerns have been raised over the biases it might introduce into studies, thereby lowering the quality as well as the efficiency of medical research [4].

A recent review [5] of methods to improve the participation rates of clinical studies did not consider informatics methods, perhaps because only randomised or quasi-randomised controlled trials were included. Yet there are relevant informatics initiatives, for example: 1) UK, US and EU E-Science projects such as PsyGrid (www.psygrid.org) and Open-CDMS (www.opencdms.org); 2) “bureaucracy-busting IT” initiatives such as England’s National Institute for Health Research Information Systems Programme (www.nihr.ac.uk/systems); and 3) initiatives building on national healthcare information systems such as the National Health Service (NHS) Connecting for Health Research Capability Programme (www.connectingforhealth.nhs.uk/systemsandservices/research). There is a need for studies of whole-system informatics to support the clinical research cycle (Figure 1):-

![Figure 1. The clinical research cycle](image)

Here we propose an information system architecture that links feasibility assessment with recruitment for clinical studies. We examine the issues of privacy and consent in such integrated systems. We report on our exploratory prototypes and our future plans for production quality implementation.
2. Background

2.1. Status Quo: Ad Hoc Feasibility Assessment and Recruitment

Study feasibility is often assessed on an ad hoc basis by asking clinical staff for estimates of the numbers of patients with particular characteristics they might expect to see in a given time period. With short deadlines and difficulties in accurately searching patient records, over-estimation is common [2].

Once a study commences, eligible patients must be identified and approached. This may be via the patient’s attending clinician during a clinical encounter or by notification such as a letter. Patients are typically identified by searching databases or paper records, which may or may not be systematic. In addition, clinical teams may deal with a number of concurrent studies, each with specific procedure for seeking informed consent during recruitment. So the process is laborious and ad hoc.

2.2. Efficiency need: Automated feasibility assessment and recruitment

In order to speed-up studies and reduce selection bias, there is a need for systems of rapid, accurate identification of patients eligible to participate [4,6]. This applies both to feasibility assessment and recruitment. For a given research protocol, the ideal research information system would parse the protocol, form a search query, and enable the study sponsor to assess the potential recruitment in a specific population while varying the inclusion and exclusion criteria. Thus the study design extends into feasibility, which might benefit the design. When the study is approved, the well-understood protocol and search mechanisms are employed in the same e-infrastructure to aid recruitment. Only at the recruitment stage is it necessary to know the identity of an individual patient, and this disclosure can be restricted to the attending clinician and the patient, as required.

2.3. Framework for privacy and consent

The natural framework for issues of privacy and consent is the clinical information governance plus the research governance for any defined population. In the UK, Local Research Ethics Committees tend to map to such populations, as do the commissioners and local providers of care services – in other words ‘local health communities’. Relevant laws, regulations and guidance may operate at a higher level, but the key trust relationships for clinical studies operate locally.

A specific issue for feasibility analysis and recruitment within a general framework of privacy and consent is consent-for-consent. This refers the requirement to seek an individual’s consent to search their person-identifiable records to determine whether or not they are eligible to be invited to participate in a study. Draft guidance has been issued for this in the UK [3].

2.4. Related work

Existing reports focus on alerting clinicians and/or researchers to patients who might be eligible to participate in studies. For example Dugas and colleagues [7] report the
design and implementation of a workflow system involving email alerts in a tertiary care setting. Weiner and colleagues [8] report an increase in trial participation rates after introducing a paging alert system for research in an emergency medicine setting. We could find no studies that map to the natural governance setting, which is the population.

Technical advances in distributed system security and semantic technologies for knowledge management have enabled new research in this area [9]. The VOTES project (Virtual Organisations for Clinical Trials and Epidemiological Studies) has prototyped a system based on Grid middleware [10] that can perform distributed data queries across multiple clinical data sets from multiple independent organisations using a dynamic trust model. These queries can be used to find patients matching trial eligibility criteria. The ePCRN have also developed a Grid based system for distributed data queries [11]. In addition they have developed an ontology-driven query builder system to simplify the construction of complex eligibility criteria.

To date only prototype systems have been reported by VOTES and ePCRN. However, both systems are technology driven solutions to trial feasibility planning, and do not address the process of trial protocol development and trial enrolment employing consent-for-consent. We believe a process-centric design methodology to be crucial for the development of systems used in a clinical care setting; if the system cannot be accommodated into existing processes then it will fail. In a similar way, the NHS adopts a cautious approach to emerging technology and as a consequence systems such as VOTES and ePCRN that employ Grid technology will face many barriers to adoption.

2.5. Our approach: e-Lab integrated

We envision a system of feasibility assessment and recruitment that integrates fully with the study cycle, merging with study design phases leading to feasibility, and study management phases following recruitment.

We think of information systems to enable research using anonymised personal information from a defined population as electronic laboratories, or e-Labs, bringing together data, data processing methods and expertise in a secure environment. We see feasibility analysis and recruitment as specific use case for a more general e-Lab, within the same framework of privacy and consent. This fits particularly well with consent-for-consent. For research organisations working across a number of populations, we see a federation of e-Labs, reflecting local trust relationships and consistently interpreting research protocols.

In the following sections we describe the requirements for such a system, its architecture and initial prototype implementations.

3. Requirements

Our requirements capture process started with interviews with clinical experts and protocol development experts from the Local Research Networks and was followed up with iterative design based around user interface and system prototypes.

The primary requirements for the FARSITE (Feasibility Assessment and Recruitment System for Improving Trial Efficiency) system were identified as
preserving the consent-for-consent model for clinical trial recruitment; to improve the
efficacy of the clinical trial protocol design process; and, with reference to Figure 1, to
automate as much as possible of the workflow from “hypothesis generation” through to
“recruitment”. We address each of these three fundamental requirements in detail in the
following sections.

3.1. Preserving consent-for-consent and patient-clinician relationship

It is essential that the system preserves: i) the privacy of the patient in respect of
queries to identify patients eligible to be invited to participate in studies; and ii) the
clinician-patient relationship in respect of protecting patients from inappropriate
invitations to participate in studies – for example when a patient is grieving. The
system must be flexible to accommodate changing interpretations of consent-for-

3.2. Improving research protocol design interactively

When designing a research study protocol it is important to balance the need for tightly
defined eligibility criteria against the need to get sufficient numbers of participants to
achieve the required statistical power. A system that enables the user to progressively
test and refine eligibility criteria is required until the correct balance is found. An
automated system that performs a parameter sweep by testing all combinations of all
eligibility criteria within bounds and increments specified by the user would be
efficient.

The definition of the eligibility criteria must allow the user to select clinical codes
quickly and easily. It must also allow for complex combinations of criteria using
Boolean operators. For individual criteria it must be possible to require an exact value
or to specify an upper and/or lower bound.

We recognise that some study criteria may not be recorded (accurately) in
electronic health records [12].

3.3. A unified process model

Existing approaches [5,13,14] have either focused on the trial protocol design or trial
recruitment, without making the connection between the two. We argue that this is
essential to ensure that efficiency gains made do not compromise best practice for
privacy and consent, and any solution must recognise a clear distinction is required
between the actors involved in clinical trial protocol design, and those involved in

We have captured our subject matter experts’ views of the ideal trial protocol
development and recruitment process as a simple flow chart. This is shown in Figure 2.

The initial step is to draft the study protocol. This defines the bounds for each of
the eligibility criteria. The analyst then iteratively refines the eligibility criteria against
the information available and the results of queries. The refined criteria are then
examined by a clinical expert for plausibility and practicality. Issues such as known
miscoding of clinical data may be uncovered at this stage. The iterations converge when analyst and clinician are satisfied that the results are stable and reasonable. The expected number of eligible subjects and the agreed protocol are then submitted to the study sponsor for approval. If the sponsor approves recruitment begins. The attending clinicians of the patients identified by the queries are notified that they are seeing patients who might be eligible to participate in the study. The clinician elects to see which of their patients meet the eligibility criteria, and makes a judgement about whether or not it is appropriate to invite specific patients to participate. The clinician may notice a coding error that has led to the incorrect identification of eligibility. Or the clinician may feel that a particular circumstance, such as a patient undergoing a divorce, makes it insensitive to invite them to participate in a study. The invitation letters and information leaflets are automatically printed – the clinician may choose to print only for selected patients, or print for all of those patients turned up by the study.

**Figure 2.** The clinical study protocol development and recruitment process
protocol query and weed out the inappropriate invitations. For some protocols, extra information, not held in the patient record, may be needed from the clinician to complete the assessment of eligibility. As the study progresses, if ongoing recruitment is required, the system must autonomously run the eligibility queries and notify the clinician if any of their patients are found to be newly eligible to invite into the study.

4. Architecture

Our analysis of the requirements identified the need for the system to distinguish between those who design and coordinate studies, and those who recruit participants. The former are typically researchers or administrators not involved in the direct care of the patient, and the latter are typically the attending clinician of a patient in a context relevant to the study. The clinical researchers should not have unnecessary access to patient-identifiable information – and indeed this is not necessary for their role in protocol design and refinement. They simply need to identify the number of patients that meet a set of eligibility criteria. It is useful to consider the ‘clinical care boundary’ as shown in Figure 3 that divides the clinicians and the researchers. On the clinical side of the boundary there is access to identifiable patient data, whilst on research side there can only be access to anonymised data. Furthermore, no identifiable patient information can cross the boundary. Therefore the system must include an anonymised copy of the electronic health record system that can be queried by clinical researcher as part of the trial protocol design process. The analyst progressively develops the trial protocol by issuing queries against the anonymised repository to determine who many patients will meet the specified eligibility criteria. Although these queries only return an integer count of matching patients, allowing users to issue a sequence of queries leaves the system open to deductive disclosure [15]. For example, if I know that my next door neighbour has only one leg and is asthmatic, and if the query returns a count of one for “+one-leg +asthma”, then I can issue a query of the form “+one-leg +asthma +alcoholic”, that will tell me that my neighbour is alcoholic if the count is one. To counteract this, the trial protocol design system will filter the results of queries to ensure that counts less than five are returned as five.

In order to preserve consent-for-consent the attending clinician must run the query that was constructed by the analyst against the electronic health record system to identify potential recruits. This raises a number of technical issues that need to be resolved. Because the anonymised repository does not contain patient identifiable information the identity of the attending clinician is not known. So, the query must be transmitted across the clinical care boundary and then it must be autonomously rewritten to identify all clinicians with potential recruits. Using this information, the query is rewritten again, this time a specific query for each clinician with eligible subjects is created. This query is constructed to return only the eligible patients for whom the clinician is responsible. The query is stored for future execution. The clinician is notified by email that a trial is active and that they have patients that are eligible. The email contains a HTTP link which when clicked on, executes the query for that clinician, assuming that they can successfully authenticate with the FARSITE system. The results of the query are presented to the clinician as a form in their web browser. At this stage the clinician selects the subjects that are suitable and submits this information back to the system. The system collates the responses and the projected
number of participants is emailed to the trial protocol designer. This may cause the trial protocol to be redesigned. If the number of participants is acceptable, then another email is sent to the clinicians informing them that trial recruitment can begin. The clinician can then log on to FARSITE through the web browser and generate personalised letters and information sheets for each patient.

![Diagram](Image)

**Figure 3.** The clinical care boundary and the information flows in the FARSITE system

Once the trial is registered with the system, the system will autonomously run the query to test if new patients have become available since the last execution.

5. Implementation Plans

5.1. Trial Protocol Development Tool

We have developed a prototype Trial Protocol Designer tool that enables user to construct eligibility criteria queries and retrieve the number of eligible subjects in an anonymised database of diabetic patients. The web based eligibility criteria builder interface is shown in Figure 4. We have also incorporated the same style of query interface into the openCDMS (www.opencdms.org) system to enable identification of eligible trial subjects from prospective cohort study data, which enables users to save their eligibility queries.
In response to our prototype our users have suggest improvements to the user interface. They have request that for each one of the eligibility criteria we present the count of patients that satisfy it. From this they can easily see the weighting of each criteria and its impact on the overall total. Planned future developments include the ability to quickly and easily find the correct clinical code, where the system will provide the user with a suggested list of codes based on their eligibility criteria so far and using an ontology to make informed suggestions [16].

5.2. FARSITE in Salford NHS in collaboration with the Greater Manchester Comprehensive Research Network (GMCRN)

Our first deployment of the FARSITE system will be in collaboration with the Greater Manchester Clinical Research Network (GMCRN) and the NHS in Salford, and this is shown in Figure 5. Salford NHS has one of the most advance Electronic Health Record systems in England, known as the Salford Integrated Record (SIR), which integrates primary and secondary care data to form a single patient record for each citizen of the city of Salford. Salford is also deploying an e-Lab [17], a secure information system for assembling data, methods and expertise around study protocols and defined populations. The Salford e-Lab contains an anonymised repository of patient data extracted from the SIR system. We plan to mount the Trial Protocol Design (TPD) tool within the Salford e-Lab to enable trial protocol design to be performed by the researchers of the GMCRN. We will develop the Trial Recruitment Tool (TRT) as a standalone web application inside the clinical care boundary accessible only to clinicians. The TRT will require two factor authentication, and we hope to be able to use the NHS National Programme for IT (NPfIT) authentication infrastructure which has issues all NHS staff with smart cards. This should enable us to provide a Single Sign On (SSO) solution to the TRT for clinicians. The TRT will interact directly with the SIR system to execute queries on behalf of clinicians to find the identities of eligible patients. The interface between the TPD will be web services using HTTP secured with mutually authenticate SSL.
5.3. Integration with HealthGrid

The model way have presented can be scaled to national or international populations using Grid technologies, whilst preserving the key trust relationship between patient and clinician that operates at a local level. We plan to build a Grid-based federation of e-Labs, initially across the North-West of England, and eventually much wider geographies and populations. OGSA-DQP [18] will be incorporated into the Trial Protocol Designer, and the e-Lab anonymised repository will be exposed through an OGSA-DAI [19] interface, effectively creating a virtualised population-level database, whilst maintaining local ownership and governance. The link between the TPD and TRT will become one to many, but crucially the TRT, and the relationship between clinician and patient that it embodies will be left unchanged.

![Diagram](image)

**Figure 5.** Proposed implementation of FARSITE in Salford NHS for Greater Manchester Comprehensive Research Network

6. Conclusion

We have presented a novel architecture, FARSITE, for integrating clinical and research information systems to facilitate the assessment of feasibility and recruitment in research studies, while preserving consent-for-consent. The system is designed to minimise clinicians’ work-loads in identifying and recruiting their patients into studies, whilst preserving clinical oversight and ability to protect their patients from inappropriate or insensitive approaches. This and other informatics initiatives to
enhance clinical studies should undergo controlled trials as there is a major gap in the evidence base [5]. The need to maximise the population-level utility of health record information, while preserving the privacy of individuals, is not unique to FARSITE – we call the generic architecture “e-Lab”.

Acknowledgments

We are most grateful to Dr John New and Dr Martin Gibson of Salford Royal Foundation NHS Trust for their help in leading the clinical activities in identifying the requirements for the FARSITE system.

References

Publication 2: e-labs and work objects: towards digital health economies.
e-Labs and Work Objects: 
Towards Digital Health Economies

John D. Ainsworth* and Iain E. Buchan

School of Community Based Medicine,
University of Manchester,
Manchester Academic Health Science Centre,
Manchester, M13 9PL, United Kingdom
{john.ainsworth,iain.buchan}@manchester.ac.uk

Abstract. The optimal provision of healthcare and public health services requires the synthesis of evidence from multiple disciplines. It is necessary to understand the genetic, environmental, behavioural and social determinants of disease and health-related states; to balance the effectiveness of interventions with their costs; to ensure the maximum safety and acceptability of interventions; and to provide fair access to care services for given populations. Ever expanding databases of knowledge and local health information, and the ability to employ computationally expensive methods, promises much for decisions to be both supported by best evidence and locally relevant. This promise will, however, not be realised without providing health professionals with the tools to make sense of this information rich environment and to collaborate across disciplines. We propose, as a solution to this problem, the e-Lab and Work Objects model as a sense-making platform for digital health economies - bringing together data, methods and people for timely health intelligence.

Keywords: Health Intelligence, Collaboration, Work Objects, e-Lab, Digital Economy, Health Economy, Analysis Workbench.

1 Introduction

In the 1970s Archie Cochrane and colleagues alerted the medical profession to the need to weed out subjectivity and anecdote from clinical practice [1]. At the same time there was a move to improve the safety of medicines. Since then the evidence-based care movement has grown and is now accepted by most healthcare professionals to be best practice. However, there are serious problems with the evidence on which healthcare and public health practice is based: it is expensive to produce; it takes a long time to produce; it takes a long time to influence professional practice; it is crude, relating to the average participant and simple treatment definitions under ideal conditions – in other words, it gives a low-resolution picture of how a patient might respond to treatment or a how a subgroup of the community might respond to a public health intervention. There is

* Corresponding author.
also a lack of public benefit from investments in science and public services, due
to fragmentation of communities, data and analytical methods. In other words,
silos of research that could be more effective and efficient if the researchers had
easy ways to find and share resources when they need them. The divisions are
common between disciplines, for example social vs. biomedical science investiga-
gations of obesity. But they also exist within disciplines, for example between
biomedical scientists investigating nutritional vs. physical activity components
of obesity. Most of the health informatics literature on electronic health records
and putting evidence into practice is about weaving the existing evidence-base
into healthcare decision-making. The role of clinical information systems in im-
proving the evidence-base, however, has been neglected, but they are essential
to providing a timely and more flexible evidence base for future healthcare. This
future could be called high resolution healthcare; it would enable personalised
medicine, efficient and opportunistic clinical trials, complex (including genomic)
epidemiology, and tactical development of local services based on local environ-
mental factors and outcomes at the population level. High-resolution care and
research requires information systems to link relevant data, methods and people
in a clear and timely fashion.

The history of public health intelligence shows rapid advancement in the dis-
cipline through the application of information technology [2], [3] and [4]. Increas-
ingly complex analysis methods requiring High Performance Computing (HPC)
resources are being used. Simultaneously, there has been a rapid increase in the
range of data sources available to the public health practitioner, encompassing
electronic health records, research databases, geographical information systems
and socio-demographic profiles. Ubiquitous connectivity and middleware enables
HPC resources to be shared, and data collections to be accessed from anywhere.
However the applications used to make sense of these electronic resources them-
selves tend to be very specific to the problem being addressed resulting in iso-
lation of outputs and duplication of effort when the same problem is solved for
each discipline [5].

2 Related Work

Over the course of the past decade, we have witnessed the growth of e-Science [6]
and much progress in developing the middleware required for sharing resources,
both computational and data. The plethora of Grid frameworks [7] and grid de-
ployments represents the main thrust of these efforts, but it has not become the
universal infrastructure envisioned by its pioneers [8]. In fact the most successful
Grid deployments are actually as part of a complete vertical application such as
the CERN Large Hadron Collider Grid. Service Oriented Architectures (SOA),
usually realised through Web Services, offer an alternative approach to sharing,
typically by providing a workflow tool for orchestration [9]. The e-Science move-
ment has also spawned numerous Virtual Research Environments [10] drawing
on the collaboratory concept [11], but no generic, reusable, electronic equivalent
of the laboratory workbench or lab notebook has emerged. The Open Provenance
Model [12] provides a standard way of capturing the history of the production of digital objects, with the goal of providing repeatability of *in-silico* experiments. myExperiment [13] draws on the social networking paradigm to provide a platform for curating and sharing scientific workflows. myExperiment also contains an aggregation mechanism known as a “pack”, which enables user to bind related artefacts together. This capability is further developed as Research Objects in [13]. The Open Archives Initiative (OAI) have developed a standard for aggregating web-based resources through the Object Reuse and Exchange protocol [14], which is being widely adopted within the digital repositories community. The concept of Boundary Objects, as a means of cross-discipline communication, was first identified by Star and Griesemer [15] two decades previously.

3 Motivating Use Cases

The use cases presented below serve to illustrate the need for an electronic laboratory for health.

3.1 Obesity Investigations

The obesity epidemic [16] and its potential to break financial models of healthcare has raised the urgency of understanding the epidemiology of obesity and the effectiveness of large-scale measures to tackle it. However, identifying the determinants of obesity, which are very complex, requires understanding social and behavioural as well as biomedical mechanisms [17]. Obesity-relevant information is contained in a number of large surveys, such as Health Surveys for England and the British Household Panel Survey. However, these surveys are difficult to navigate, and are under-used in obesity research. The difficulty arises from the number of variables measured in each survey, and subtle differences in measurement techniques and variable names, which can only be resolved by digging through supporting documentation. Researchers fail to learn from one another about finding, extracting and analysing relevant data. Furthermore, individual researchers may be unable to reproduce an analysis, based on a complex survey after they have forgotten the steps they took. The statistical analysis is usually encapsulated in scripts, but this is not usually chained to the data extraction. Surveys that are repeated on a regular basis, for example the annual Health Survey for England, may have differences in measurement, sampling, or simply labelling of variables, which makes analysis across surveys difficult. It is unsurprising therefore that social and health scientists asking similar questions using HSE would usually in isolation from one another. Social researchers don’t usually know where or how to get at the full range of data relevant to obesity research, for example data collected by healthcare services or schools. And for obesity research in the public health service, there is often a lack of analytical capacity, for example to resolve spatial or temporospatial patterns of obesity from geocoded data sets.
3.2 Genetic Epidemiology

Understanding the genetic basis for disease, and how genetic factors interact with environments and behaviours is a grand challenge for science. Biotechnologies are providing vast amounts of genetic and gemonic data. For example, out of the three million or so genetic factors that vary between people, half a million factors can now be measured on a blood sample for around two hundred dollars. These points of variation, or Single Nucleotide Polymorphisms (SNPs), are usually studied for their relation to disease states by running statistical analyses over tens of thousands of study subjects, hundreds of thousands of genetic factors and a handful of other factors such as age. This is a computationally expensive task [18], even with the crudest types of analysis. Ideally more relaistically complex analyses, such as seeking clusters of interacting genetic factors, would be commonplace, but this is restricted by statistical and computational limits at present. The development and/or application machine learning methods may make the more complex analyses tractable. Validation of the causal relationship between a genetic variation and a disease state, must take into account environmental exposures of individuals as these may contribute significantly. This information can be acquired through a clinical study of the cases or from medical records. The successful interpretation of genotype and phenotype data requires a specialist understanding of the disease. The ideal genomic research information system would enable collaboration between methodologists (bioinformaticians, biostatisticians and biomathematicians), domain experts (clinicians, epidemiologists and biologists) and computer scientists. The system would provide a timely thinking space for teams of experts to co-develop insights into the genetic basis of disease from a combination of perspectives.

3.3 Pharmacovigilance

Post-marketing surveillance of medicines (also known as Phase IV of clinical trials) is required to assess the safety, and to some extent the effectiveness, of newly licensed medicines ‘in the wild’, outside the artificial environments of clinical trials. Phase III clinical trials do not usually include all of the types of patient, for example women of child bearing age or patients with other diseases taking other medicines, who might be eligible for treatment with the drug after it is licensed. Therefore the evidence from clinical trials does not provide a full picture of the public health implications of the drug. Regarding safety: a system of Adverse Event Reporting (AER) is employed, which relies on clinicians identifying, and reporting harmful effects to a central authority. It may be the case that adverse reactions are not identified as being caused by a particular medicine and so not reported. Important signals about the safety and effectiveness of newly licensed medicines could be extracted from electronic health records. For example, if patient A has the same indication for new medicine X as patient B, but patient A’s physician is not yet prescribing X, then a natural experiment takes place - the challenge is to identify appropriate natural control patients like A and make careful statistical analyses to compare X with existing treatment ‘in the wild’. However, there is no central database that can be analysed; the
data is held within multiple systems that not only cover a subset of the population but it further fragmented by the type of care being provided, typically primary and secondary care. The difficulty of combining the relevant data is further compounded by the need to preserve patient privacy and to comply with the information governance requirements of each organisation that holds a part of the patient’s overall the health record. An ideal system would enable analysts to extract anonymised data across a federation of electronic health record databases, effectively treating it as a single virtual population data set. Effective analysis requires a combination of statistical method expertise and clinical expertise to interpret the findings [19].

3.4 Modelling Healthcare for Populations

Long-term conditions, such as Coronary Heart Disease (CHD), consume the largest proportion of healthcare budgets, and are a major focus of public health initiatives. Moving interventions ‘up stream’ to earlier stages of disease would reduce the amount of suffering over the average lifetime and save money. Health policy makers and those planning and managing local health services are poorly served by over-simple estimates of the potential public health impacts of making changes to the pathways of care or taking preventive public health measures. These estimates are often unreliable [20], because the models do not represent the complexity of the disease, population or care over time. It is possible to construct graphical models [21] and to use Discrete Event Simulation to model a disease in a population [22]. Such a simulation would enable the user to test various different scenarios, with the ability to modify both clinical and public health interventions, and measure both the effectiveness based on clinical outcomes and costs. Larger simulations, in terms of the population size, results in better accuracy but require greater computational resources. Discrete event simulations are amenable to parallelisation, and so there is a benefit to employing HPC resources. The construction of models requires collaboration between health economists, epidemiologists, biostatisticians and typical decision-makers/leaders (public health professionals, healthcare managers, and clinicians). The execution of simulation scenarios is of interest to public health professionals, clinicians and service commissioners and the results of simulations are used to inform policy decisions. The ideal system would enable user to construct and share models around ‘what if scenarios’ easily; to execute individual simulations quickly; and to share simulations and their results.

3.5 Use Case Summary

From these domain specific use cases, we can identify a set of common requirements. The electronic laboratory must:

1. Provide a mechanism for organising work, such that it can be shared, repeated, audited, reused and reviewed.
2. Provide easy access to resources such as data sets and computational resources.
3. Provide support for the scientific method such that investigations can be planned, constructed, executed recorded and repeated.

4. Provide support for collaboration through the formation of ad-hoc communities of interest, both within and between disciplines.

Our goal is to support both the reuse of content and the reuse of software. The curation and discovery of content via Work Objects within an e-Lab can serve to act as an organisational memory, as training materials, as an accelerant to the discovery process, and as means to reduce duplication. Within health services we envisage a key benefit of the e-Lab/Work Object paradigm to be analytical capacity building among the workforce. The reuse of content between e-Labs will require a standard interexchange format for Work Objects to be developed. The e-Lab software architecture must foster the reuse of functionality and interoperability, but allow specialisation for domain specific tasks.

4 The e-Lab

An e-Lab is an information system for bringing together people, data and analytical methods at the point of investigation or decision-making. It provides a secure environment for managing, exploring and analysing data from anonymised, integrated health records. The functional architecture of the e-Lab is shown in Figure 1. The e-Lab provides access to three different types of workspace for each user: personal space that is private to the user; group collaboration spaces that are visible only to members of the group; and public space that is visible to all e-Lab users. Collaboration facilities – such as people search and messaging – are provided, as is the capability to organise communities of interest around Work Objects. Syndication is available for users to track the development of Work Objects. The e-Lab enables access to computational and data resources. Computational resources may range from private compute clusters, required for secure processing of medical records, genomic data and images, to national and international Grids. The e-Lab embeds anonymised clinical data and enforces information governance policies. The e-Lab provides the capability to link across data sources, to perform statistical analysis and visualise the results. Users may upload their own data sets – retaining full control over access rights – and the e-Lab will add it to the data resource catalogue so that it can be used in the same way as the embedded data resources. We distinguish between ‘expert users’ and ‘routine users’. Expert users are able to create and publish methods to support the knowledge discovery process into the e-Lab as Work Objects. These Work Objects can then be re-used by routine users to accelerate their own knowledge discovery. The e-Lab is secured through both technical and operational governance procedures. Maintaining privacy and confidentiality of individuals whose anonymised medical records are stored in the e-Lab is paramount. Privacy preserving data linkage [23] and statistical disclosure control [24] is used. Furthermore, users are only permitted to access data for which they have the approval of the governance board and full audit trails of all activity in the e-Lab are maintained. The e-Lab employs a Service Oriented Architecture (SOA), which
enables both reuse of software and reuse of operational services between e-Lab deployments. We define the core set of e-Lab services to be a Work Object repository, data set repository, metadata catalogue, statistical analysis, visualisation, governance, and access control.

5 Work Objects

Work Objects are central to the e-Lab, providing the capability to curate and share information, which in turn builds analytical capacity and organisational memory. Work Objects are collections of digital content assembled to support a specific work task or a series of work tasks – for example to provide a persistent record of an investigation, to publish to a community of interest a statistical method for reuse, or to group together training examples for a tutorial.

Repeatability. A useful analogy can be drawn between a Work Object and a scientific paper. In theory, the paper should give the community all the information necessary to reproduce the results of the research, however there is rarely sufficient information in the paper for another scientist exactly to recreate the investigation. A Work Object representing an investigation can capture all the information necessary to reproduce the results, by recording each step in the process, the data sources used, any transformations applied, the analysis methods and models used, and the commentary underpinning the interpretation of results.
Reuse. Furthering the analogy with scientific papers, a Work Object must be able to reference other Work Objects, in a similar fashion to citations in papers and these references must be navigable. However the Work Object concept goes further. It is possible to embed a Work Object inside another Work Object. For example a Work Object containing a method of statistical analysis could be used inside any number of Work Objects each representing an investigation.

Permanence. A Work Object must provide a persistent record of activity and the associated findings. The process of publishing a Work Object into the public domain must cause a permanent record to be made. A Work Object contains metadata that enables searches to be made over a collection of Work Objects.

Typing. A Work Object must provide a mechanism that enables constraints to be place on its contents, to define application specific content types, and to describe relationships between the content items. This mechanism enables Work Objects to be typed, and consequently systems that are aware of the type of Work Object that they are producing or consuming can provide a richer user experience. The typing of a Work Object requires the specification of the allowed content items, their format and the required number of each; it requires specification of the precedence of content items, for example “data set A and method B must be populated before results C”; it requires specification of production relationship between contents items, for example “executing query I on data source J produces data set K”. This specification defines a Work Object’s lifecycle that compliant systems will enforce. As an example a Research Object must contain a definition of a research question; the design of the investigation; the ethical approval; the measurements; a record of the steps used to transform the data into results; the results; finished documents about the results. The typing mechanism is extensible, allowing for new types of Work Object to be created as and when required by a community of users.

Graceful degradation of understanding. All systems producing and consuming Work Objects must implement the Work Object as a container; it is not necessary to understand any specific Work Object type. We term these systems Work Object Compliant. An example of this type is a Work Object Repository that is able to store Work Objects, and provides the capability to search for specific Work Objects by querying the metadata. Systems that produce/consume Work Objects and understand one or more types are application specific but are able to reuse components that are Work Object Compliant. Furthermore, Work Objects inherit from OAI ORE [14], and so any system that it Work Object Compliant is also ORE compliant as shown in Figure 2.

Content Items contained in a Work Object maybe embedded directly or indirectly referenced by URI. There are pros and cons associated with either approach. Embedded Content Items can be guaranteed to be immutable and are always accessible. There can be no such guarantees with Reference Content Items, although it may be possible to enforce this through service level agreements with the content provider. It is impractical to embed some content items because of their size, for example genomic data sets, and impossible for others
where they are subject to copyright. A published Work Object is considered to be in the public domain, however it is possible to restrict access to Content Items. Any Content Item may be encrypted so that it is not visible without prior arrangement with the author. This applies to both embedded content items and the URI of a referenced content item. Furthermore, although the URI of a Referenced Content Item may be visible, the content provider may apply access control.

6 Discussion and Future Work

The e-Lab and the Work Object work together to provide a solution to the problems of resource access, collaboration, reuse and organisation of work. We expect that it will be used in the UK NHS to build analytical capacity and accessible organisational memory. The e-Lab model will be fully developed in the North
West e-Health (NWeH) project, a collaborative effort between the University of Manchester, Salford Primary Care Trust and Salford Royal Foundation Hospital Trust. NWeH is developing the e-Lab and the Work Object software. The first operational e-Lab will be deployed in Salford in 2009, with further deployments following across the North West of England. These community e-Labs will be federated creating a virtual e-Lab for large-scale population-based research containing data on over 2 million people across the North West of England, and Work Objects contributed from NHS personnel from all members of the federation – tapping into the existing culture of sharing across the NHS. The core e-Lab software will be further developed across a range of projects in the Northwest Institute for Bio-Health Informatics (http://www.nibhi.org.uk) including the Shared Genomics Project [18], the Obesity e-Lab [25] and the Manchester Collaboration for Leadership in Applied Health Research and Care (Systems Research Theme, which is producing new methods for care pathway modelling and simulation).

We have presented e-Labs, and their enclosed Work Object repositories. This model can be extend to enable sharing of resources and sharing or Work Objects between communities centred around an e-Lab. We introduce the Resource Bus and the Work Object Bus as a means of federating e-Labs (Figure 3). If a community wishes to trust another community it can export a Work Object, which could be used by the receiving community to accelerate service provision or research and ensure that it is consistent across communities. A published Work Object is considered to be in the public domain, however it is possible to restrict access to it contents. Any content item may be encrypted so that it is not visible without prior arrangement with the author or if the content item is indirectly referenced, the content provider may apply additional access control. The Resource Bus enables the sharing of data and computational resources between Health Economies. The Resource Bus enables users to discover the resources that are available to them from other e-Labs, contingent on the trust relationships that exist between any two e-Labs. These resources can then be used as part of an investigation. For example, e-Labs can expose their embedded health data resources, derived from integrated electronic health records, onto the Resource Bus, creating a single virtual database of the entire population from the participating health economies. The virtual population database can then by accessed in the same way as any embedded e-Lab data resources. This model of distributed collaboration ensures that access control and governance arrangements of each e-Lab are maintained at a local level, which is not possible with traditional approaches that utilise a central data warehouse.

The e-Lab Technical Architecture Group at the University of Manchester was established to bring together projects from disciplines outside of health such as bioinformatics and chemistry. The goal of this group is to standardise Work Objects and define a common, reusable e-Lab infrastructure.

**Acknowledgments.** We would like to thank the members of the Obesity e-Lab (funded by the UK Environmental and Social Research Council), Shared Genomics (funded by Microsoft), the Manchester Collaboration for Leadership in
Applied Health Research and Care (funded by UK National Institute for Health Research), and North West e-Health (funded by the UK Northwest Development Agency) projects for their input and the e-Lab Technical Architecture Group at the University of Manchester for many insightful discussions.

References

   http://www.liv.ac.uk/PublicHealth/sc/bua/IMPACT-Model-Appendices.pdf
Publication 3: Why linked data is not enough for scientists.
Why linked data is not enough for scientists

Sean Bechhofer a,*, Iain Buchan b, David De Roure d,c, Paolo Missier a, John Ainsworth b, Jiten Bhagat a, Philip Couch b, Don Cruickshank c, Mark Delderfield b, Ian Dunlop a, Matthew Gamble a, Danius Michaelides c, Stuart Owen a, David Newman c, Shoaib Sufi a, Carole Goble a

a School of Computer Science, University of Manchester, UK
b School of Community Based Medicine, University of Manchester, UK
c School of Electronics and Computer Science, University of Southampton, UK
d Oxford e-Research Centre, University of Oxford, UK

A R T I C L E   I N F O

Article history:
Received 8 March 2011
Received in revised form 18 July 2011
Accepted 5 August 2011
Available online 19 August 2011

Keywords:
Research object
Linked data
Reproducibility
Reuse
Sharing
Publishing

A B S T R A C T

Scientific data represents a significant portion of the linked open data cloud and scientists stand to benefit from the data fusion capability this will afford. Publishing linked data into the cloud, however, does not ensure the required reusability. Publishing has requirements of provenance, quality, credit, attribution and methods to provide the reproducibility that enables validation of results. In this paper we make the case for a scientific data publication model on top of linked data and introduce the notion of Research Objects as first class citizens for sharing and publishing.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Changes are occurring in the ways in which research is conducted. Within wholly digital environments, methods such as scientific workflows, research protocols, standard operating procedures and algorithms for analysis or simulation are used to manipulate and produce data. Experimental or observational data and scientific models are typically “born digital” with no physical counterpart. This move to digital content is driving a sea change in scientific publication, and challenging traditional scholarly publication. Shifts in dissemination mechanisms are thus leading towards increasing use of electronic publication methods. Traditional paper publications are, in the main linear and human (rather than machine) readable. A simple move from paper-based to electronic publication, however, does not necessarily make a scientific output decomposable. Nor does it guarantee that outputs, results or methods are reusable.

Current scientific knowledge management serves society poorly, where for example the time to get new knowledge into practice can be more than a decade. In medicine, the information used to support clinical decisions is not dynamically linked to the cumulative knowledge of best practice from research and audit. More than half of the effects of medications cannot be predicted from scientific literature because trials usually exclude women of childbearing age, people with other diseases or those on other medications. Many clinicians audit the outcomes of their treatments using research methods. This work could help bridge the knowledge gap between clinical trials and real-world outcomes if it is made reusable in wider research [1].

As a further example from the medical field, there are multiple studies relating sleep patterns to work performance. Each study has a slightly different design, and there is disagreement in reviews as to whether or not the overall message separates out cause from effect. Ideally the study-data, context information, and modelling methods would be extracted from each paper and put together in a larger model – not just a review of summary data. To do this well is intellectually harder than running a primary study – one that measures things directly. This need for broad-ranging “meta-science” and not just deep “mega-science” is shared by many domains of research, not just medicine.

Studies continue to show that research in all fields is increasingly collaborative [2]. Most scientific and engineering domains would benefit from being able to “borrow strength” from the outputs of other research, not only in information to reason...
over but also in data to incorporate in the modelling task at hand. We thus see a need for a framework that facilitates the reuse and exchange of digital knowledge. Linked Data [3] provides a compelling approach to dissemination of scientific data for reuse. However, simply publishing data out of context would fail to: (1) reflect the research methodology; and (2) respect the rights and reputation of the researcher. Scientific practice is based on publication of results being associated with provenance to aid interpretation and trust, and description of methods to support reproducibility.

In this paper, we discuss the notion of Research Objects (ROs), semantically rich aggregations of (potentially distributed) resources that provide a layer of structure on top of information delivered as Linked Data. An RO provides a container for a principled aggregation of resources, produced and consumed by common services and shareable within and across organisational boundaries. An RO bundles together essential information relating to experiments and investigations. This includes not only the data used, and methods employed to produce and analyse that data, but also the people involved in the investigation. In the following sections, we look at the motivation for linking up science, consider scientific practice and look to three examples to inform our discussion. Based on this, we identify principles of ROs and map this to a set of features. We discuss the implementation of ROs in the emerging Object Reuse and Exchange (ORE) representation and conclude with a discussion of the insights from this exercise and critical reflection on Linked Data and ORE.

2. Reproducible research, linking data and the publication process

Our work here is situated in the context of e-Laboratories, environments that provide distributed and collaborative spaces for e-Science, enabling the planning and execution of in silico and hybrid studies–processes that combine data with computational activities to yield research results. This includes the notion of an e-Laboratory as a traditional laboratory with on-line equipment or a Laboratory Information Management System, but goes well beyond this notion to scholars in any setting reasoning through distributed digital resources as their laboratory.

2.1. Reproducible research

Mesirov [4] describes the notion of Accessible Reproducible Research, where scientific publications should provide clear enough descriptions of the protocols to enable successful repetition and extension. Mesirov describes a Reproducible Results System that facilitates the enactment and publication of reproducible research. Such a system should provide the ability to track the provenance of data, analyses and results, and to package them for redistribution/publication. A key role of the publication is argumentation: convincing the reader that the conclusions presented do indeed follow from the evidence presented.

De Roure and Goble [5] observe that results are “reinforced by reproducibility”, with traditional scholarly lifecycles focused on the need for reproducibility. They also argue for the primacy of method, ensuring that users can then reuse those methods in pursuing reproducibility. While traditional “paper” publications can present intellectual arguments, fostering reinforcement requires inclusion of data, methods and results in our publications, thus supporting reproducibility. A problem with traditional paper publications, as identified by Mons [6] is that of “Knowledge Burying”. The results of an experiment are written up in a paper which is then published. Rather than explicitly including information in structured forms however, techniques such as text mining are then used to extract the knowledge from that paper, resulting in a loss of that knowledge.

In a paper from the Yale Law School Roundtable on Data and Code Sharing in Computational Science, Stodden et al. [7] also discuss the notion of Reproducible Research. Here they identify verifiability as a key factor, with the generation of verifiable knowledge being scientific discovery’s central goal. They outline a number of guidelines or recommendations to facilitate the generation of reproducible results. These guidelines largely concern openness in the data publication process, for example the use of open licences and non-proprietary standards. Long term goals identified here include the development of version control systems for data; tools for effective download tracking of code and data in order to support citation and attribution; and the development of standardised terminologies and vocabularies for data description. Mechanisms for citation and attribution (including data citation, e.g. Data Cite) are key in providing incentives for scientists to publish data.

The Scientific Knowledge Objects [8] of the LiquidPub project describe aggregation structures intended to describe scientific papers, books and journals. The approach explicitly considers the lifecycle of publications in terms of three “states”: Gas, Liquid and Solid, which represent early, tentative and finalised work respectively.

Groth et al. [9] describe the notion of a “Nano-publication”—an explicit representation of a statement that is made in scientific literature. Such statements may be made in multiple locations, for example in different papers, and validation of that statement can only be done given the context. An example given is the statement that malaria is transmitted by mosquitoes, which will appear in many places in published literature, each occurrence potentially backed by differing evidence. Each nano-publication is associated with a set of annotations that refer to the statement and provide a minimum set of (community) agreed annotations that identify authorship, provenance, and so on. These annotations can then be used as the basis for review, citation and indeed further annotation. The Nano-publication model described in [9] considers a statement to be a triple – a tuple of three concepts, subject, predicate and object – which fits closely with the Resource Description Framework (RDF) data model [10], used widely for (meta)data publication (see the discussion on Linked Data below). The proposed implementation uses RDF and Named Graphs. Aggregation of nano-publications will be facilitated by the use of common identifiers (following Linked Data principles as discussed in Section 7), and to support this, the Concept Web Alliance develop a ConceptWiki, providing URIs for biomedical concepts. The nano-publication approach is rather “fine-grain”, focusing on single statements along with their provenance.

The Executable Paper Grand Challenge was a contest for proposals that will “improve the way scientific information is communicated and used”. For executable papers, this will be through adaptations to existing publication models to include data and analyses and thus facilitate the validation, citation and tracking of that information. The three winning entries in 2011 highlight different aspects of the notion of executable papers. Collage [11] provides infrastructure which allows for the embedding of executable codes in papers. SHARE [12] focuses on the issue of reproducibility, using virtual machines to provide execution. Finally, Gavish and Doh [13] focus on verifiability, through a system consisting of a Repository holding Verifiable Computational Resources.
Results (VCRs) that are identified using Verifiable Result Identifiers (VRIs). We note, however, that none of these proposals provide an explicit notion of “Research Object” as introduced here. In addition, provenance information is only considered in the third proposal, where Gavish and Donoh suggest that the ability to re-execute processes may be unnecessary. Rather, understanding of the process can be supported through providing access to the computation tree along with inputs, outputs, parameters and code descriptions.

2.2. Linked data

Benefits of explicit representation are clear. An association with a dataset (or service, or result collection, or instrument) should be more than just a citation or reference to that dataset (or service, or result collection). The association should rather be a link to that dataset (or service, or result collection, or instrument) which can be followed or dereferenced explicitly. Such linking provides access to the actual resource and thus enactment of the service, query or retrieval of data, and so on, fostering reproducibility.

The term Linked Data is used to refer to a set of best practices for publishing and connecting structured data on the Web [3]. Linked Data explicitly encourages the use of dereferenceable links as discussed above, and the Linked Data “principles” – use of HTTP URIs for naming, providing useful information when dereferencing URIs, and including links to other URIs – are intended to foster reuse, linkage and consumption of that data. Further discussion of Linked Data is given in Section 7.

2.3. Preservation and archiving

The Open Archival Information System (OAIS) reference model [14] describes “open archival information systems” which are concerned with preserving information for the benefit of a community. The OAIS Functional Model describes a core set of mechanisms which include Ingest, Storage and Access along with Planning, Data Management and Administration. There is also separation of Submission Information Packages, the mechanism by which content is submitted for ingest by a Producer; Archival Information Package, the version stored by the system; and Dissemination Information Package, the version delivered to a Consumer.

OAIS considers three external entities or actors that interact with the system. Producers, Management and Consumers, to characterise those who transfer information to the system for preservation; formulate and enforce high level policies (planning, defining scope, providing “guarantees”) and are expected to use the information respectively. OAIS also considers a notion of a Designated Community, a subset of consumers that are expected to understand the archived information.

2.4. Scientific publication packages

One notable precursor to the notion of Research Object presented in this paper is the idea of Scientific Publication Packages (SPP), proposed in 2006 by Hunter to describe “the selective encapsulation of raw data, derived products, algorithms, software and textual publications” [15].

SPPs are motivated primarily by the need to create archives for the variety of artefacts, such as those listed above, that are produced during the course of a scientific investigation. In this “digital libraries” view of experimental science, SPPs ideally contain not only data, software, and documents, but their provenance as well. As we note here, the latter is a key enabler both for scientific reproducibility, and to let third parties verify scientific accuracy. Thus, SPPs are essentially containers that, unlike standard file packaging tools such as tar, or zip, adopt a specific terminology to provide a description of their content.

Such terminology is an e-science specific extension of the ABC class hierarchy [16], previously proposed by the same authors as a generic taxonomy of terms for recording events in the lifecycle of digital objects in a library. Examples of specialisations include terms such as Experiment and Simulation (both types of Event), as well as Model and Theory (a type of Work). Although the taxonomy is simple and does not include terms to describe the relationships amongst the artefacts within a SPP, this proposal pre-dates the idea, common to our Research Objects, of combining containers with vocabularies for expressing a rich description of content.

To the best of our knowledge, the interesting preservation architecture designed around SPPs has remained at a prototype stage, making this more of an interesting point of reference than a baseline for a concrete implementation of an RO assembly and sharing toolkit.

2.5. Content vs. container

In terms of the conceptual models that can support the scientific process, there is much current interest in the representation of Scientific Discourse and the use of Semantic Web techniques to represent discourse structures (e.g. see [17]). Ontologies such as EXPO [18], OBI [19], MGED [20] and SWAN/SIOC [21] provide vocabularies that allow the description of experiments and the resources that are used within them. The HyPER community is focused on infrastructure to support Hypotheses, Evidence and Relationships. The Semantic Publishing and Referencing (SPAR) Ontologies [22] also provide facilities for describing the component parts of documents and the scholarly publishing process.

In the main, however, this work tends to focus on the details of the relationships between the resources that are being described—what we might term the content rather than the container.

2.6. A motivating scenario

We use a scenario to motivate our approach and to illustrate aspects of the following discussion.

Alice runs an (in silico) analysis that involves the execution of a scientific workflow over some datasets. The output of the workflow includes results of the analysis along with provenance information detailing the services used, intermediate results, logs and final results. Outside the workflow she may add background information and interpretation of results. She collects together and publishes this information as an RO so that others can (1) validate that the results that Alice has obtained are fair; and (2) reuse the data, results and experimental method that Alice has described. Alice also includes within the RO links/mappings from data and resources used in her RO to public resources such as the ConceptWiki [23] or Linked Life Data [24], providing additional context. Finally, Alice embeds the RO in a blog post so that others can access it.

Bob wants to reuse Alice’s research results and thus needs sufficient information to be able to understand and interpret the RO that Alice has provided. Ideally, this should require little (if any) use of backchannels, direct or out-of-band communication with Alice. Bob can then deconstruct Alice’s RO, construct a new experiment by, for example, replacing some data but keeping the same workflow, and then republishes on his blog, including in the new RO a link to Alice’s original.

---

7 http://conceptwiki.org/.
8 http://linkedlifedata.com/.
The OAIS model considers three external entities or actors that may interact with the system, producers, management and consumers. In our scenario here, Alice is playing the role of producer, while Bob is a consumer. This desire to reduce the use of backchannels corresponds to the OAIS notion of [preserved] information being independently understandable in the sense that the information can be understood by users without the assistance of the information producer. Bob is thus a member of the Designated Community in OAIS terms.

In order to support this interaction, common structures for describing the resources and their relationships are needed. In addition, we require support for navigation/reference to external resources (such as ConceptWiki entries).

Importantly, ROs may contain references to data that is stored elsewhere. A number of data preservation initiatives are currently in place to ensure the long-term storage and reusability of scientific data on a large scale. While this assumption pushes all data stewardship problems to dedicated data architectures, it also raises the new issue of resolving data references with no guarantee that the target has not been removed. In the RO model we take a best-effort approach that is similar to that of the Web architecture: there is indeed no guarantee that all links inside an RO can be resolved. On the other hand, unlike simple Web pages, ROs maintain a state, as described later in Section 6.3. Among other things, the state reflects the integrity of an RO with respect to the external resources it links to, at the time those resources are accessed.

2.7. Package, publish and preserve

We can identify at least three distinct processes or stages in the scenario described above.

Packaging. In conducting and describing her investigation, Alice brings together a number of different resources, for example a description of her hypothesis; datasets; workflows, scripts or analysis pipelines that she may have used to perform the investigation; intermediate and final results; and dissemination materials relating to the investigation, e.g. “the paper” (in a traditional sense) or presentation materials. These resources (some owned by Alice, some under the control of third parties) are brought together into a single package.

Publishing. Once materials are collected together, they can be exposed in a way that is then (re)usable by others. By publication here we refer to a process which involves the exposure or advertising of results. This could include aspects of “traditional” publication channels but is not limited to this. In the scenario described above, the embedding of an RO in Alice’s blog is publication.

Preservation. Packaging and Publication make information available to others. Preservation aims to ensure that resources are made available in the future. Preservation may also require an element of curation and management of metadata or annotations relating to the preserved objects. In our scenario, once Bob has conducted his own investigation or experiment, making use of the resources and results packaged up in Alice’s RO, he can repackage it along with any additional results of methods that he may have used into a new RO for archiving.

We explicitly consider here publication (exposure) as distinct from preservation. Although preservation is an important consideration in any approach supporting reusability, our focus here is chiefly on the packaging mechanisms (although see discussion of the Wf4Ever project in Section 8).

Fig. 1 shows some of the interactions involved in the scenario, along with the different stages identified above.

2.8. Linked data is not enough!

Through the use of HTTP URIs and Web infrastructure, Linked Data provides a standardised publishing mechanism for structured data, with “follow your nose” navigation allowing exploration and gathering of external resources. For example, [23] uses a Linked Data approach to publish provenance information about workflow execution. The use of RDF (and thus associated representation machinery such as RDF Schema and OWL) offers the possibility of inference when retrieving and querying information.

What Linked Data does not explicitly provide, however, is a common model for describing the structure of our ROs and additional aspects that are needed in order to support the scholarly process—factors such as lifecycle, ownership, versioning and attribution. Linked Data thus says little about how that data might be organised, managed or consumed. Linked Data provides a platform for the sharing and publication of data, but simply publishing our data as Linked Data will not be sufficient to support and facilitate its reuse.

Jain et al. [24] also question the value of “vanilla” Linked Data in furthering and supporting the Semantic Web vision. Their concerns are somewhat different (although complementary) to ours here—with a focus on how one selects appropriate datasets from the “Linked Data Cloud”, a concern about the lack of expressivity used in datasets (thus limiting the use to which reasoning can be usefully employed), and the lack of schema mappings between datasets. The nano-publications of Groth et al. [9] are also looking to add additional shared content on top of the Linked Data approach in terms of minimal annotations. Here we focus more on the need for a (common) aggregation model.

Note that this is not intended as a criticism of the Linked Data approach—simply an observation that additional structure and metadata is needed that sits on top of the Linked Data substrate and which then supports the interpretation and reuse of that data. Furthermore there is a need for the metadata to link the structure of the research resources with the function of the research process. A somewhat simplified picture is shown in Fig. 2 with the RO Layer providing a structured “view” on the underlying resources that can then be consumed by RO aware services.

What is missing, then, is a mechanism to describe the aggregation of resources, which through sufficient description of the contribution of these resources to the research and their relationships to each other, captures the additional value of the collection, and enables its reuse through the exchange of a single object. Scientific publication requires the representation of provenance, versioning, attribution, credit and the flow of intellectual rights.

Our notion of Research Object is intended to supply these aggregations and provide a container infrastructure, facilitating the

---

9 One of these is the NSF-sponsored DataONE project (http://www.dataone.org), which caters to the Earth Sciences community and aims at preserving Observational data.
sharing and reuse of scientific data and results. Such a common model then facilitates the construction of services for the creation, manipulation and sharing of our research results.

3. Characterising reuse

In our scenario, we assert that Bob wants to reuse Alice’s results and observe that the term “reuse” can be used to describe a range of activities. Reuse can come in many different forms, particularly when we consider reuse not just of data but also of method or approach. Thus an experiment or investigation may be repeated, enacting the same sequence of steps, or perhaps repurposed, taking an existing sequence of steps and substituting alternative data or methods in order to arrive at a new, derived, experiment. Objects can be reused as they can be decomposed and then recomposed in different ways. If they encapsulate processes, these processes can be re-enacted or previous executions of the process can be examined. As introduced above, reproducibility is key in supporting the validation of research.

Below, we introduce a number of principles intended to make explicit the distinctions between these kinds of general reuse, and identify the particular requirements that they make on any proposed e-Laboratory infrastructure.

We anticipate that due to the context of e-Laboratories, Research Objects will often encapsulate an enactable experiment or investigation. Thus some of our principles are driven by this assumption and refer in some way or other to being able to reuse or repeat the process.

Reusable. The key tenet of Research Objects is to support the sharing and reuse of data, methods and processes. Thus our Research Objects must be reusable as part of a new experiment or Research Object. By reuse here, we refer to a “black box” consideration of the Research Object where it is to be reused as a whole or single entity.

Repurposeable. Reuse of a Research Object may also involve the reuse of constituent parts of the Research Object, for example taking a study and substituting alternative services or data for those used in the study. By ‘opening the lid’ we find parts, and combinations of parts, available for reuse. The descriptions of the relationships between these parts and the way they are assembled are a clue as to how they can be reused. To facilitate such a disaggregation and recombination, Research Objects should expose their constituent pieces. Thus our Research Object framework also has need of an aggregation mechanism.

Repetable. There should be sufficient information in a Research Object for the original researcher or others to be able to repeat the study, perhaps years later. Information concerning the services or processes used, their execution order and the provenance of the results will be needed. Repetition may involve access to data or execution of services, thus introducing a requirement for enactment services or infrastructure that can consume Research Objects. In the extreme, this may require, for example, virtual machines that recreate the original platform used to enact an analysis or simulation. In addition, the user will need sufficient privileges to access any data or services required.

Reproducible. To reproduce (or replicate) a result is for a third party to start with the same inputs and methods and see if a prior result can be confirmed. This can be seen as a special case of Repeatability where there is a complete set of information such that a final or intermediate result can be verified. In the process of repeating and especially in reproducing a study, we introduce the requirement for some form of comparability framework in order to ascertain whether we have indeed produced the same results. As discussed above, reproducibility is key in supporting the validation and non-repudiation of scientific claims.

Replayable. If studies are automated they might involve single investigations that happen in milliseconds or protracted processes that take months. Either way, the ability to replay the study, and to study parts of it, is essential for human understanding of what happened. Replay thus allows us to “go back and see what happened”. Note that replay does not necessarily involve execution or enactment of processes or services. Thus replay places requirements on metadata recording the provenance of data and results, but does not necessarily require enactment services.

Referenceable. If ROs are to replace (or augment) traditional publication methods, then they (and their constituent components) must be referenceable or citeable. Thus mechanisms are needed for unambiguous reference to versions of ROs and which support discovery and retrieval.

Reveable. The issue of provenance, and being able to audit experiments and investigations is key to the scientific method. Third parties must be able to audit the steps performed in the research in order to be convinced of the validity of results. Audit is required not just for regulatory purposes, but allows for results to be interpreted and reused. Thus an RO should provide sufficient information to support audit of the aggregation as a whole, its constituent parts, and any process that it may encapsulate.

Respectful. A key aspect of e-Laboratories is user-visibility, credit and attribution. The paper citation count is an important metric in measuring the visibility and impact of published work. If we move to RO based publishing, we will require a re-engineering of reward structures for scientists—citation counts are no longer enough if derived works are being built through the reuse or repurposing of data and methods. Explicit representations of the provenance, lineage and flow of intellectual property associated with an investigation are needed.

4. RO principles, behaviours and features

The main purpose of Research Objects is to provide a class of artefacts that can encapsulate digital knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge.

The variety of reusabilities can be seen as a collection of behaviours that we expect our shareable objects to exhibit—these then place requirements on the ways in which our models are defined, and this in turn informs the features of the Research Object Model and the services that will produce, consume and manipulate ROs.

The principles stated above describe properties or constraints on the way in which we see ROs being used or behaving. Below, we outline a number of features that can facilitate the delivery of this functionality.

Aggregation. ROs are aggregations of content. Aggregation should not necessarily duplicate resources, but allow for references
to resources that can be resolved dynamically. There may also, however, be situations where, for reasons of efficiency or in order to support persistence, ROs should also be able to aggregate literal data as well as references to data.

Identity. Fundamental to Information Retrieval Systems is the ability to refer uniquely to an object instance or record by an identifier that is guaranteed to be unique throughout the system in which it is used. Such mechanisms must allow reference to the Object as a whole as well as to the constituent pieces of the aggregation. Identity brings with it the requirement for an account of equivalence or equality. When should objects be considered equivalent? Alternatively, when can one object be substituted for another? This will be context dependent; for example, in a given context, two objects may not be considered equivalent, but may be substitutable (e.g. either could be used with the same results).

Metadata. Our e-Laboratory and RO framework is grounded in the provision of machine readable and processable metadata. ROs will be annotated as individual objects, while metadata will also be used to describe the internal structures and relationships contained within an RO. Metadata can describe a variety of aspects of the RO, from general “Dublin Core” style annotations through licensing, attribution, credit or copyright information to rich descriptions of provenance or the derivation of results. The presence of metadata is what lifts the RO from a simple aggregation (e.g. a zip file) to a reusable object.

Lifecycle. The processes and investigations that we wish to capture in the e-Laboratory have a temporal dimension. Events happen in a particular sequence, and there are lifecycles that describe the various states through which a study passes. ROs have state, and this state may impact on available operations. For example, a study may go through a number of stages including ethical approval, data collection, data cleaning, data analysis, peer review and publication. At each stage in the process, it may be possible to perform different actions on the object. Thus a principled description of RO lifecycle is needed in our framework (see Section 6).

Versioning. In tandem with Lifecycle comes Versioning. ROs are dynamic in that their contents can change and be changed. Contents may be added to aggregations, additional metadata can be asserted about contents or relationships between content items and the resources that are aggregated can change. ROs can also be historical, in that they capture a record of a process that has been enacted. Thus there is a need for versioning, allowing the recording of changes to objects, potentially along with facilities for retrieving objects or aggregated elements at particular points in their lifecycle (see Section 6).

Management. The management of ROs will require operations for Creation, Retrieval, Update, Deletion (CRUD) of those objects. Storage is also a consideration.

Security. ROs are seen as a mechanism to facilitate sharing of data, methods and expert guidance and interpretation. With sharing comes issues of access, authentication, ownership, and trust that we can loosely classify as being relevant to Security.

Graceful Degradation of Understanding. Finally, we outline a principle that we believe is important in delivering interoperability between services and which will aid in reuse of ROs, particularly serendipitous or unanticipated reuse—“Graceful Degradation of Understanding”. RO services should be able to consume ROs without necessarily understanding or processing all of their content. ROs contain information which may be domain specific (for example, properties describing relationships between data sources and transformations in an investigation). Services should be able to operate with such ROs without necessarily having to understand all of the internal structure and relationships. This places a requirement of principled extensibility on the RO model.

This notion of Graceful Degradation of Understanding can also be observed in the layering approach used, for example, in Semantic Web representation languages. Triple store infrastructure can be used to store data represented using RDF graphs. Such graphs may include the use of vocabularies or representations—for instance, descriptions could be applied to resources making use of OWL[25] ontologies. The underlying triple store does not necessarily need to “understand” the semantics of OWL in order to provide useful functionality. For example a number of triple stores support hierarchical classification using simple RDF(S)[26] reasoning. Of course, if applications do understand upper layers, they can provide additional functionality or services.

5. Representation and implementation

In practice, during the lifecycle of an investigation (which spans activities including planning, execution of experiments or gathering of observational data, analysis of data and dissemination/publication) scientists will work with multiple content types with data distributed in multiple locations. Thus scientists use a plethora of disparate and heterogeneous digital resources. Although potentially useful individually, when considered collectively these resources enrich and support each other and constitute a scientific investigation [27].

These resources may vary widely depending on domain, discipline and the particular investigations being performed. We can, however, identify how individual resources constitute familiar parts of an investigation, and these are among the pieces that will make up our ROs.

Questions around a research problem, with or without a formal hypothesis. Descriptions or abstracts.

Organisational context. Ethical and governance approvals, investigators, etc. Acknowledgements.

Study design encoded in structured documents.

Methods scientific workflows or scripts, services, software packages.

Data from observations or measurements organised as input datasets.

Results from analyses or in silico experiments. Observations, derived datasets, along with information about their derivation or capture—provenance, algorithms, analyses, instrument calibrations.

Answers. Publications, papers, reports, slide-decks, DOIs, PUBMED ids etc.

A number of different projects have already been developing what one might describe as RO frameworks. These projects are “e-Laboratories” — environments providing a distributed and collaborative space for e-Science, enabling the planning and execution of in silico and hybrid experiments; i.e. processes that combine data with computational activities to yield experimental results.

Here we discuss this work and how it relates to our overall vision of ROs.

5.1. myExperiment

The myExperiment Virtual Research Environment [10] has successfully adopted a Web 2.0 approach in delivering a social web site where scientists can discover, publish and curate scientific workflows and other artefacts. While it shares many characteristics with other Web 2.0 sites, myExperiment’s distinctive features to meet the needs of its research user base include support for

---

credit, attributions and licensing, and fine control over privacy. myExperiment now has around 3000 registered users, with thousands more downloading public content, and the largest public collection of workflows. Over time, myExperiment has embraced several workflow systems including the widely-used open source Taverna Workflow Workbench. Created in close collaboration with its research users, myExperiment gives important insights into emerging research practice.

In terms of our reuse characterisations, simply sharing workflows provides support for repurposing, in that workflows can be edited, and re-run. myExperiment recognised [28] that workflows can be enriched through a bundling of the workflow with additional information (e.g. input data, results, logs, publications) which then facilitates reproducible research. In myExperiment this is supported through the notion of “Packs”, collections of items that can be shared as a single entity.

The pack allows for basic aggregation of resources, and the pack is now a single entity that can be annotated or shared. In order to support more complex forms of reuse (for example, to rerun an investigation with new data, or validate that the results being presented are indeed the results expected), what is needed in addition to the basic aggregation structure, is metadata that describes the relationships between the resources within the aggregation. This is precisely the structure that ROs are intended to supply, the basic pack aggregation being enhanced through the addition of metadata capturing the relationships between the resources—for example the fact that a particular data item was produced by the execution of a particular workflow. The pack (or RO) then provides a context within which statements can be made concerning the relationships between the resources. Note that this is then one viewpoint—other ROs could state different points of view regarding the relationships between the (same) resources in the RO. We return to a discussion of representation in myExperiment in Section 7.

5.2. SysMO SEEK

Systems Biology of Microorganisms (SysMO) is a European trans-national research initiative, consisting of 91 institutes organised into thirteen projects whose goal is to create computerised mathematical models of the dynamic molecular processes occurring in microorganisms. SysMO-DB is a web-based platform for the dissemination of the results between SysMO projects and to the wider scientific community. SysMO-DB facilitates the web-based exchange of data, models and processes, facilitating sharing of best practice between research groups.

SysMO SEEK is an “assets catalogue” describing data, models, Standard Operating Procedures (SOPs), and experiment descriptions. Yellow Pages provide directories of the people who are involved in the project.

SysMO SEEK provides a retrospective attempt to share data and results of investigation along with the methods that were used in their production. The implementation is built upon, and specialises, generic components taken from the myExperiment project.

A number of challenges characterise SysMO-SEEK. Users want to keep their current, bespoke data formats, with a significant support for spreadsheets. Consequently, individual projects are responsible for keeping their own data in separate repositories requiring a framework which allows for references to data that can be resolved upon request.

Projects are also cautious about data access, sharing and attribution, resulting in a sophisticated model of sharing and access control where data and models can be shared with named individuals, groups, projects, or the whole community at the discretion of the scientists. This supports not only the publication of results, but also collaborative working and sharing.

The information in SysMO SEEK is structured using a model called JERM (Just Enough Results Model) which allows the exchange, interpretation and comparison of different types of data and results files across SysMO. JERM is based on the ISA (Investigation/Study/Assay) format. Within SysMO, experiments are described as Assays, which are individual experiments as part of a larger Study. These Studies themselves are part of a much larger Investigation. The aim is that the JERM will move towards linking Models (Biological models, such as SBML) together with the experimental data that was used to both construct and test the model, within the context of one or more Assays. The JERM model extends ISA and provides specific relationships appropriate to the domain. The ISA format is, however, somewhat “top down”, allowing for the packaging of data relating to specific investigations or studies, but less appropriate for re-assembling or reusing data for differing sources.

ROs would then encapsulate the Model together with information about its simulation environment, parameters and data thereby providing a third party with everything they need to reproduce and validate the model, along with the hypothesis and provenance behind its creation. An addition, this description of the Experimental Narrative is a feature that we are likely to see needed in other scenarios.

In the Systems Biology community, the requirement for ROs has already been recognised. Emerging standards and markup languages, such as SBVRML (Systems Biology Results Markup Language) extend the SBML model format to allow scientists to encapsulate experimental data links with their models. This allows both the representation of computational simulation results and experimental results in the context of a particular model.

Returning to our characterisation of reuse, many of the processes currently described within SysMO are actually wet-lab experiments. As a result, traceability and referenceability are the key kinds of reuse that are needed within SysMO, allowing for validation of the results. With greater use of workflows in the future, repeatability and replayability will begin to play a part.

5.3. MethodBox and NHS e-Lab

MethodBox is an environment for finding variables from data archives for cross disciplinary research effectively “turning data archives into data playgrounds”. It is part of the Obesity e-Lab project addressing the public health need for greater understanding of obesity across biomedical and social perspectives, where researchers from different disciplines may use common data archives but do not usually cross-fertilize their work. The generic MethodBox environment is built on the concept of a social network of researchers “shopping for variables”. A variable is a vector of data measured or observed in some way about a factor such as age, sex, body mass index, etc. The elements of the vector usually relate to an individual taking part in a study. Archives such as the UK Data Archive contain millions of variables grouped into sets such as annual surveys, for example the Health Surveys for England. The supporting documentation for each survey typically contains important metadata about variables. Researchers

11 http://www.sysmo.net/
12 http://www.sysmo-dlb.org/
13 http://www.sysmo-dlb.org/seek/
14 http://www.methodbox.org/
15 http://www.data-archive.ac.uk/
may take days to wade through supporting documentation and large datasets to extract the variables and metadata they need. MethodBox reduces the time required from days to minutes. It does this by mapping variables to metadata from: (1) relevant parts of supporting documentation; (2) sets of variables extracted by users; (3) user-contributed scripts and guidance for deriving, transforming or using variables. A derived variable might be a categorisation of socio-economic status based on household income and other factors, or a categorisation of obesity based on body mass index. The social scientist may have more expertise in measuring socio-economic status and the biomedical researchers expertise focuses on obesity. Thus a cross-talk between disciplines may emerge at a very early stage of research by sharing methods applicable to variables of common interest. Users are able to share their expertise over particular survey variables such as the way questionnaire responses about smoking can be made “research ready” and then analysed appropriately. Scripts for extracting sets of variables, transforming multiple variables into one and building research models are the currency of sharing. The sharing of scripts leads to repurposing of study methods.

The sets of variables in MethodBox “shopping baskets” may be seen as incomplete ROs intended to seed ROs for analysis in external e-Laboratories. In addition, ROs may be initiated elsewhere before being populated with data preparation methods and data extracts in MethodBox. So MethodBox is taking the “Research Object on the inside” approach, anticipating future value of reuse and audit of the semantic aggregation of research entities.

NHS e-Lab\[8\] is an e-Laboratory for socially networked “sense-making” over health related datasets. It operates within the UK National Health Service firewall and introduces the notion of a federation of e-Laboratories sharing ROs across organisational boundaries after checks that the ROI does not contain material that might identify a patient. In addition to security, there is a strong requirement to increase the consistency and efficiency with which NHS analysists perform analyses. The repurposing of ROs encourages sharing of templates instead of duplication of similar analytical processes; the realisability enhances information governance; the repeatability builds organisational memory; and the respectfulness helps to build a reward environment, which can be linked to continuing professional development credits.

In order to “borrow strength” from academia, NHS e-Lab is designed to import ROs from MethodBox where national survey data is needed by those planning local NHS services. Attribution, sharing and audit logs will become particularly important for cross organisation as well as cross discipline sharing.

6. RO stereotypes and versioning

In this section we characterise ROs in terms of a small number of stereotypes, i.e., common patterns of resource aggregation that emerge from an examination of our projects involved in e-Laboratory related activities. Stereotypes characterise ROs according to the two orthogonal dimensions of state and purpose.

More specifically, we introduce lifecycle stereotypes, describing states that ROs can transition into and out of as part of their evolution in time (Section 4), and functional stereotypes which describe the role of the RO in the context of data sharing. We then describe RO evolution in terms of updates and versioning operations that affect state.

6.1. Lifecycle stereotypes

Live objects (LO) represent a work in progress. They are thus mutable as the content or state of their resources may change, leading to the need for version management. LOs are potentially under the control of multiple owners and may fall under mixed stewardship, raising issues of security and access control.

Publication objects (PO) are intended as a record of past activity, ready to be disseminated as a whole. This is in line with our key motivation for ROs, namely to support “rich publication” by moving from traditional paper based (linear) dissemination mechanisms, to aggregations of related and interlinked pieces of information. POs are immutable, and their multiple successive versions are considered as distinct objects. They must be citeable, and credit and attribution are central aspects of the publication process as they are key to providing rewards, and thus incentives, for scientific publication. POs may also make use of ontologies for the representation of the rhetorical or argumentation structure in the publication (see Section 2.5).

Archived objects (AO) encapsulate aggregations that represent the endpoint of an RO’s life, either because it is now deprecated, or has reached a version that the author prescribes to be final. AOs are therefore immutable, with no further changes or versions allowed. For example, an AO may represent a historical record for resources used in an experiment which has concluded, or has been abandoned.

With this simple state classification, we can describe the lifetime of an RO in terms of its evolution from LO, to either PO or AO (the “terminal states”), while at the same time multiple versions of an LO may be created, each evolving independently into POs or AOs.

6.2. Functional stereotypes

Work objects encompass ROs and extend the applications beyond research, for example to business intelligence and audit—where repeatability, replayability and repurposing are key aspects [1].

Exposing objects are wrappers that provide a standardised metadata container for existing data. For example, spreadsheets may be gathered together and aggregated along with the methods used to produce them. This aggregation can be seen as an RO, but it can also be a smaller component, exposing the spreadsheet collection to the RO thereby setting it in a reproducible research context. The Exposing Object provides a Wrapper [32] that allows the spreadsheet to be seen as an RO, facilitating its exposure and integration into the Web of Linked Data.

View/context objects can provide a view over some already exposed data. It is here that ROs can interact with data that is exposed or published using Linked Data principles [3], providing a “Named Graph” for those resources.

Method objects contain methods and descriptions of methods—enabling methodological research to be exposed in an RO and consumed by other ROs in applied, as distinct from methodological, research. This may help to propagate methodological integrity and avoid translation errors for methods.

The OAIS model [14] also identifies variants of aggregation such as dissemination and archival information packages, corresponding loosely to our notion of publication or archived objects.

6.3. Evolution and versioning

At any given point in time, an RO is characterised by (i) its lifecycle stereotype, defined earlier; (ii) its version, and (iii) its value, defined as the union of the values of all its components. Note that when internal components are known by reference, i.e., via
their URIs or other Open Data links, the value of the referenced content is represented by the value of the reference. This means that the value of an RO does not change if an update to any of its components is not reflected by providing a new reference to it.

RO evolution may occur in three possible ways:

by state transition: the legal transitions involving the lifecycle stereotypes are shown in Fig. 3(a);
by in-place update: the update operation produces a new value for the same RO and retains its identity;
by versioning: the versioning operation produces a new RO with a new identity and a new value.

Fig. 3(b) shows a possible evolution of a Live Object X. Version 1 of X, denoted $X_1$(LD), is updated multiple times prior to being published at time $t_1$, as $X_1$(PO). A new version $X_2$(LD) of $X_1$(LD) is then created, which is itself updated multiple times, prior to being archived as $X_2$(AO). Independently, $X_1$(PO) is also archived at some other time $t_2$. Note that, according to the state diagram, neither $X_1$ nor $X_2$ can evolve further after reaching their AO state.

A key characteristic relates to the (im)mutability of both the resources described and the relationships between them. Neither Linked Data nor OAI-ORE tackle the issue of versioning explicitly. In the case of an Archived Object, when a scientist returns to it, it should refer to the same versions of the data that were originally used. Live Objects, however may have contents that are updated.

Mechanisms such as the Probity service\(^\text{17}\) that allow a client to detect when changes have occurred have a role to play here as does the Memento framework of Van de Sompel\cite{33}. Probity allows for recordings of “checksums”, providing some minimum guarantees as to whether resources have changed (but not necessarily providing solutions as to what to do when changes have occurred). Memento provides a versioning mechanism for resources that allows “follow your nose” style navigation as widely used in the Linked Data approach. Considering OAIS again, we can see that our Archive Objects relate to Archival Information Packages which should contain information relating to the preservation (provenance, fixity, etc.).

7. Implementing ROs: linked data and OAI-ORE

Although our argument here is that Linked Data alone is not enough to support scientific publication (and the reuse, reproduction or validation of results), Linked Data does offer a substrate for data publication that can then be used by additional layers providing the necessary aggregations, provenance, attribution, etc.

7.1. Linked data

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web\cite{3}, intended to foster reuse, linkage and consumption of that data. The principles can be summarised as follows:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up or dereferences a URI, provide useful information, using standard representations (for example RDF).
4. Within that useful information, include links to other URIs, so that clients can discover more things.

In the five years or so since the first discussions of the Linked Data approach, the amount of linked data published has been increasing rapidly. The Open Data movement has seen successful pressure on governments to expose and open up data sets—in many cases this is being done using a Linked Data approach. An example of this is data published by the UK Ordnance Survey\cite{18} which provides a SPARQL\cite{34} endpoint (allowing query against an RDF triple store) to data describing administrative geography in the UK. Similar government initiatives are also in place in other countries including the US.

Within the scientific community, datasets are also being exposed using a Linked Data approach. Bio2RDF\cite{35} provides “rdfized” access to information from data sources such as Kegg, PDB, MGI and HGNC. The Linking Open Drug Data (LODD)\cite{20} activity of W3C’s Health Care and Life Sciences Interest Group is publishing and interlinking information relating to drugs using Linked Data. Linked Life Data\cite{21} provides a platform for integrating a number of data sources including UniProt, UMLS, Drug information, PubMed.

Other sources exposed as Linked data include species data\cite{22}, Clinical Trials\cite{23}, MeSH\cite{24} and the ConceptWiki\cite{25} as discussed above.

Our intention is that the basic concept of ROs should be independent of the mechanism used to represent and deliver those objects. However the Linked Data approach has a good fit with the notion of ROs.

Within a (semantic) Web context, the term information resource is used to distinguish those resources (things that might be identified by URIs) for which it is the case that their essential characteristics can be conveyed in a message. Non-information resources are those things that might be identified by URIs, but for which this is not the case. Thus web pages, PDF documents, JPG images are examples of information resources, while people are non-information resources. A number of patterns have been identified\cite{36} using techniques such as content negotiation and HTTP redirection, that support the description of non-information resources.

Thus the separation of the identity of an RO from serialisations of the description of its content reflects the handling of non-information resources—we consider a particular RO to be a non-information resource which may have alternative concrete representations. See below for further discussion of the use of non-information resources within myExperiment.

7.2. Aggregation

The idea of aggregation in a web context has already been addressed by the Open Archives Initiation Object Reuse and Exchange Specification (OAI-ORE, or ORE\cite{37}). ORE defines a data model and a number of concrete serialisations (RDF, Atom and RDFa) that allow for the description of aggregations of Web resources. The key concepts in ORE are the notions of Aggregation, which represents an aggregation of a number of resources; and ResourceMap, which provides a concrete representation of the elements in the aggregation (AggregatedResources) and relationships between them.

The ORE model is agnostic as to the semantics of such aggregations—examples are given which include aggregations of

\footnotesize
\begin{itemize}
\item \footnotesize[17] \url{http://www.probity.org/}.
\item \footnotesize[18] \url{http://data.ordnancesurvey.co.uk/}.
\item \footnotesize[19] \url{http://www.w3.org/wiki/HCLSIG/LODD}.
\item \footnotesize[20] \url{http://linkedlifedata.com/}.
\item \footnotesize[21] \url{http://lod.geospecies.org/} and \url{http://www.taxonconcept.org/}.
\item \footnotesize[22] \url{http://linkedct.org/}.
\item \footnotesize[23] \url{http://www.nlm.nih.gov/mesh/}.
\item \footnotesize[24] \url{http://conceptwiki.org/}.
\end{itemize}
favourite images from Web sites, the aggregation of a number of different resources to make up a publication in a repository, or multi-page HTML documents linked with “previous” and “next” links.

ORE provides a description of Resource Map Implementations using RDF [38], which integrates well with current approaches towards the publication of Linked Data [21].

7.2.1. Aggregations in myExperiment

Work in myExperiment makes use of the OAI-ORE vocabulary and model in order to deliver ROs in a Linked Data friendly way [39]. Although specific to myExperiment, the following discussion is pertinent to the other e-Laboratories.

In myExperiment, packs are created using a shopping basket (or wish list) metaphor. Typical packs contain workflows, example input and output data, results, logs, PDFs of papers and slides. To explore the extension of packs to richer ROs a service has been deployed which makes myExperiment content available in a variety of formats. Following “Cool URI” guidelines,25 entities in myExperiment are considered as Non-Information Resources and they are given URIs. Content negotiation is then used to provide appropriate representations for requests, separating the resources from their explicit representations. RDF metadata is published according to the myExperiment data model which uses a modularised ontology drawing on Dublin Core, FOAF, OAI-ORE, SWAN-SIOC, Science Collaboration Framework, and the Open Provenance Model (OPM)26. In addition to this “Linked Data” publishing, myExperiment content is also available through a SPARQL endpoint27 and this has become the subject of significant interest within the community. It is effectively a generic API whereby the user can specify exactly what information they want to send and what they expect back—rather than providing query/access mechanism via specific API functions. In some ways it has the versatility of querying the myExperiment database directly. But with the significant benefit of a common data model which is independent of the codebase, and through use of OWL and RDF it is immediately interoperable with available tooling. Exposing data in this way is an example of the “cooperate don’t control” principle of Web 2.0.

This brings myExperiment into the fold of the other SPARQL endpoints in e-Science, especially in the life sciences area [27] and we are beginning to see workflows that use the data provided by such endpoints. In minutes a user can assemble a pipeline which integrates data and calls upon a variety of services from search and computation to visualisation. While the linked data movement has persuaded public data providers to deliver RDF, we are beginning to see assembly of scripts and workflows that consume it — and the sharing of these on myExperiment. We believe this is an important glimpse of future research practice: the ability to assemble with ease experiments that are producing and consuming this form of rich scientific content.

7.2.2. Extended aggregation vocabularies

Publishing the myExperiment data using Linked Data principles facilitates the consumption of that data in applications, but needs further shared infrastructure to support the description of the RO structure. An RO is essentially an aggregation of resources, and we are using ORE as the basis for describing our RO. As we have mentioned, however, ORE only provides a general vocabulary for describing the relationships between resources in terms of an aggregation, but says nothing about the particular semantics of the relationships between resources. Thus there is no way, for example, of distinguishing between an aggregation of resources in a publication, and the constituent pages in a multi-page HTML document. To enable the description of specific relationships between the aggregated resources, the set of ORE relationships must be extended.

We consider two such extensions here. Firstly, the Research Objects Upper Model (ROUM) provides basic vocabulary that is used to describe general properties of RO that can be shared across generic e-Laboratory services. For example, the basic lifecycle states of ROs (as described in Section 4) are described in this upper model. Secondly, Research Object Domain Schemas (RODS) provide application or domain specific vocabulary for use in RO descriptions. For example, an RO may contain a reference to a service and a data item, along with an assertion that the data was produced through an invocation of the service. Applications which are aware of the intended semantics of the vocabulary used for these assertions can exhibit appropriate behaviour. It is important to stress here that applications that are not aware of these vocabularies will still be able to operate on the overall aggregation structure. This layered approach therefore helps meet our principle for Graceful Degradation of Understanding across e-Laboratory services (see Section 4). OAI-ORE has also been used in other efforts aimed at providing aggregations over scientific data such as SCOPE [41].

The interaction with a Linked Data view of the world is two-fold here. Firstly, one could view the RO as “Named Graphs for Linked Data”, through the definition of an explicit container. The concept of Named Graphs in Semantic Web architecture and languages allows for the identification of selected subgraphs within a single triple or RDF graph. This helps to overcome some of the difficulties.

---

25 http://www.w3.org/TR/cooluris/.
26 http://openprovenance.org/.
in using a simple triple model, for example by being able to assert provenance or trust information with a particular collection of statements in a graph without resorting to reification. Named graphs extend the RDF data model and are supported in query languages such as SPARQL [34].

This also facilitates the exposure or publication of digital content as linked data. Secondly, the RO may also be a consumer of linked data, with linked data resources being aggregated within it.

Fig. 4 shows an enriched view of the layers presented earlier, following a common pattern of exposing content through a protocol layer to a collection of content aware services. The Linked Data Services provide a common mechanism exposing resources in the Data Space. The common protocols adopted here (use of Web architecture, HTTP, URIs, etc.) facilitate access to those resources, but are agnostic as to the content of those resources. Research Object services support the bundling together of resources into ROs (ingest) and the subsequent interpretation of those ROs (access).

8. Discussion

This paper sets out what can be seen as a manifesto for the Research Object concept and approach. We have discussed ways in which information can be repurposed, reused or validated in order to support the scientific research process. These ideas are currently being pursued in a number of research projects.

A challenge common to all emerging collaborative environments that promote open science and the rapid exchange of experimental and pre-publication data and methods is one of trust. As an identifiable container, Research Objects allow us to attribute a measure of trust to the object itself [42], with potential to apply and extend methods for modelling and computing social trust [43], trust in content [44] and trust based on provenance information [45].

The provision of reproducible results requires more than traditional paper publication—or even electronic publication but following the “paper metaphor”. Linked Data provides some of the infrastructure that will support the exposure and publication of data and results, but will not alone enable reusable, shared research and the reproducibility required of scientific publication. Additional mechanisms are needed that will allow us to share, exchange and reuse digital knowledge as (de)composable entities. Our solution to this is ROs, semantically rich aggregations of resources that bring together the data, methods and people involved in (scientific) investigations.

The RO concept provides a layer of aggregation structure that is consistent with the Linked Data view of the world. ROs are both: (1) resources accessible via linked data principles; and (2) will aggregate linked data resources.

As discussed in Section 2.4, previous work has defined the notion of the Scientific Publishing Packages (SPP). Where SPPs diverge from ROs is more in the intent than in the structure: while SPPs are essentially designed for archival purposes, the lifecycle of a Research Object is instead centred around the notion of partial sharing, reusing, and possibly repurposing, making the issue of self-consistency of an RO central to our model.

A number of existing projects are already beginning to apply the RO approach to organising and publishing their data. In particular, myExperiment and NHS e-Lab have notions of prototypical ROs, and the capability to export them using Linked Data principles. By reflecting on how such aggregations play a part in the scientific process, we have proposed a set of principles and features.

Our next steps are to further refine these principles and features and provide implementations that support the lifecycle stages as identified here. The Wf4Ever (“Workflow for Ever”) project,28 aims to support the preservation and efficient retrieval and reuse of scientific workflows. The RO approach is central to Wf4Ever, with workflows being the prime content of the ROs generated. The ROs will be used as containers to package together workflows with data, results and provenance trails, with a key consideration being to support preservation of results. Approaches for validating integrity and mitigating against workflow decay are particular areas of interest for the project—this introduces requirements for aspects such as Aggregation, Versioning and Lifecycle as discussed in Section 4 and will allow us to further investigate issues of preservation which are not explicitly considered here. Two contrasting domains are being explored, Astronomy and Genomics—in the initial use cases, validation and verification of results will be the focus. Current explorations include investigation into more detailed stereotypes and the production of a “Research Object Zoo” identifying concrete examples of the broad classifications introduced in Section 6. In addition, particular use case scenarios within those domains are being used to identify different user roles and their interactions with Research Objects at different stages in the lifecycle as discussed in 6; and identify more specific requirements for content of Research Objects and the vocabularies needed to describe relationships between that content. Use cases cover varying scenarios include the analysis of existing gene expression data from wet lab experiments (genomics) and calculation of luminosities for galaxies.

In closing, we believe that the RO approach will enable us to conduct scientific research in ways that are: efficient, typically costing less to borrow a model than create it; effective, supporting larger scale and deeper research by reusing parts of models; and ethical, maximising benefits for the wider community, not just individual scientists, with publicly funded research.

Acknowledgments

This paper reflects the experiences within several e-Laboratory projects represented by the authors. We wish to thank the project funders and the many other members of the project teams who have contributed indirectly to this work.

References


Iain Buchan is Professor of Public Health Informatics at the University of Manchester. His research interests cover tools and interfaces to support the use of knowledge representation languages. He has developed applications, editors, parsers, APIs and interfaces to support the use of semantic technologies and participated in standardisation activities for W3C, contributing to both OWL and SKOS.

Sean Bechhofer is a Lecturer in the Information Management Group of the University of Manchester. His research interests span tools and infrastructure to support the use of knowledge representation languages. He also drives innovation to support personal and co-produced health-care decision making.
Dave De Roure is a Professor of e-Research in the Oxford e-Research Centre and UK National Strategic Director for Digital Social Research. His research projects draw on Web 2.0, Semantic Web, workflow and pervasive computing technologies and he focuses on the co-evolution of digital technologies and research methods in and between multiple disciplines.

Paolo Missier is a Lecturer, now based at the University of Newcastle. His background is in data and information management (including the semantic variety), and his current research interests include the modelling and design of data and systems architectures in support of computational science. His favourite technology areas include semantic data modelling (RDF), data mining, distributed SW architectures and cloud computing.

John Ainsworth is a Research Fellow in the School of Community Based Medicine, specialising in the application of emerging computing technologies to a wide range of health care challenges from predictive modelling of population needs to novel therapeutic interventions.

Jiten Bhagat is a core developer at the University of Manchester on the myExperiment and BioCatalogue projects.

Philip Couch is a software engineer, developing health information systems for the CLAHRC project. In this role, Philip forms part of a collaboration between the University of Manchester and the NHS, applying state of the art technologies to the analysis of NHS care pathways, including the measurement of health care inequalities. Philip is involved in the design and implementation of software that can be used to simulate the impacts of making changes to care pathways with the aim of significantly improving the health of the UK population.

Don Cruickshank is a senior research fellow in the Web and Internet Science group at the University of Southampton. He develops software for the myExperiment social networking site for scientists.

Mark Delderfield is a Technical Project Manager for the Northwest Institute for BioHealth Informatics (NIBHI). He is responsible for leading a team to develop an e-Infrastructure that provides informatics support for clinical, medical and biological researchers.

Ian Dunlop is a software Engineer at the University of Manchester working on the Obesity e-Lab project and is the lead designer of MethodBox.

Matthew Gamble is a Ph.D. student at the University of Manchester. He is currently involved in the e-Laboratories initiative helping to define Research Objects and is particularly focused on the issues of Trust (reputation, content and provenance based trust) in eScience, web-based scientific collaboration (Science 2.0), and the Semantic Web.

Danis Michaelides is a Senior Research Fellow within the Web and Internet Science group at the University of Southampton where he is involved in building e-Science applications. His main interests are distributed systems and open information systems.

Ian Dunlop is a software Engineer at the University of Manchester working on the Obesity e-Lab project and is the lead designer of MethodBox.

Carole Goble is a full professor in the University of Manchester School of Computer Science, where she co-leads the Information Management Group. She has worked closely with life scientists for many years and has an international reputation in the Semantic Web, e-Science and Grid communities. Carole is the Director of the myGrid project, a team that produce and use a suite of tools designed to “help e-Scientists get on with science and get on with scientists”.

Stuart Owen is a principle developer at the University of Manchester. He is the lead designer of Sysmo-DB, a systems biology project using the Ruby on Rails framework.

David Newman is a researcher in the Intelligence, Agents and Multimedia research group at the University of Southampton. He is currently working on the NeuroHub project and has research interests including e-Science/e-Research, Social Networking, Semantic Web Technologies and Question-Answering Systems.

Shoaib Sufi is a Project Portfolio Manager at the University of Manchester. He is currently responsible for the Technical Software Project Management across a number of projects including the MethodBox system developed as part of the UK ESRC funded Obesity e-Lab project.
Publication 4: IMPACT: a generic tool for modelling and simulating public health policy.
IMPACT: A Generic Tool for Modelling and Simulating Public Health Policy

J. D. Ainsworth¹; E. Carruthers¹; P. Couch¹; N. Green¹; M. O’Flaherty²; M. Sperrin³; R. Williams¹; Z. Asghar¹; S. Capewell¹; I. E. Buchan¹
¹North-west Institute for Bio-Health Informatics, School of Community Based Medicine, University of Manchester, UK; ²Division of Public Health, University of Liverpool, UK; ³Department of Mathematics and Statistics, Lancaster University, UK

Keywords
Discrete event simulation, decision support, public health, policy modelling, computer simulation

Summary
Background: Populations are under-served by local health policies and management of resources. This partly reflects a lack of realistically complex models to enable appraisal of a wide range of potential options. Rising computing power coupled with advances in machine learning and healthcare information now enables such models to be constructed and executed. However, such models are not generally accessible to public health practitioners who often lack the requisite technical knowledge or skills.

Objectives: To design and develop a system for creating, executing and analysing the results of simulated public health and healthcare policy interventions, in ways that are accessible and usable by modellers and policymakers.

Methods: The system requirements were captured and analysed in parallel with the statistical method development for the simulation engine. From the resulting software requirement specification the system architecture was designed, implemented and tested. A model for Coronary Heart Disease (CHD) was created and validated against empirical data.

Results: The system was successfully used to create and validate the CHD model. The initial validation results show concordance between the simulation results and the empirical data.

Conclusions: We have demonstrated the ability to connect healthcare policy-modellers and policy-makers in a unified system, thereby making population health models easier to share, maintain, reuse and deploy.

1. Introduction

Long-term conditions, such as Coronary Heart Disease (CHD), consume the largest proportion of healthcare budgets, and are a major target for public health initiatives. Moving interventions “up-stream” to earlier stages of the natural histories of diseases would delay or prevent subsequent events, thereby reducing the amount of suffering over the average lifetime, and potentially saving substantial costs. Health policy-makers and those planning and managing local health services are, however, poorly served by over-simple estimates of the potential public health impacts of taking preventive public health measures or making changes to the pathways of care. These estimates are often unreliable [1] because the models do not adequately represent the complexity of the population, the disease, or care over time.

Population health impact estimation is usually done by a small group of analysts synthesising evidence and producing a report for a decision-making team. For example, to quantify the potential impact of reducing CHD in a defined population over five years, local policy-makers might ask, “how should the balance be struck between investments in statin tablets vs. smoking cessation vs. physical activity promotion?”

There are several problems with this approach:

a) The available data and literature to consider is vast, complex and increasing;
b) A static report is relatively inflexible and does not enable ‘what if’ scenario planning; thus, relatively few options are appraised;
c) There are not enough analysts to support current decision-making needs, yet it is unlikely that health systems could afford to employ more analysts. Furthermore they are in short supply;
d) Most healthcare commissioning groups do not have the skills or time to build realistically complex models which take all reasonable factors into consideration – decisions may therefore be biased by where a narrowly defined model focuses, which may reflect the interests of service providers more than the needs of the population served.

It is possible to construct visual representations of disease and healthcare pathways, and to use the resulting networks to simulate outcomes for populations. Such a simulation system would enable the user to compare different intervention scenarios, with the ability to modify both clinical and public health interventions, and measure the effectiveness based on both clinical outcomes and costs. The system could bring together public health professionals, clini-
cians and service commissioners in interactive scenario planning activities to inform policy decisions. The ideal system would enable users to construct and share models around what-if scenarios easily; to execute individual simulations quickly; and to interpret simulation results collectively. Larger simulations, in terms of the population size, provide greater accuracy but consume more computational resources. The construction of the best models requires collaboration between epidemiologists, biostatisticians, health economists and typical decision-makers/leaders (public health professionals, healthcare managers, and clinicians).

In this paper we report on the IMPACT system that has been designed to enable this approach, by bringing together model builders, model users and computational resources to participate in shared decision-making.

2. Background

There are many examples of computer modelling and simulation of a disease using a range of methods reported in the literature [2–4]. Coronary heart disease (CHD) is one of the most extensively modelled diseases, so it was chosen the focus for designing a generic system for modelling health impacts in defined populations.

A recent systematic review [5] of cardiovascular disease policy models concluded that models vary widely in their depth, breadth, quality, utility and versatility; with few models being adequately validated or replicated in different settings. Moreover, few were either available for inspection or transparent enough to enable full understanding of the underpinning methods and assumptions. As such, the strengths and limitations of most models were poorly defined; therefore, few were acceptable for use in policy making. For example, a recent model published by the English Department of Health to support cardiovascular screening appears both over-simple and opaque [6]. Out of 70 modelling attempts identified in this area, fewer than 10% published more than one paper, and very few have functioned for a decade or more.

The first IMPACT model [7] used an attributable risk fraction approach and was implemented in a spreadsheet with over 44,000 cells. However, it required extensive training of users and was difficult to deconstruct for validation. Here we report a new approach to the IMPACT model, separating the generic modelling challenge from its application to CHD. Furthermore, we separate the computation of the model from interaction with users, and address the generic problem of simulating public health impact.

3. Objectives

The mathematical methods and computing technologies required to unify model building and use are available [8, 9]. The aim of this work was to harness these methods for health policy making. The objectives were to: 1) develop a versatile, flexible, valid and credible quantitative system for executing population disease models; 2) provide a single framework for domain experts to collaborate on model design and validation; and 3) to provide a decision support capability that enables health professionals to interact with the models.

4. Method

4.1 System Requirements and Analysis

Taylor-Robinson et al. conducted an extensive consultation exercise with policy-makers on their attitudes to modelling and simulation [10]. The findings of that research were used to inform our requirements for the system.

4.1.1 Versatile and Flexible

Our principal objective is to provide a generic system for simulating public health interventions, enabling users to find, ask, and reuse ‘what-if’ questions about options for preventive and clinical interventions in a population’s health. This can be contrasted with the prevailing use of bespoke models often implemented with spreadsheet applications. Consequently, the system must contain a generic execution engine, that can instantiate a given model and perform the simulation. To create models, a model design tool is required that guides the end user through model creation and ensures valid models are created. What constitutes a valid model is intrinsically linked to the design and implementation of the model execution engine. The model alone cannot be executed; it must be configured with additional parameters that define a simulation. Thus a simulation is the combination of the model and the data that characterises the population, the environment, and the interventions being considered. Therefore the system must provide a tool that enables users to define simulations for a given model. We must also consider what the system will be used for. The IMPACT system is intended for answering five types of question:

- How will the burden of disease change over time?
- What will be the impact of specific treatment interventions/technologies?
- What will be the impact of population level/public health interventions?
- In terms of life expectancy is prevention more effective than treatment?
- Are interventions targeted at high-risk groups more effective than whole population level interventions?

The system must provide a tool that enables the results of a simulation to be analysed and visualised, and for comparisons to be made between simulations.

4.1.2 Transparency

Transparency was identified as a key requirement for users to be able to trust and subsequently act on the results of simulations. By transparency we mean that the system must be open to inspection at all levels. Consequently:

- The system software must be open source, so that it can be inspected and critically appraised. The source code must have companion documentation that describes its architecture, algorithms and implementation that is accessible from the system.
- The statistical theory and algorithms underpinning the models and their
execution must be formally documented and accessible.

- For each model, the model builders are required to supply descriptive metadata that describes: the risk factors and disease groups; data sources and main assumptions; the relative risk reductions of interventions; the uptake (availability and adherence) of interventions; the nodes of the graphical model; the edges of the graphical model, defining transition probabilities between health states; the observable outputs of the model and terminology.

- For each simulation, the system must enable users to inspect the configuration that defines the population, environment, and interventions.

### 4.1.3 Accessible

To achieve widespread adoption, access to the system must be as easy as possible for the end user. Thus we are delivering the IMPACT model as a web application that requires no end user installation, configuration or maintenance.

The user interface must be simple and intuitive to use. In order to achieve this different classes of user are defined in terms of their intended use of the system, such that the functions and features available in each user class provides a different view of the system. This enables the complexity of the system to be hidden from the user interface if it is not required. Basic users can create and execute simulations, perform simulation comparisons, and share their results. Advanced users have access to a suite of model building tools enabling them to create new models for wider consumption.

### 4.1.4 Usable for Collaborative Model Creation and Decision Making

The development and validation of models requires collaboration between statisticians/modellers, epidemiologists and health economists. Health policy-making is also a multi-disciplinary process. Web-based social computing technologies are widely deployed and used across many different disciplines [11] for collaborative working. This again favours a web application such that a shared workspace can be created and technologies for storage, retrieval and search of work products can be leveraged. In essence the system must bring people, data and methods together if it is to meet our objectives.

### 4.2 Model and Execution Engine

The life courses of the population of interest are modelled statistically through a two-stage procedure. The first stage is called the population model, which simulates disease incidence. The second stage is called the clinical model, which simulates the progression of deceased individuals to death. The priorities for the population and clinical models are different, so different types of model are used. The overall modelling platform is designed to be a flexible sandbox, allowing various ‘what-if’ scenarios to be trialled in the population.

The population model uses an accelerated failure time (AFT) approach to model the age of onset of the disease of interest [12]. That is, the covariates measured on an individual are assumed to act multiplicatively on the time scale and so affect the rate at which an individual proceeds to become an incident case. Risk factors such as cholesterol, smoking status and blood pressure are incorporated as covariates into the regression. These risk factors are allowed to change over time. The approach can be generalised to allow downstream risk factors to be controlled by upstream risk factors such as diet and exercise. It is also possible to generalise to a multivariate approach, allowing multiple diseases to be considered simultaneously. All incident cases generated by the population model are passed to the clinical model with their associated characteristics at time of incidence.

Interventions in the population model are modelled as changes to the distributions of risk factors. For example, a population level intervention on healthy eating may reduce average salt intake, and the model will propagate this automatically to downstream risk factors such as BMI and blood pressure. Through the AFT model this will thus reduce the speed of progression towards becoming an incident case. Alternatively, a targeted or medical intervention such as a change in statins prescribing trends may reduce, for example, cholesterol levels amongst those with existing high cholesterol. The population model can be run for various potential interventions, and then the incidence distribution of disease cases compared to address questions about the impact of specific interventions and the burden of disease.

The clinical model uses discrete event simulation: for each individual, a sequence of events, from a possible set of events defined by a multi-state model, occur chronologically [13–15]. Various disease states are included as nodes in a graph, and edges represent permitted transitions between the disease states. For example, the multi-state model for CHD is given in Figure 1. A subset of states are specified as entry states, where an individual has presented with CHD symptoms, and another (possibly overlapping) subset of states are defined as sink states, corresponding to death events. A continuous time multi-state model is developed here, which allows higher fidelity simulation and more flexible, realistic intervention policies than a discrete time analogue.

Transitions between nodes in the clinical model are controlled by hazard functions, which describe the instantaneous risk for a given individual making a transition between two disease states. The topology of the graph determines the competing risks for any given disease state. For non-specialists in particular, the graph representation makes constructing and editing a new model relatively straightforward.

Interventions in the clinical model affect patients indirectly, meaning that the results of a simulated intervention do not immediately alter the state of an individual (an impulse intervention) but rather alter the hazard of entering a given state. This is achieved by implementing interventions as proportional adjustments to the transition hazard functions. It is an assumption of the model that the adjustment in the hazard should be proportional. For example, a patient suffering from chronic angina and taking statins will have a reduced hazard of experiencing myocardial infarction, compared to an otherwise identical patient with chronic angina who is not taking statins. It
is possible to specify the uptake and availabilities of various interventions for different disease states.

The clinical model can be run with numerous different interventions applied, such as adjusting the uptake or availability of a particular drug, or even adding a new drug. Since the clinical model simulates patients to death, with two separate nodes, one for death from the disease of interest, one for death from other causes, the effect of these intervention strategies can be analysed on the whole life course. For simulations run under different conditions, various powerful and easy to use tools are available to statistically compare outputs to address policy questions. For example, we may interrogate the number of deaths, counts per state, life tables, time to events, common pathways, time in state, period prevalence and transition probabilities.

A major benefit of this model is the integration of the population and clinical models. This allows policy makers to answer questions such as “should I invest my money in smoking cessation as a preventive measure, or instead spend the money by prescribing more statins to diseased individuals?”

This approach is essentially a generic approach to modelling any non-communicable disease epidemiology and its control. Most diseases and conditions can be described temporally in two phases: an initial clinically not detectable phase, where most preventative interventions aimed at disease risk factors are targeted; and a clinically evident phase where therapeutic interventions are used. For most common diseases, data on risk factors and clinical effectiveness of preventative and therapeutic interventions is available from the literature or national surveys.

A model such as this requires fitting, so that the results it produces are evidence-based, robust, and reflect the population of interest. Furthermore, the model fitting procedure used is able to synthesise evidence from a range of sources.

To fit the parameters of the CHD population model, effect sizes of the risk factors on time to disease onset are estimated from various US [16] cohort studies through AFT regression. The model is also tuned against estimates of the incidence distribution of CHD in the population of interest, where this is available.

The clinical model is able to combine information from cohort studies and elicited expert opinion. For simplicity and tractability, the information obtained is converted into a collection of transition probabilities, or, more generally, constraints that the model attempts to replicate. An example of a transition probability constraint is:

$\Pr[55 \text{ year old male in state B at time } t | \text{ was in state A at time } t-1].$

This model fitting construction is designed to be very flexible, to allow evidence from disparate sources to be synthesised. Constraints can be weighted depending on the value assigned to each source of evidence. These can be formal weights, depending on the statistical variability of data, or subjective evaluations of belief in different sources.

Both models are fitted using methods similar to simulated annealing [17]. For the clinical model, for example, we attempt to maximise the fit of the model to the supplied constraints by minimising the Jensen-Shannon divergence [18], which reflects how well the fitted model reproduces the constraints.

### 4.3 System Architecture

The system was designed around a number of architectural principles. In the interests of transparency, open source technologies were used and the IMPACT Simulator has Service Oriented Architecture to provide a clean separation between components with a view to minimizing the impact of future development and to enable scaling through flexible deployment across a range of hardware platforms. The system is composed of four components: Presentation, Data Management, Broker Service and Simulation Service (Fig. 2).

The Presentation component is a web application that provides the interface for users to interact with the system and is developed to be conformant to the Model-View-Controller design pattern. Web pages offer users functionality in the form of editors and management and reporting tools. Simulations can be configured using editors that perform general create, read, update and delete operations on simulation data. The Model Editor allows users to define models by specifying graphical structures that represent the disease pathways of interest. Graph nodes represent the disease states and arcs represent the allowed transition possibilities.
tions between those states. This editor also allows the assignment of Quality Adjusted Life Years (QALY) weights to each disease state for use in post-simulation analysis. Users can only partially specify models using this editor and model fitting has to be performed before a model can be used with a simulation. The Model Fit Editor allows users to specify initial hazard function parameter values for transitions between disease states and to optimise these by fitting to a set of inputted constraints.

Each model can be fitted many times with users controlling the fitting process by inputting simulated annealing fitting parameters. One criticism of simulated annealing is the difficulty with which fit parameters are chosen. The system provides default values that the developers have found to be sensible for most scenarios. The editor presents diagnostic information following the completion of a fit to inform the user about the suitability of these parameters, including information about the convergence and closeness of fit. The fitting process is completed separately for males and females and the most suitable fit for each gender can be marked for ‘publication’. Published fits provide the hazard functions used for a model during a simulation. The Intervention Editor allows users to define new interventions and to characterise their effect on models. This effect is represented as a reduction in one or more transition risks, inputted by the user as a set of relative risks. Regional variation in the availability of an intervention and the concordance of patients to their use (uptake) is modelled using Regions. The Region Editor allows users to define new regions characterised by uptakes and availabilities for interventions and a population size. Disease incidence in the simulated population is determined using either the Population Model or a cohort specified using the Cohort Editor. A cohort is defined by the age, gender and incident disease state of each member and can be inputted by tabular entry or, more conveniently for large cohorts, by file upload. Finally, the Simulation Editor is used to associate all of the information required to fully specify a simulation for execution. Users will need to select a fitted model, region, cohort and duration. The editor allows intervention uptakes and availabilities, selected through the region, to be altered and new interventions to be defined for the duration of the simulation. This is designed to facilitate straightforward exploration of ‘what if’ intervention scenarios. Any modifications are registered as a text description that can be retrieved later as simulation metadata. The simulation editor can then be used to execute the simulation. Management tools monitor simulations and model fits, providing the current status and allowing users to stop execution. These web pages make use of asynchronous methods and automatically update to provide user notifications such as the completion of a process. Reporting tools offer visualisation and analysis functionality, supporting the effective interpretation of simulation results. Both tabular and graphical representations are employed and full provision is made for large data sets by allowing results to be exported as Excel spread sheets. Existing reports include:

i. The number of disease and non-disease related deaths for the duration of the simulation;
ii. The number of individuals in specific disease states at a specified time;
iii. Life table;
iv. QALY;
v. The time taken for each individual to reach specific disease states;
vi. The disease state of each individual at a particular time;

vii. The complete disease history for each individual (suitable for off-line analysis);

viii. Common disease pathways;
ix. The percentage of individuals in each disease state at each age;
x. Time spent in each disease state (per visit) box plot;
xii. The prevalence of each disease state for specific years;
xiii. The number of individuals that make each transition and the probability of transitions from each disease state for the duration of the simulation;
xiv. The probability of being in a particular disease state one year after being in another disease state (by direct and in-direct transitions);
xiv. Comparison of simulated time to event data with time to event data uploaded by the user (comparisons include Cox regression, Kaplan–Meier curves, age at event distributions and QQ plot).

Many of the reports can be obtained for an inputted list of age ranges and for either or both genders. Comparisons between simulations can be made simply by incorporating information about multiple simulations in the same report. Tools make use of R packages using a COM-based interoperability layer to perform some of the statistical analysis.

User accounts and role-based access control are also managed through the Presentation component. Users are required to log on with a username and password before using the simulator. An initial registration process is required to obtain a user account with default ‘Guest’ privileges. Users with Guest privileges are able to run simulations and generate reports but are not able to perform more advanced tasks such as creating and editing models, interventions and regions. Elevated privileges may be obtained by contacting the site administrator. To provide consistency of functionality and appearance across a range of web browsers and therefore ensuring availability to a wide community of users, graphical components are expressed in both the Vector Mark-up Language (VML) and Scalable Vector Graphics (SVG).

An important feature of the user interface is its ability to allow simulation data to be associated with metadata. Users are able to make general comments about data and provide references for data sources. This metadata could provide, for example, references for intervention hazard adjustments or a description that clarifies the meaning of a disease state. As discussed earlier, this information is seen as critical for the correct interpretation and re-use of the simulation data. The system automatically records the users that create metadata and the dates that this information is provided. The metadata can be updated at any time and this update history is recorded. The IMPACT system supports a form of peer review by allowing users to verify and approve data provided by others. This is achieved by users marking data for review and adding to the data comments. The review date and reviewer user name are automatically captured by the system.

The Data Management component provides data storage, retrieval and validation services for other system components. It exposes an interface that abstracts away details relating to the physical storage of data,
allowing data to be managed more simply in terms of an object model. The creation of this shared component allows the localisation of the business logic to a single piece of software, reducing the likelihood of inconsistency across the system. This component makes use of the NHibernate Framework to map the domain object model to a relational model, persisting data in a SQL Server database. A full recovery model is employed to backup the database, permitting recovery of all data to any point in time following a failure. Further, the databases are regularly and frequently copied to a remote site to allow disaster recovery should any physical damage occur to the system.

The Presentation component interacts with the Broker Service to execute simulations and model fits. This service uses the Data Management component to retrieve information required to configure simulations and model fits and to persist the results. The Broker Service is a web service and requires each client to identify itself using an X.509 certificate, which is validated by comparing it to existing client public certificates held on the host. The use of certificates enables the service to guarantee the integrity of the transferred data, by including a digital signature in the request message, and to authenticate the client. The Simulation Engine is the sub-component that performs the simulation – it was developed in Python and uses some mathematical functions from the SciPy library. The other system components were developed to use the Microsoft .NET Framework. In order to support the execution of simulations in a wide range of operating environments, a channel adapter was developed. The adapter exposes a web service interface to the Simulation Engine, allowing simulations to be configured and results to be returned using SOAP messages over HTTP. This architecture allows multiple Simulation Services to be deployed to execute simulations, ensuring the future scalability of the system to many concurrent simulations and improving the tolerance of the software to problems such as hardware failures and network faults.

5. Validation

The system has been tested by using it to implement and validate the IMPACT model of CHD. The validation process is an integral part of model development. It helps in identifying issues with model implementation, data and assumptions. More important, it is a key element in increasing the model value to policy makers. However, this aspect of model development has been frequently overlooked in cardiovascular disease modelling [5].

We validated the model by using it to simulate the SLIDE cohort, a cohort of survivors of acute coronary syndromes in Scotland [20] (n = 80,241). Our aim was to compare the actual CHD mortality experience of the acute myocardial infarction sub-cohort, with that predicted by the model. For this, we simulated a population with the same age and gender structure as the real sub-cohort, and constructed the simulation taking into account historically plausible treatment effects, using data from systematic reviews and randomized clinical trials [21]. Treatment uptakes were obtained from surveys and audits from a contemporary source to the SLIDE cohort [22]. We generated censoring times in the simulated data with the same distribution as the censoring times in the actual data.

As an example, it can be seen that the model produces an age distribution of CHD deaths that resembles the actual cohort (Figs. 3 and 4). These results are very encouraging, although further work on model constraints is still ongoing. Treatment uptakes have been updated and most of the edges are satisfactorily parameterised. The Kaplan-Meier survival function produced using the model output is similar to the observed (Fig. 5), where the outcome of interest is CHD death. These preliminary results demonstrate that the system is usable for model creation and execution and that the predictions of such models are consistent with observed data. More validation work is needed, specifically regarding comparisons with different cohorts and populations. Because the system can generate many outcome measures that can be derived from time to event data (e.g. life years and quality adjusted life years), as well as prevalence of individuals at individual states at specified times, comparisons with different modeling approaches producing those outcome (such as Markov models for health economics analyses) will be possible to cross compare and to explore the strengths and limitations of different approaches. In addition, the validation process will offer valuable insights towards improving the model ability to produce more accurate estimates of the number of CHD deaths, data visualization and model functionality.
6. Case Study

In this section, we step through the key elements of using the policy model.

The first step is to define the research context and question motivating the use of the model. For the population model, define the necessary population input data. Birth statistics are taken from the Office of National Statistics UK National Statistics database. The risk factor distributions and the correlation between risk factors are obtained from the Health Survey for England (HSE) database. Risk factors included are gender, Body Mass Index (BMI), Systolic Blood Pressure (SBP), cholesterol, diabetes and smoking prevalence. Estimates of the effect of these risk factors, including sex, are obtained using an available US cohort dataset [23]. It is assumed that the effect of these risk factors is the same across populations [24–26].

Fit the model to this data, including the effect of the risk factors, using AFT regression.

The remaining model parameters are then calibrated, using the iterative simulated annealing-type approach, so that it achieves the correct disease incidence for the UK population. That is, the parameters of the model are tuned to obtain correct mortality age distributions and total number of fatalities, taken from the General Practice Research Database (GPRD).

Specify any population level or targeted interventions, and their effects on the risk factor distributions. Interventions are assumed to adjust risk factor distributions. For example, a smoking cessation policy will reduce the probability of an individual being affected by this risk factor.

For the clinical model, set the simulation duration and the initial population either from a specified cohort or the population model output. For exposition, we will investigate the effect on a population of size 10,000 of different levels of preventative treatments regimes. Define the discrete states for the disease of interest: for example, Figure 1 shows the CHD graph structure and Table 1 gives the node descriptions from this graph. Derive the constraint values from relevant data sources and expert elicitation. Fit the model
through the calibration of the hazard function parameter values on each edge against the constraints using simulated annealing. Define the interventions and the related states on which they act; define the probability of adherence of each individual to the intervention regime, which may depend on, for example, their age and gender.

We now illustrate a workflow in the clinical model by considering a toy example. Suppose that, due to budget cuts, a policy maker wishes to consider the removal of statins prescriptions for 50-year-old males who present with heart attacks (note 50-year-old males are considered in isolation here for simplicity of exposition. In reality, one would consider a representative population). In order to do this, a policy maker runs simulations for the policy model under two scenarios: 1) statins set at current uptake levels; 2) statins uptake set to zero. We simulate 10,000 males who have an AMI at age 50 under each scenario, and compare their outcomes. A simple outcome we may wish to compare is how many of the males ultimately die of CHD under each scenario. A graph of this output is given in ►Figure 6. The overall number of deaths is, under current statin levels, 4144 (95% confidence interval [4018.8, 4272.1]); and under statins removed, 4723 (95% confidence interval [4589.3, 4859.7]). So we conclude that removal of statin prescriptions for 50-year-old males suffering AMI would lead to an increased cumulative CHD death rate of 579 per 10,000. It would then be up to the policy maker to trade off the increased death rate against the financial saving, and possibly consider other ways to make the financial saving that may have less impact.

7. Discussion

We have shown that policy modelling can be made accessible and transparent via a web-based system. Unified modelling frameworks such as the one described here may encourage epidemiologists, biostatisticians, modellers, health economists and public health practitioners to contribute to open, accessible policy models rather than creating a blizzard of niche models.

A limitation of the described approach is that the clinical model only considers the public health burden of the specified disease(s) in isolation, and not the overall public health burden. This means that caution is needed in interpreting the results of interventions since demonstrated benefit in the context of a single disease may has simply transferred the burden elsewhere.

Furthermore, the clinical model and the population model in particular are highly dependent on data. The calibration approach adopted for the clinical model does provide flexibility about the type of data suitable for this but there is still a need for a relatively large amount of data, or at least reliable expert opinion, to obtain models with good face and predictive validity.

In on-going work we are returning to the community of planners and policy makers [10] to assess the usability, accessibility and utility of the system and the IMPACT CHD model. The uptake and usage of the system will be monitored, as these will be the key measures of success.

Future work is planned to parallelise the simulation engine to take advantage of multi-core and cluster computing. This will dramatically reduce the simulation run-time making the system more usable for complex models and large populations. The modular nature of the architecture enables the use of cloud computing infrastructure in the future.

The IMPACT simulator will be integrated into the nascent e-Lab population health information system [27]. This will lever electronic health record data to refine, extend and localise models. The e-Lab platform provides the Work Object [27] mechanism as a way of exchanging knowledge between federated e-Labs in different localities. The IMPACT Simulator already has the capability to export IMPACT Simulation Work Object Archives for a specific simulation, to demonstrate proof of concept. These are ZIP archives containing files that represent all of the simulation data and metadata in a semantically explicit way. The metadata includes that inputted by the user, such as the references to data sources, descriptions and general comments. It also includes metadata automatically captured by the system, such as information about users and dates. In addition, the system also provides information about the statistical methods used and their implementation details, such as software versions. Semantic Web standards are adopted to ensure that the meaning of the data is unambiguous. RDF is used to represent the data, relating it to concepts that are specified in an ontology. This ontology is being developed for use with the simulator and builds on other conceptual models including that proposed for Research Objects [28].

In conclusion, rising computing power coupled with advances in machine learning and healthcare information now enables more user-friendly models to be constructed and executed. The IMPACT Simulator represents a system for creating, executing and analysing the results of simulated public health and healthcare policy interventions, which is accessible and usable by modellers and policy-makers alike.

The IMPACT simulator is deployed and available on the Internet at http://www.impactsimulator.org.uk.

Acknowledgments

This work was supported by funding from the UK NIHR Greater Manchester CLAHRC programme; UK MRC IMPACT project, the UK NPR13 scheme and the EU MedCHAMPS project.

References

6. Response from the Faculty of Public Health to the Department of Health Consultation on Economic Modeling of a Policy of Vascular Checks


Publication 5: eLab: Bringing together people, data and methods to enhance knowledge discovery in healthcare settings.
eLab: Bringing Together People, Data and Methods to Enhance Knowledge Discovery in Healthcare Settings

John AINSWORTHa,1, James CUNNINGHAMA and Iain BUCHANa
a The Northwest Institute for Bio-Health Informatics, University of Manchester, UK

Abstract. The discovery of knowledge from raw data is a multi-stage process, that typical requires collaboration between experts from disparate disciplines, and the application of a range of methods tailored to the research question. The aim of the eLab is to provide a web-based environment for health professionals and researchers to access health datasets, share knowledge and expertise and to apply methods for analysis and visualization of the results. The eLab is built around the core concept of the Research Object as the mechanism for preserving, reusing and disseminating the knowledge discovery process. The possible range of applications of the eLab is vast, and so the consideration of the trade off between specificity and generality is an important one, that is reflected in the requirements. The architecture and implementation of the eLab is described, and we report on the deployment of eLabs for applications in primary care, long-term conditions management, bariatric surgery and public health.

Keywords. Collaboration, knowledge discovery, meaningful reuse, electronic health records, research object

Introduction

An eLab is an information system for bringing together people, data and analytical methods for knowledge discovery [1]. The problem motivating the development of eLabs is the shortage of public benefit from investments in science and public services, due to fragmentation of communities, data and analytical methods. In other words, ‘silos’ of research that could be more effective and efficient if the researchers had easy ways to find and share resources when they need them. The divisions are common between disciplines, for example social vs. biomedical science investigations of obesity. But they also exist within disciplines, for example between biomedical scientists investigating nutritional vs. physical activity components of obesity.

1 Mr John Ainsworth, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK. john.ainsworth@manchester.ac.uk
The challenges of remote collaboration in scientific endeavour have been well characterized [2], and include assumptions about terminology and processes, and preserving ownership and credit for work done; informatics can provide tools to address these problems. The eLab provides an environment in which researchers can share not just the end results of their analyses but the intermediary components, for example data cleaning scripts, assumptions behind analysis decisions, allowing researchers who are not co-located to form a deeper understanding of their colleagues’ activities and making tacit expertise tangible.

The concept of Virtual Research Environments (VRE) (also referred to as a Collaboratory), defined very broadly as electronic space to enable researchers to share ideas, infrastructure, data and tools, were proposed almost 20 years ago [3]. In the intervening period, numerous VREs have been created for many different disciplines. They also vary greatly in their capability, from very generic “portal” frameworks, for example Microsoft SharePoint, to cloud-based data repositories, for example figshare [4], to domain specific applications, for example myExperiment for sharing workflows [5]. Two examples from the healthcare domain are I2B2 [6] and SAIL [7].

The risk is that VREs are either too general or they are too specific. For the former, the effort required to customize them to a domain in order to make them useful to end users is prohibitive. Similarly, application specific VREs must undergo considerable rework to reapply to another problem space. Although the eLab shares much heritage with other VRE systems, it is distinguished by (a) the rigorous application of information governance necessary for handling sensitive medical information (b) providing comprehensive support for recording, exchanging and reusing the artefacts of the knowledge discovery process. The eLab is specifically designed for the health domain, but is flexible enough to be customized easily to a wide range of applications therein.

1. Methods

Previous work analysing use cases from multiple domains [8], identified the following core features for the eLab:

1. Provide a mechanism for organising work, such that it can be shared, repeated, audited, reused and reviewed.
2. Provide support for the scientific method such that investigations can be planned, constructed, executed recorded and repeated.
3. Provide support for collaboration through the formation of ad-hoc communities of interest, both within and between disciplines.
4. Provide easy access to resources such as data sets and computational resources.

Research Objects (RO) [9] is the mechanism used in the eLab to fulfil 1 and 2 above. We introduce Method Objects (MO) as a mechanism for automating the manipulation of Research Objects. The eLab itself satisfies requirements 3 and 4. In the following sections, we describe the architecture of the eLab, the design and implementation of Research and Method Objects, and the information governance and security approach.
1.1. eLab Architecture

The eLab employs the Service Oriented Architecture design paradigm and Web Services are used to implement functionality where appropriate. The user interface to the eLab is delivered as a web application. The rationale for this approach is that web applications are zero-configuration on the client side, which is important when end users are widely distributed thus making support difficult. Service Oriented Architecture promotes modularity, loose coupling and hence reuse. It also enables rapid extension and customisation of the eLab through the provision of additional services. This architectural approach also allows tighter security control over access to Research Objects and will facilitate federation of eLabs in the future.

![eLab Architecture Diagram](image)

**Figure 1.** eLab architecture showing major functional components and their interactions

The principal components of the eLab and the main interactions are shown in **Figure 1.** The Workbench is a web-based application that manages local security and social networking elements such as messages, groups and user profiles. The application structure of the workbench has distinct tiers of functionality:

- **Web layer** – This includes the controllers and views for the Model-View-Controller pattern employed by the application and the HTML required for display.
- **Entities** – This describes the models used by the application and the mappings from the objects to the relational database used for persistence.
- **Services** – The business logic layer mediates the retrieval of research objects from the repository and provides services to perform application tasks.
- **Repository** – The repository provides the persistence layer to the database and hides specific data access functionality from the other tiers of the application.
The Research Object Server is a web service for viewing, manipulating and managing Research Objects.

Data can be loaded into the eLab either by users uploading a dataset into a RO or by the system administrator publishing them into the eLab Data Repository. A standalone tool is provided for importing data into the eLab, and the tool can be configured to automatically refresh the data in the eLab from the source at a defined frequency. The tool also provides the capability to pseudonymise or anonymise the data during the import. For each column in the source data, a transformation can be applied before it is written into the eLab. The available transformations include cryptographic functions (SHA-1, SHA-2, AES-256, public-private key, random number), date transformations (e.g. convert a full data to a year), and transformations for geo-locators. The Data Classification and Integration module applies automated routines for data cleaning, aggregation and meta-data extraction. This module can also be accessed through the system administration interface, where these routines can be mainly applied and data can be annotated.

The eLab was developed in C# and JavaScript and use the Microsoft .net (v3) execution environment. All the databases, except the Data Repository, use Microsoft SQL Server, although the use of nHibernate and the Spring framework make the database transparent to the application. MonogoDB [10] is used to provide a No-SQL document data store for the Data Repository. Microsoft Internet Information Services is used to host the web applications and services. Choosing to develop on the Microsoft software stack was not made for solely technical reasons; it was primarily an operational decision. We anticipated that the majority of eLab deployments would be within the UK National Health Service (NHS), where Microsoft technology is pervasive. It is an unrealistic to expect NHS IT departments to have the skills or resources to manage other technology platforms. However, the Service Oriented Architecture and conformance to web service standards enables the integration of services built on other technology platforms.

1.2. Research Objects

The implementation of ROs in the eLab is realised by representing each RO as a graph. Each node in the graph is a content item of the RO and the edges of the graph define the relationship between content items. Each edge is labelled with the relationship type. The set of relationship types is shown in Table 1.

<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is a research question of</td>
<td>Used for attaching the motivating research question</td>
</tr>
<tr>
<td>Is a hypothesis for</td>
<td>Used for attaching the hypotheses under investigation</td>
</tr>
<tr>
<td>Is derived from</td>
<td>Specifies dependency between two content items</td>
</tr>
<tr>
<td>Is input to</td>
<td>Identifies that the content item is used by a function content item</td>
</tr>
<tr>
<td>Is output from</td>
<td>Identifies the content item as output from a function content item</td>
</tr>
<tr>
<td>Is a dataset for</td>
<td>Used for attaching a dataset</td>
</tr>
<tr>
<td>Is a part of</td>
<td>Defines an aggregation relationship between content items</td>
</tr>
<tr>
<td>Is a function of</td>
<td>Identifies the content item as a function</td>
</tr>
<tr>
<td>Is a governance content item for</td>
<td>Used for attaching governance approvals</td>
</tr>
<tr>
<td>Is a governance submission application</td>
<td>Used for attaching governance applications</td>
</tr>
<tr>
<td>Is a visualization of</td>
<td>Used for attaching visualizations of results</td>
</tr>
</tbody>
</table>
The advantage of using a graphical representation is that it allows arbitrarily complex ROs to be constructed from a simple schema, are easily manipulated, can be naturally described by RDF which will enable export, exchange and reuse ROs. An example of a typical RO graph is shown in Figure 2.

Figure 2. An instance of a Research Object graph, showing nodes as content items and edges as annotated relationships between them specifying the type of relation.
1.3. Method Objects

Where a Research Object encapsulates a representation of input and activity surrounding a piece of research, a Method Object (MO) encapsulates a means of augmenting an existing Research Object with additional information in a way that automates the potential use of an eLab system. In a sense a Method Object is a script for recording research actions in the context of an eLab. A Method Object is executed against a Research Object, and that Research Object is then augmented with the information specified in the Method Object. The implementation of Method Objects is based around the notion of proxy objects, which specify a template for matching against existing eLab content, or creating new content. These proxy objects reflect the content of a Research Object; there are Content Item Proxies and Relationship Proxies. A Method Object specifies a set of these proxy objects, each of which specifies a series of steps. When a Method Object is executed each step is executed in turn. When a step is executed the set of proxy objects defined in that step is matched against existing content in the eLab. For any proxy that does not match existing content, new content is created such that the proxy will subsequently match it. In certain cases the addition of new content will require user intervention in order to select between multiple possible instantiations of a proxy object. In such cases the execution of the Method Object is suspended until user input is given. Where there exists a partial match between an item in the eLab and a proxy described in a Method Object, information is added to the eLab content such that it then matches the proxy object.

As an example, a Method Object could be defined that would automatically find any data set within a Research Object that contained geographical information and to produce a map for each such dataset that plotted the geographical region represented overlaid with a summary of the associated information in the corresponding dataset.

1.4. Information Governance and Security

Information Governance in the eLab is guided by a core set of principles. These are: (i) maintain the link between data ownership and control; (ii) expose only the minimum amount of data the user needs; (iii) enforce information governance through technology wherever possible; (iv) require operational safeguards to compliment technical measure. The technical safeguards include access control based on privileges, full audit trails and monitoring, automated reporting of suspicious activity, and anonymisation of data. Operationally, users must sign the terms and conditions of use for the eLab and health professionals are bound by employment contracts that stipulate their responsibilities.

1.5. Research Object Security

The eLab system aims to be as open as possible in terms of access to Research Objects with the goal of creating a collaborative research environment. Access to Research Objects within an eLab is controlled through a membership system. Research Objects have members, where a member is a user of the eLab. Additionally each Research Object in an eLab is classified as being either public or private. All users of an eLab
can view the contents of a public Research Object whereas the contents of a private Research Object can only be viewed by its members. Adding, removing and editing the content of a Research Object can only be done by its members. When a Research Object is initially created the user that created it is assigned as a member. Any member of a Research Object can invite another user of the eLab to become a member of that Research Object. This creates an invitation to become a member of that research object, which that user can choose to accept. A member of a Research Object can choose to give up that membership at any time. On doing so, an open invitation to rejoin that Research Object is left with that user. When a Research Object is classified as public a user can request an invitation to join as a member. This pushes a message to all existing members of that Research Object asking them to invite that user. Any member of a Research Object can switch its classification between public and private at any point in time.

2. Results

In this section we report on four representative deployments of the eLab framework and describe the application specific requirements and how they were satisfied.

2.1. GPLab

The primary requirement for the GPLab was to provide practice level information, such as the demographics, density of population and deprivation measures about the population served by the practice (Figure 3). Each practice is instantiated as a Research Object with its own dataset, but common analysis and visualisation encoded as a Method Object.

![Figure 3](image.png)

**Figure 3.** Screenshot of eLab showing construction of a function to calculate average cost per item of prescribing spend for a primary healthcare provider.
Maps are used as the primary visualisation technique, showing for example population density as in Figure 4. In the future the GPLab will be extended to provide benchmarking across the practices, and to analyse the health state of the population for Cardio Vascular Disease to assist in planning services targeted to need.

Figure 4. eLab screenshot showing geographic mapping of catchment area centred on primary health care provider.

2.2. Public Health eLab

The Greater Manchester Public Health (PH) Community adopted the eLab as their main communication platform for discussing problems, finding solutions and sharing expertise as a way to overcome the communication difficulties that will arise when public health analysts will move from Primary Care Trusts (PCTs) into Local Authorities. The PH eLab is accessible by Greater Manchester public health analysts, allowing them to ask and answer questions; and share project work. Currently, in the GM area there are 30-40 Public Health Analysts.

A requirements analysis with members of the Greater Manchester Public Health Network identified a core set of eLab requirements for their specific use. The PH eLab platform should provide the ability for PH analysts and members from the wider communities to:

- Login to a secure environment;
- Ask and respond to questions;
- Share work (including data, analysis and reports) with others; and
- Search the contents of the entire system.

The eLab core framework was extended to support the additional social networking features required by the PH Community, and consequently these features are available in all eLab deployments.
2.3. Bariatric Surgery eLab

The Bariatric Surgery eLab project was undertaken as a joint project alongside Salford Royal NHS Foundation Trust. The trust is one of two centres in the North-West of England that specialise in the provision of bariatric (weight loss) surgery. The aim of this eLab deployment was to collate the large amounts of raw data that had been collected by the trust who, however, lacked the resources to process, into a form that allowed for easy presentation and processing by members of the bariatric surgery team and the regional commissioners.

The eLab instance that co-ordinates this information uses a Method Object to take the raw quality of life questionnaire data and the raw co-morbidity information from the patient record and to transform it into a report that summarised the key information across the various areas being analysed. From the user's perspective this eLab acts simply as a report generation tool, and the complexity of the underlying research object server/workbench implementation is hidden from them. It is worth noting however that the full functionality of the system is still available within this eLab it is just hidden from standard users.

2.4. Long Term Conditions eLab

The Long Term Conditions eLab has the aim of producing a system that allows both individual GPs and GP practices along with area-based health analysts to monitor the treatments being given for an array of long-term conditions across groups of patients in the Greater Manchester area. The system centers around the use of a traffic-light scoring system which classifies the various treatments and drugs being given to patients on a three-tier, red, amber green scale measuring how well the treatments being given correspond to NHS assigned targets for these conditions.

The eLab designed to capture this information uses the approach that for each long-term condition to be monitored a single Research Object is created for each GP practice and a separate RO for analysts. This contains the data necessary to display a traffic light monitoring content item, developed specifically for this eLab deployment, and to contain tables of data representing the individual patient level data. In this case, because the data is 'live' and the research object aims to represent the current state of the world rather than a snapshot at single point in time, a specialised sub-class of data content item was created for this eLab which polls for new data and updates any related content items, such as the 'traffic light' content items, that are related to that data.

The presentation layer for the users of this eLab was specialised to display a dashboard view of datasets and traffic light content items on a single unified page. This 'dashboard' view represents a simplified view of the content of the eLab in that it hides a lot of options from the user and is presented as a 'read only' layer, but can be used to access the underlying advanced functionality of the eLab for users who wished to do so.

3. Discussion

The eLab provides a powerful and flexible framework for knowledge discovery in healthcare settings. By designing the eLab around privacy and confidentiality requirements, we have created a system that is fit for purpose and consequently this
reduces the risk to data owners of deploying an eLab. This has been a consistent theme across all eLab deployments to date. The RO mechanism provides the foundation for repeatable and shareable work. When combined with social networking capabilities, this adds a new dimension to the discovery and management of collaborative research.

In all settings, with the exclusion of academic research, exposing the full eLab to users overwhelmed them. They wanted an eLab pre-configured to meet their specific needs because they simply did not have the time to learn how to use the eLab. We were able to achieve this by leveraging the RO and MO capabilities. By pre-defining these for users we were able to meet their needs and did not need to undertake additional software development. The RO and MO can be thought of as the mechanism that enables eLab polymorphism. The user’s interaction with the eLab was simplified, as they could access the information they needed with a simple sequence of mouse clicks, however, within the eLab it is always possible to examine the MO or RO, if desired.

Building on the RO/MO professional social network paradigm, we envisage the next two evolutionary phases of e-Lab development as: 1) "finding and reuse" where ROs enable faster, more consistent replication of analyses across organisational boundaries; and 2) "radical sharing and assisted reasoning" where networks of investigators shape emergent findings across organisations via ROs. This evolution could support larger-scale, faster and lower cost production of knowledge about health risks and healthcare outcomes. The future of healthcare is inevitably more data-intensive, but requires smarter assembly of data, methods and people if it is to optimise its knowledge production.

References

COCPIT: A Tool for Integrated Care Pathway Variance Analysis

John AINSWORTH and Iain BUCHAN

The University of Manchester, UK

Abstract. Electronic Health Record (EHR) data has the potential to track patients’ journeys through healthcare systems. Many of those journeys are supposed to follow Integrated Care Pathways (ICPs) built on evidence based guidelines. An ICP for a particular condition sets out “what should happen”, whereas the EHR records “what did happen”. Variance analysis is the process by which the difference between expected and actual care is identified. By performing variance analysis over multiple patients, patterns of deviation from idealised care are revealed. The use of ICP variance analysis, however, is not as widespread as it could be in healthcare quality improvement processes – we argue that this is due to the difficulty of combining the required specialist knowledge and skills from different disciplines. COCPIT (Collaborative Online Care Pathway Investigation Tool) was developed to overcome this difficulty and provides clinicians and health service managers with a web-based tool for Care Pathway Variance Analysis.

Keywords. Integrated Care Pathways, Care Pathway Variance Analysis, Clinical Audit, Healthcare Quality Improvement

Introduction

Integrated Care Pathways (ICP) [1] are multi-disciplinary care plans that detail the diagnostic and treatment steps for patients presenting with a particular condition. They are used to aid clinical decision-making, as they effectively implement (national) clinical guidelines, with customisation possible at a local level to reflect service provision. They also help to ensure quality standards are met and to reduce variation in practice. In the UK Map of Medicine [2] and NICE Pathways [3] provide ICPs for many different conditions for use within the National Health Service (NHS). ICPs define a chronological sequence of steps, mostly commonly diagnostic or treatment, to be followed in providing care for a patient. An ICP may branch as a result of a diagnostic result and they may be cyclic where repetition of a sequence of steps is required for the maintenance of a health state. There will typically be many paths through an ICP, and each patient will follow one path. Flow charts are the predominant formalism for representing ICPs.

Where patients are cared for by a healthcare service that implements an ICP, the patient’s medical record will capture the details of the patient’s progression along or around the ICP. By comparing the patient’s medical record against the ICP, we can identify where the care given has diverged from the care expected. This is termed Care Pathway Variance Analysis (CPVA). CPVA can be calculated for individual patients.
and then aggregated according to the group of interest, such as hospital or clinician to provide a view of performance over a defined time frame. This may be further analysed by examining specific patient groups defined by, for example, gender, age, ethnicity or socio-economic status. The core method of CPVA can then be extended to a range of applications: clinical process audit; clinical outcome audit; inequality audit; identifying inefficiencies in provision of care across multiple morbidities; service redesign and economic evaluation of care pathways. Performing CPVA requires (i) the ICP to be defined in a computable form in terms of the clinical codes used in the corresponding medical records; (ii) electronic medical records accessible and available for use in a way that maintains patient privacy and confidentiality; (iii) computer codes for the analysis of CPVA and presentation of the results. For an organisation to perform CPVA considerable investment in time and people with specialist skills is required. Consequently, despite the considerable benefits that can be derived from CPVA, it has not been widely developed \[4\], \[5\].

In this paper we describe the functionality of a tool named COCPIT that begins to address the issues with CPVA identified above and report on some first experiences of using the tool. Our aim was to remove the barriers to performing CPVA, by developing an easy to use, web-based tool targeted at an end user community of clinicians and health service managers.

1. Methods

Throughout we used an agile, user-centred design methodology beginning with initial requirements capture using paper-based storyboards progressing through successive iterations of software development and review. A user reference group was established with representatives from primary care, secondary care, public health, service managers and commissioners to ensure a balance of views and requirements, and this group provided feedback on the user interface and prioritisation of tasks for the next iteration. COCPIT has three principle components: first, a data management framework, providing access to individual-level medical records whilst ensuring appropriate information governance; second, a visual editor for designing ICPs, which can then be stored; and third, an analysis and visualization component implementing the methods and techniques for users to perform CPVA and summarise the results for a range of applications.

1.1. Data Preparation

COCPIT was developed as a plugin tool for the eLab framework [6]. The eLab is a web-based software framework for data-intensive health research and knowledge generation. It provides secure access to health datasets and enforces information governance policies. Through the Research Object (RO) mechanism it provides the capability for users to record all the activity of their investigation such that it can be repeated or reused by others. For example, a RO can encapsulate an ICP definition that can then be discovered, copied, modified and employed by other users. The eLab content discovery mechanism employs metadata on ROs, combined with social networking techniques to actively suggest content for users. The eLab also provides data pre-processing tools that can reformat patient level data into a format that COCPIT analysis subsystem can understand. For example, transforming a journal table of
clinical events for multiple patients to a table containing a list of clinical events for each patient. COCPIT imposes minimal requirements on the input data, but each event should be time stamped so that it can be chronologically ordered for each patient. COCPIT is developed in the Microsoft Silverlight web application framework to provide the interactivity required, whilst remaining web-based.

1.2. ICP Representation

In order for ICPs to be computed and persisted, a suitable machine-readable representation was required. Having reviewed the plethora of guideline languages available [7], we concluded that there was no agreed standard and so we should define our own representation to be used internally to COCPIT, since the most important criteria at this stage was to be able to control and experiment with the definition language, such that it was tuned to our exact needs. Eventually, COCPIT could be extended to read and write other formats to allow reuse and exchange between CPOCPIT and other systems. We chose to represent ICPs as directed graphs, where each node represents a step on the pathway and the edges represent allowed transitions. Any valid path through the graph corresponds to a possible patient journey. Associated with each node is a statement that defines the event(s) that, if recorded in the patient’s medical record, mean that the node on the pathway has been entered. Events may be defined as clinical codes from a terminology (e.g. SNOMED-CT [8]) or may take on any user-defined value. This allows the ICP to be customised to the dataset over which it will operate. Complex combinations of events can be created using logical operators AND, OR, NOT and nested statements are permitted. Complex combinations are needed where there is variation in coding practice or where a terminology set is ambiguous. Expressions may include temporal relations between events, such that a particular event recurs a defined number of times in a defined period. Nodes may have validity conditions, which can determine whether a transition between a pair of nodes was expected or not. These are defined in terms of a conditional expression on an event(s) with an associated value, such that the value is equal, less than, greater than or falls within a range. A node may also have values associated with each exit that defines the proportion of patients that will follow the edge. Edges between nodes may have an associated transition time value that can define either the upper limit on the transition time between nodes or defines the expected transition time. Our initial implementation of the ICP representation used XML Schema, though this was subsequently changed to RDF as XML proved insufficiently flexible.

1.3. ICP Editor

The COCPIT ICP Editor presents the user with a visual canvas on which nodes can be added. For each node, a statement must be defined, and COCPIT provides a visual statement builder that simplifies the construction of complex statements. The statement builder utilises the eLab terminology service to lookup clinical codes from a search string that can be used to populate the event definitions. Edges are simply drawn between nodes on the canvas to build up the graph of the ICP. Transition time values can be optionally added to the edge definition. If a new exit edge is added to a node that already has one or more exit edges, then the exit condition editor is opened that allows the user to define the expression for each edge.
1.4. CPVA Processing

Having defined an ICP and selected a data source the user can now perform CPVA. The quality of real-world data is highly variable and this has consequences for data analysis. There will be a set of patients whose medical record does not completely match a path through an ICP. There could be many reasons for this, for example: (i) the event has not been recorded in the patient’s record; (ii) care professionals did not carry out one or more steps on the ICP; (iii) the patient refused treatment; (iv) the patient did not attend clinic; (v) the event was miscoded; or (vi) the ordering of events was reversed when recorded, which is particularly common when events occur in close proximity in time. COCPIT cannot determine why data is missing, but its matching algorithms must be able to cope with incomplete data to produce the best possible fit between observed and expected. Another implication is that patients may enter or exit a node in the ICP graph without following a defined edge when there is no record of entering a preceding or a succeeding node.

For a set of patient records and an ICP, COCPIT can currently perform the following analysis:

- Count of matches to each node
- Count of matches of unique patients to each node
- The point in time a patient matches to each node

From these values we can also then calculate values such as time between events, and summary statistics for number of node visits and duration.

The results can be presented in different ways:

1. Overlay the results on the ICP graph, by incrementing the count to for each node visited and each transition made. Display the value for each count on the node/edge as a proportion of the total population
2. Generate histograms and Kaplan-Meier (time to event) plots for the traversal times across each edge.
3. Full node statistics for patients entering, leaving, appearing, disappearing, proportion following branch vs. expected, and transition times.

COCPIT provides the capability to segment the result by population characteristics. Any combination of age (including user defined ranges), gender, ethnicity and deprivation of their local area can be selected to create a data series, so that comparisons between patient groups can be made. This can be used to investigate inequity in the provision of healthcare to sub-populations.

2. Results

The COCPIT tool was implemented over seven development iterations, with input from the user reference group at each stage. This has ensured that the software meets the needs of the end users. The ICP Editor has been used to implement ICPs for the Greater Manchester Stroke Service and for Chronic Kidney Disease (CKD) in the English NHS. The system has been tested with simulated patient data and real world patient data from the Salford Integrated Record [9] database for patients with CKD. The tool has been successfully run over a database of 100,000 unique patients with a total of 2.5 million events. COCPIT was applied to the survival analysis of stroke patients at Salford Royal Foundation Trust, UK. The aim was to verify that COCPIT...
would produce the same results as produce by a statistician using the R statistical analysis software package independently. As the input data set did not use a clinical coding system, the meta-data describing allowed values was extracted prior to loading into the eLab. There was one difference in the results, caused by an inconsistency in the way that R and Microsoft .net handled daylight savings time changes.

3. Discussion

A limitation of COCPIT is that it depends on high quality data to produce meaningful results. Further exploration of matching algorithms is required to improve the handling of missing data. However, this is a potential application for COCPIT, that was not previously envisaged: improving data quality and clinical coding of electronic health records. It is ideally suited to this task, as an ICP definition essentially defines what should be coded and how. Future development will provide the functionality in COCPIT to calculate the financial costs of variance from expected care, evidence that can be used to support the case for service redesign. COCPIT will also be extended to show how a service changes over time, supporting a “natural experiment” approach to continuous quality improvement, especially where multiple health systems compare and contrast their evolution using this tool. We envisage that COCPIT could be used to explore care quality at multiple levels: patient; clinician; clinical team; provider; health system/community; and profession. Eventual machine-to-machine communication of such quality metrics will have profound implications for managing health systems – the ethical and legal issues need to be examined.

4. Acknowledgements

Ainsworth and Buchan are partly funded by the NIHR CLAHRC for Greater Manchester.

References

Publication 7: Using string metrics to identify patient journeys through care pathways.
Using String Metrics to Identify Patient Journeys through Care Pathways

Richard Williams, BA1,2, Iain E. Buchan, MD, FACMI1,2, Mattia Prosperi, M.Eng., Ph.D1,
John Ainsworth, BSc, MSc1,2
1Centre for Health Informatics, 2Greater Manchester Primary Care Patient Safety
Translational Research Centre, University of Manchester, Manchester, UK.

Abstract

Given a computerized representation of a care pathway and an electronic record of a patient’s clinical journey, with potential omissions, insertions, discontinuities and reordering, we show that we can accurately match the journey to a particular route through the pathway by converting the problem into a string matching one. We discover that normalized string metrics lead to more unique pathway matches than non-normalized string metrics and should therefore be given preference when using these techniques.

Introduction

When faced with a patient’s electronic health record (EHR) and a prescribed care pathway it is useful to know if that patient’s care has deviated from the expected route through the pathway1. The degree of deviation from a pathway calculated with a distance metric, when combined with outcome data, could lead to the discovery of instances where the standard of care has been suboptimal leading to adverse outcomes, and also to instances of localized practice that lead to better outcomes.

However, before determining distance from a given route, we need accurately to determine which route through the pathway was traversed by the patient. This is a problem because routinely collected patient information is often poorly recorded with missing data, incorrect coding practice and data recorded out of sequence.

String metrics provide the distance between two strings and are usually based on algorithms for matching strings to patterns, with various degrees of approximation. They typically involve performing operations such as insertion, deletion and substitution. The string metric can be normalized2,3 or non-normalized4–6.

We attempt to discover the routes patients took through a care pathway by using string matching methods in a novel way with electronic health records from Salford, UK.

Related Work

Representing a care pathway in a format that can be readily interpreted by a computer is essential for analysis and also enables health information systems to provide decision support to health care professionals7. Computer-interpretable guidelines (CIGs) are computer representations of the clinical knowledge in a clinical guideline and are usually networks of tasks that occur over time8. A recent review of CIGs shows there is ongoing work on CIG modelling languages, their integration with EHRs, validation and verification of CIGs, compliance monitoring and sharing9. Most CIG modelling is based on Task-Network Models8,9 of which our graph-based approach is a general case.

There is also a large body of work on process mining10,11, frequent pattern mining, and the use of hidden Markov models for trajectory clustering12 for healthcare data, which has been reviewed by Lakshmanan et al.13 However, each of these techniques begin with the healthcare data and attempts to interpolate the pathways taken, whereas our approach differs by starting with a well-defined care pathway and attempts to discover the route taken.

Background

Care Pathways

Care pathways are structured guidelines for the assessment, diagnosis, and treatment of patients with a given condition1,14–16. They provide the ideal care that a patient should receive and are often represented as a flow chart1,14. In the UK, “NICE Pathways” (National Institute for Health and Care Excellence) offers pathways for over 150 conditions17.

More formally, a care pathway flow chart can be represented as a directed graph, \( G = (V, E) \), with \( V \) a set of nodes that represent clinical events such as diagnoses, measurements, procedures and treatments, and \( E \) a set of directed
edges that correspond to the permitted transitions between nodes. A transition can occur in a determined amount of
time. Figure 1 shows an example of a care pathway represented as a directed graph, defined a priori by experts.

**Figure 1:** A graphical model of a simplified, coded care pathway. Clinical codes in parentheses.

**SINAP**

The Stroke Improvement National Audit Programme (SINAP)\(^8\) is a data collection process for the purposes of clinical audit. It collects data about the care provided to stroke patients and includes several index events and the times they occurred. Here we examine data from Salford Royal Foundation Trust (SRFT) on 1078 patients with suspected strokes between 2010 and 2011. Figure 2 shows the approximate pathways that can be followed when a patient is admitted to hospital with a suspected stroke, covering the events recorded in the SINAP dataset. This is a simple pathway with only two decision points following when the patient is first seen and also after the patient has undergone brain imaging. The alphanumeric characters associated with each node in the pathway will be used later.

**Figure 2:** Stroke Improvement National Audit Programme (SINAP) pathway nodes as characters.

*Electronic Health Record*
A patient’s EHR is typically a list of coded events and states describing their care. In the UK a variety of coding schemes are used, such as Read Codes v2\textsuperscript{19}, CTV3\textsuperscript{19}, ICD-10\textsuperscript{20} and SNOMED\textsuperscript{21}. The processes described in this paper can be used with any coding system: here we use the SINAP dataset that employs custom codes.

**Method**

**Process**

We first assign an alphanumeric character to each node in the graph. By using the Unicode\textsuperscript{22} character set we can manage care pathways with up to 65,536 nodes. We then extract every possible route through the pathway as a string made up of the characters assigned to each node. For a graph \( G \) with \( n \) possible routes we construct the set \( R = \{ R_1, R_2, \ldots, R_n \} \), where each \( R_i \) is a string representing one of the \( n \) possible routes. For acyclic graphs such as the stroke pathway for the SINAP dataset this is straightforward via recursion. For a directed graph with cycles it is possible to repeat a cycle indefinitely so the number of possible routes is infinite. To avoid this we only allow each cycle to be repeated a finite number of times.

Due to the nature of our data, the events recorded are all covered by the pathway. In general, however, when using records from primary or secondary care, they may not be consistent with a care pathway event/transition graph. For a single patient we therefore extract all timed events from their record that occur on the pathway of interest, convert the events to characters, and concatenate the characters into strings according to their date-time order. The strings then represent the patient’s journey through the care pathway.

If our dataset contains patients with multiple interactions with the pathway, we must then distinguish between distinct interactions with the care pathway by specifying a cut-off time. If ever the gap between adjacent patient events is greater than the cut-off, then we assume that the patient has left the pathway and any subsequent events form part of the patient’s next visit to the pathway. This works well when the timescale of a pathway is shorter than the distances between them.

We then use the following string metrics to determine the distance between a patient pathway and each possible route through a care pathway.

**Longest Common Subsequence**

Formally, given two sequences \( A = a_1a_2 \cdots a_m \) and \( B = b_1b_2 \cdots b_n \) (\( m \leq n \)) we say that \( A \) is a subsequence of \( B \) if there are indices \( 0 < j_1 < j_2 < \cdots < j_m \leq n \) such that \( a_i = b_{j_i} \) is true for \( i = 1, 2, \ldots, m \).

Given two sequences \( X \) and \( Y, Z \) is a common subsequence if it is a subsequence of both \( X \) and \( Y \). \( Z \) is the longest common subsequence (LCS) if \( |Z| \geq |Z'| \) for all common subsequences \( Z' \), where \( |X| \) is the length of \( X \). The LCS is not necessarily unique.

We are interested in which route through the pathway a patient took so we need to decide on a distance metric to convert the LCS into something more meaningful. An initial algorithm for a single patient is as follows:

1. Create a list of all the possible routes \( R_1, \ldots, R_n \) through the care pathway
2. Filter the patient’s events to just include pathway events and apply the time cut-off time to give an event sequence \( E = E_1 \cdots E_m \)
3. For each route \( R_i \) calculate \( L_i = \text{LCS}(R_i, E) \)
4. If \( L_i > 0 \) calculate the distance \( d_i = \max(|R_i|, |E|) - |L_i| \)
5. Return the set of routes with the smallest distance

However, this only considers the discrepancy between the LCS and the pathway route; it doesn’t take into account the length of the LCS. We can normalize the distance by either dividing by the LCS, or by dividing by the combined length of the two strings and step 4 above becomes either:

4. If \( L_i > 0 \) calculate the distance \( d_i = \frac{\max(|R_i|, |E|) - |L_i|}{|L_i|} \)

or

4. If \( L_i > 0 \) calculate the distance \( d_i = \frac{\max(|R_i|, |E|) - |L_i|}{|R_i| + |E|} \)
We call these two methods LCS1 and LCS2 respectively.

**Simple Edit Distance (Levenshtein Distance)**

An alternative to the LCS is to consider the edit distance or Levenshtein distance. The edit distance between two strings $X$ and $Y$ is the minimum number of operations required to convert $X$ into $Y$ where an operation is either: insert a character, delete a character or replace a character. When switching is allowed ($ab \rightarrow ba$) the algorithm is the Damerau-Levenshtein. The costs of inserting, deleting and replacing are given as $W_I, W_D,$ and $W_R$ respectively. It holds that $W_R \leq W_D + W_I$, as we can always delete and then insert instead of substituting. By default the cost of each operation is 1.

The algorithm for our problem would be:

1. Create a list of all the possible routes $R_1, ..., R_n$ through the care pathway
2. Filter the patient’s events to just include pathway events and apply the time cut-off to give an event sequence $E = E_1 ... E_m$
3. For each route $R_i$ calculate the distance $d_i = LEV(R_i, E)$
4. Return the set of routes with the smallest distance

Similarly we can do this for the Damerau-Levenshtein distance which we will notate as $d_i = DAM(R_i, E)$.

**Levenshtein Variants**

Several versions of the Levenshtein Distance normalized to the length of the strings have been suggested. We notate the following as $NLEV^2$.

$$NLEV(X,Y) = \frac{LEV(X,Y)}{|X| + |Y|}$$

Also a normalized Levenshtein distance that satisfies the triangle equality and is therefore a true distance metric:

$$NLD(X,Y) = d_{N-GLD}(X,Y) = \frac{2 \cdot LEV(X,Y)}{\alpha \cdot (|X| + |Y|) + LEV(X,Y)}$$

where $\alpha$ is whichever cost is greater out of insertion and deletion. However, when $\alpha = 1$, as is the case when all the weights are set to 1 by default, although the distances produced by NLD and NLEV will differ, the ordering of the matches will always be the same.

Finally, we consider a normalized version of the Damerau Levenshtein distance.

$$NDAM(X,Y) = \frac{DAM(X,Y)}{|X| + |Y|}$$

We compare and contrast the different distance measures: LCS1, LCS2, LEV, DAM, NLEV, NLD and NDAM.

**Data cleaning**

Right censoring of the data is unlikely as once in hospital all end points are recorded. Most times in the data seem to be rounded to the nearest 10 or 15 minutes. This may potentially result in events appearing simultaneously or even out of order. There is also a risk of recollection or estimation bias as the data is often captured after the event. When events occur at the same time there are several options available. The patient can be ignored, but this would result in a lot of data being excluded from the analysis. An alternative would be to perform the analysis on the data ordered randomly and let the string matching methods correct any discrepancies. However as we are interested in discovering the actual path the patient took, we can assume where possible the events occurred in the correct order.

For two events A and B on a pathway there is either: a one-way path from A to B, a one-way path from B to A, a path from A to B and B to A, or it is impossible to get from one to the other. For a group of events occurring at the same time if it is possible to order them in a unique way then we choose that as the order of the events. If it is not possible, because of a cycle or an unreachable node, then we discard that patient. For datasets where this is commonplace it may be better to include the patients discarded here and randomise the order of the cotemporaneous events. Alternatively we could just discard the events rather than the patient.
Similarly, events of unknown time, or those with just a date and not a time, can be inserted at the correct point of a patient record, if possible, or discarded if contradictions arise.

**Data Management and Analysis Environment**

The SINAP dataset was transferred to us via an encrypted external hard drive in CSV format. This was then uploaded to a Microsoft SQL Server 2008 database for analysis. Sequence matching was performed with C#.NET and all statistical analysis was done using R. The sm library was used for plotting density curves and the pROC package was used for comparing Receiver Operating Characteristic (ROC) curves.

**Results**

**Data Characteristics**

The SINAP dataset contains 1078 patients of which 549 are female and 529 are male.

Table 1 shows the number of records that were cleaned using the above data cleaning process. Only 1 patient’s route could not be uniquely re-ordered.

**Table 1.** Data cleaning results

<table>
<thead>
<tr>
<th>Total patients</th>
<th>1078</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midnight events – able to insert</td>
<td>424</td>
</tr>
<tr>
<td>Simultaneous events – able to order</td>
<td>3</td>
</tr>
<tr>
<td>Midnight and simultaneous events – able to order</td>
<td>648</td>
</tr>
<tr>
<td>No midnight or simultaneous events – no need to order</td>
<td>2</td>
</tr>
<tr>
<td>Midnight events – unable to insert</td>
<td>1</td>
</tr>
</tbody>
</table>

There are 46 distinct pathways taken by the 1077 patients following time reordering. Table 2 shows the frequency of the top 10 patient pathways. The pathways that match the ICP are in bold. The route of GHDB should be a valid route however there are no patients in our cohort who followed this – suggesting this is not a valid route and the care pathway could be altered.

**Table 2.** Top 10 pathways – character sequences from figure 2.

<table>
<thead>
<tr>
<th>Patient Record</th>
<th>Count</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHDEFIB</td>
<td>275 (26%)</td>
<td>Valid route</td>
</tr>
<tr>
<td>GHDFIB</td>
<td>275 (26%)</td>
<td>Valid route</td>
</tr>
<tr>
<td>GHDFEIB</td>
<td>122 (11%)</td>
<td>Valid route with E/F switched – lots of people so maybe a valid route.</td>
</tr>
<tr>
<td>GDHEFIB</td>
<td>63 (6%)</td>
<td>Valid route with D/H switched – can’t be seen before you arrive.</td>
</tr>
<tr>
<td>GDHFIB</td>
<td>60 (6%)</td>
<td>Valid route with D/H switched – as above.</td>
</tr>
<tr>
<td><strong>GHDEAFCIB</strong></td>
<td>56 (5%)</td>
<td><strong>Valid route</strong></td>
</tr>
<tr>
<td>GHEDFIB</td>
<td>39 (4%)</td>
<td>Valid route with E/D switched – can’t be imaged before first seen.</td>
</tr>
<tr>
<td>GHDEFACIB</td>
<td>37 (3%)</td>
<td>Valid with A/F switched.</td>
</tr>
<tr>
<td>GHFDIB</td>
<td>24 (2%)</td>
<td>D/F switched – can’t arrive in specialist bed before being seen.</td>
</tr>
<tr>
<td>GDHFEB</td>
<td>24 (2%)</td>
<td>D/H and E/F switched</td>
</tr>
</tbody>
</table>

It appears that there are some valid routes that aren’t in our pathway. For those who don’t get thrombolysed there are many people who arrive in a specialist stroke bed prior to their brain scan. Also there are many people who get “First Seen” before they arrive at the hospital. This seems nonsensical but could be valid if “First Seen” applied to GPs or ambulance staff. Finally there are patients who receive thrombolysis after getting to a specialist stroke bed which could also be a valid route. All other switches appear to be mistakes – for example having a brain scan prior to being first seen.

In order to determine how well each method works we must determine for each patient the most probable route taken. As our dataset is small we can do this manually by defining rules based on the data. We first assume that events that don’t happen are rarely inserted and then classify the patients according to the following rules:

1. If a patient has thrombolysis or a follow up scan then assumes route GHDEAFCIB
2. Of those remaining, for any with a brain scan we assume route GHDEFIB
3. Of those remaining, for any with a stroke unit arrival or discharge we assume route GHDFIB
4. Of those remaining we assume GHDB

In addition to returning the correct result it is also of use if the distance measure returns a unique result. There will be situations where this isn’t possible but in general string matching methods that return more unique results are preferable.

For each method, Table 3 gives the number of unique matches and the number of correct matches where a correct match is one that is both unique and matches with the routes we assume the patients actually followed.

**Table 3. Number of unique and correct matches**

<table>
<thead>
<tr>
<th>Method</th>
<th>Unique Matches</th>
<th>Correct Matches</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEV</td>
<td>818 (75.95%)</td>
<td>645 (78.85%)</td>
<td>59.89%</td>
</tr>
<tr>
<td>DAM</td>
<td>853 (79.20%)</td>
<td>849 (99.53%)</td>
<td>78.83%</td>
</tr>
<tr>
<td>LCS1</td>
<td>882 (81.89%)</td>
<td>878 (99.55%)</td>
<td>81.52%</td>
</tr>
<tr>
<td>LCS2</td>
<td>1077 (100.00%)</td>
<td>1070 (99.35%)</td>
<td>99.35%</td>
</tr>
<tr>
<td>NLEV</td>
<td>1076 (99.91%)</td>
<td>841 (78.16%)</td>
<td>78.09%</td>
</tr>
<tr>
<td>NLD</td>
<td>1076 (99.91%)</td>
<td>841 (78.16%)</td>
<td>78.09%</td>
</tr>
<tr>
<td>NDAM</td>
<td>1076 (99.91%)</td>
<td>1068 (99.26%)</td>
<td>99.16%</td>
</tr>
</tbody>
</table>

The NLEV and NLD methods produce the same results as predicted. The ratio of correct matches to unique matches shows that the Damerau-Levenshtein and the longest common subsequence methods work excellently with >99% correct, whereas the Levenshtein variants only achieve 78-79%. It can also be seen that normalized methods are better at producing unique matches with LCS2 matching all pathways uniquely, while NLEV, NLD and NDAM only fail to give a unique answer for a single patient - actually a different patient for each method. Examining the difference between NLEV and NDAM shows that NDAM is correctly identifying pathways where events have been recorded out of sequence. As an example the patient record of GHDFEIB is correctly matched to GHDEFIB by NDAM, while NLEV matches it to GHDFIB.

When the values for unique correct matches are combined the normalized Damerau-Levenshtein and the second Longest Common Subsequence methods are best, correctly matching >99% of the patient pathways.

For these two methods we can split the pathways into two groups: correct and incorrect matches, where a correct match is when the algorithm uniquely identifies the route the patient traversed through the pathway. We then compare the groups under the null hypothesis that the mean ‘string’ distance between them is equal. The density plots in Figure 3 demonstrate the data we want to contrast are not drawn from normal or symmetrical distributions, indeed the distributions of string distances are quite different for matches compared with non-matches. Thus we make the contrast with a non-parametric (Matt-Whitney) method, demonstrating statistically highly significant differences for both NDAM (P < 0.0001) and LCS2 (P < 0.0001) metrics.
Figure 3: Box (top) and density (bottom) plots of string distances for matches (YES) and non-matches (NO) for NDAM and LCS2 metrics.

Finally, we compare NDAM, LCS2 and NLEV string distance metrics with regard to their classification accuracy for our care pathway journeys. Figure 4 shows the ROC curves for each metric with our test dataset, and the 95% confidence intervals for the areas under the curves: the more detailed comparison of the two most accurate metrics (NDAM and LCS2) is the Mann-Whitney result above.
Figure 4: Receiver Operating Characteristic (ROC) curves for NDAM, NLEV and LCS2 string distance metrics with 95% confidence intervals for the areas under the curves.

Discussion

Distance Weighting

The operations in the Damerau-Levenshtein string metric can be weighted. Given the nature of our dataset it is more likely that records were omitted or out of order, than miscoded. If we are sure of this we can change the weighting of the operations accordingly – an option that is possible with the NDAM and not the LCS2 method. By doubling the weight associated with deleting a character, therefore making it less likely that matches will feature deletions, of the 1077 patients we yield 1077 unique matches of which 1074 are correct. Weighted NDAM then becomes the most accurate way of predicting a patient’s route.

Generalization

The string matching process described here operates on a graph based representation of a care pathway. Therefore the methodology is theoretically applicable, although untested, to any process or workflow that can be represented as a graph, in healthcare and beyond.

Future work

There are several factors unstudied in this paper that will affect the overall success of the method. The size and shape of the graph is a factor, as is the quality of the data. Further work is needed to determine which graph shapes work well with this method. Finally, the next stage of our work is to determine how the distance a patient is from their care pathway predicts their outcomes.

Conclusion

String matching would seem to be a highly successful way to determine which route a patient followed in a care pathway. Normalized distance functions should be used to ensure high numbers of unique matches. For clinical data where the chance of events occurring, or being recorded, in the wrong order is high, the Damerau-Levenshtein or Longest Common Subsequence methods should be used in preference to the Levenshtein distance.
Acknowledgements

Funded by the National Institute for Health Research Greater Manchester Primary Care Patient Safety Translational Research Centre (NIHR GM PSTRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References


Publication 8: Combining Health Data Uses to Ignite Health System Learning.
Combining Health Data Uses to Ignite Health System Learning

J. Ainsworth¹; I. Buchan²

¹Health eResearch Centre, Farr Institute for Health Informatics Research, University of Manchester, UK
²Centre for Health Informatics, Institute of Population Health, University of Manchester, UK

Keywords
Health data reuse; secondary uses; meaningful use; learning health systems; adaptive health systems, intelligence pipelines; health-care evidence; population health.

Summary
Objectives: In this paper we aim to characterise the critical mass of linked data, methods and expertise required for health systems to adapt to the needs of the populations they serve – more recently known as learning health systems. The objectives are to: 1) identify opportunities to combine separate uses of common data sources in order to reduce duplication of data processing and improve information quality; 2) identify challenges in scaling-up the reuse of health data sufficiently to support health system learning.

Methods: The challenges and opportunities were identified through a series of e-health stakeholder consultations and workshops in Northern England from 2011 to 2014. From 2013 the concepts presented here have been refined through feedback to collaborators, including patient/citizen representatives, in a regional health informatics research network (www.herc.ac.uk).

Results: Health systems typically have separate information pipelines for: 1) commissioning services; 2) auditing service performance; 3) managing finances; 4) monitoring public health; and 5) research. These pipelines share common data sources but usually duplicate data extraction, aggregation, cleaning/preparation and analytics. Suboptimal analyses may be performed due to a lack of expertise, which may exist elsewhere in the health system but is fully committed to a different pipeline. Contextual knowledge that is essential for proper data analysis and interpretation may be needed in one pipeline but accessible only in another. The lack of capable health and care intelligence systems for populations can be attributed to a legacy of three flawed assumptions: 1) universality: the generalisability of evidence across populations; 2) time-invariance: the stability of evidence over time; and 3) reducibility: the reduction of evidence into specialised sub-systems that may be recombined.

Conclusions: We conceptualize a population health and care intelligence system capable of supporting health system learning and we put forward a set of maturity tests of progression toward such a system. A factor common to each test is data-action latency; a mature system spawns timely actions proportionate to the information that can be derived from the data, and in doing so creates meaningful measurement about system learning. We illustrate, using future scenarios, some major opportunities to improve health systems by exchanging conventional intelligence pipelines for networked critical masses of data, methods and expertise that minimise data-action latency and ignite system-learning.

1. Introduction

The imperative for health systems to deliver better care, for more people, from fewer resources is stronger than ever. The most recent worldwide economic crisis shrank the resources available for healthcare. [1, 2] The growth in demand for care-services, however, continues unabated, propelled by the twin engines of need and expectation: the needs of ageing populations living longer with naturally more chronic diseases and the persistence of unnatural risks such as diabetes through obesity; and the expectation of more care from the growth in healthcare technologies and historical investments in services. [3, 4] The promise of “personalised medicine” in particular raises demands for ever more specialized and costly diagnostics, therapeutics, devices and procedures. [5, 6] As patients, we rightly expect our health systems to be safe and effective. As consumers, we rightly expect those services to be accessible and acceptable. And as citizens, we rightly expect public resources for healthcare to be organized to achieve the maximum gain for the maximum number.

So, does the data flowing through health systems work to optimise care for patients, consumers and citizens? Here we address this question in general and illustrate it using the English National Health Service (NHS) – the NHS is one of the largest and longest running universal, population-based healthcare systems, where the use of personal health data for the public good is mandated by law. [7]

We use the term “reuse” to refer to any uses of individual health data outside the information system or subsystem in which the data were collected, for example total serum cholesterol levels used to inform planning for the provision of cholesterol lowering drugs (statins) in the health sys-
2. Objectives

In this article we will show by way of example that current health data reuse is too little, too late, and at too high a cost of information management. [14] By contrast we examine how innovations in data reuse have transformed social and economic life over the past two decades, exploring how these paradigms may translate to the design and development of health systems and what are the prerequisite conditions for system-learning. We argue that for health systems to operate efficiently they must adapt to the needs of the populations they serve, and that they can only do this if pipelines of evidence are established that allow bidirectional exchange between research and practice. We describe how these pipelines can be implemented and used in real health systems, drawing on existing research work. Finally, we posit tests of health system learning and suggest a maturity model to encourage progress toward this challenging goal.

3. Methods

The conceptual synthesis presented here is based on multiple literature reviews and stakeholder consultations in shaping a regional health informatics research strategy over ten years. From 2004–12, the North-west Institute of Bio-health Informatics (for Northwest England from a base at The University of Manchester) ran monthly research meetings focused on better digital phenotyping with healthcare data, resulting in “e-lab” and “research object” approaches to using health data more systematically. [10, 15] In 2013 this work expanded to a national colloquium as The Health e-Research Centre (www.herc.ac.uk) of the UK’s Farr Institute for Health Informatics Research (www.farrinstitute.org). The heuristic development of the health data reuse concepts presented herein have incorporated input from patient and public representatives, and can be seen at www.herc.ac.uk/peppi.

4. Results

4.1 Legacy Assumptions that Impede System Learning

We identified three assumptions of health systems and healthcare delivery that work against the adaptation and optimisation of health systems necessary for them to meet the needs of the communities they serve. These assumptions are out-dated in the face of modern data, methods and technologies, yet they pervade:

4.1.1 Time Invariance: Assuming that Evidence is Eternal

The Translational Medicine movement has delivered less than expected, partly due to a flawed assumption that once produced, evidence can be translated into practice ad infinitum. [16] We illustrate this flaw with the example of predictive modelling, which is used in healthcare to provide evidence to support a wide range of decisions, from individual patient risk scoring to planning long-term investment in the provision of services. These models are produced and validated at a particular point in time. The model developers are typically rewarded with a publication and don’t revisit the research for a long time. For example, the EuroSCORE model, which is used to predict the risk of death among patients undergoing coronary artery surgery or stenting, was developed in the late 90s, then it drifted so much in calibration that is now predicts more than double the observed mortality. [17, 18] This highlights a general problem that such models are treated as static when there is a need for dynamic alternatives that can be continuously recalibrated. The evidence production line is broken, not only its translation into practice.

4.1.2 Isolation: Assuming that Specialisation is Scalable

The organisation of health systems into specialities for the treatment of disease reflects the isolation assumption. Typically a patient with cardiovascular disease (CVD) may see a cardiologist, diabetologist, renal physician or stroke specialist, each working in separate organisational units of a hospital, served by separate clinical guidelines.
from separate areas of clinical research. These specialists deal with different facets of CVD. Generalist clinicians may try to knit the patient’s management together across specialty areas, but the evidence pipelines are usually separate. We are not arguing for healthcare provider organisational structures to change, it simply would not be possible for an individual physician to be an expert in all these areas. However we do contend that this assumption of isolation means that evidence is not easily combined to guide the management of co-morbidities. The coproduction of health between patient and physicians can enable holistic patient care resulting in better outcomes for patients and optimised service provision. [19]

4.1.3 Homogeneity: Assuming ‘One Size of Evidence Fits All’

The evidence base that underpins current health systems focuses on average treatment effects in the major diagnostic categories. Healthcare planning and care quality management therefore assume that ‘one size of evidence fits all’. Considerable variation in care outcomes, however, may be due to unrecognised subgroups (or endotypes) being wrongly treated as ‘the average patient’. Many common chronic diseases such as asthma are actually a collection of more specific diseases, which may have different responses to treatments. [20] Basic biomedical research, for example genomics, can unpick some of this stratification in the quest for “personalised medicine”. Other factors, however, may have greater influence on outcomes: for example, the social support available to a patient on discharge from hospital.

The same applies at the population level, where a national policy that is easily implemented in one community may be impractical to implement in another. So, there is a need for health information systems that can ‘phenotype’ patients and communities more deeply, enabling healthcare resources to be titrated more specifically to needs.

4.2 Barriers to the Data Uses Needed for Health Systems to Learn

The assumptions identified in the previous section are not necessary when healthcare data are reused to characterise the population that a health system serves. However, there exist barriers to healthcare data reuse that must be overcome in order to realise: dynamic, scalable and contextualised evidence. These barriers will not necessarily be the same in all health systems, communities or countries. Fragmented and disconnected data, fear of disclosure and the shortage of skilled human resources are commonplace across many health systems and are examined in this section.

4.2.1 Fragmentation and Disconnection of Data

We class current secondary (not for direct care) uses of clinical data as fragmented and/or detached:

- **Fragmented** repositories contain a subset of care records, for example a disease-specific register, such as a cancer registry. Fragmented repositories may have been created by extracting information from care records and/or by explicit data entry. The data contained in the repository is a partial view of the patient’s care and so its utility is limited.

- **Detached** repositories are created from electronic health records, but they are managed and governed by an agency not directly involved in the care of the patient/citizen. The removal of the data from the community creates a number of problems:
  - i. The data are under a different ethical and governance framework, therefore their uses are beyond the day-to-day control or sight of the contributing community.
  - ii. The metadata, typically contained in the tacit knowledge of people in the health system, are lost, for example the knowledge of the clinical biochemist that the laboratory assay for creatinine changed at a certain time point, after which a different formula for estimating kidney function needs to be used. [21]
  - iii. Structured data detached from unstructured data have lower value, for example the research value of clinical codes can be increased by allied mining of clinical narrative or rubrics. [22, 23]
  - iv. Once a detached data repository is created, there is no way for a patient to dissent (withdraw their consent and have their data removed).
  - v. There is no feedback loop between the analysis of the data and the provision of health services in the communities from which the data are derived.
  - vi. Anonymisation removes the capacity to link individual records across multiple data sources.

4.2.2 Fear of Disclosure of Person Identity

The prevailing attitude in healthcare is to minimise disclosure risk. [24] This has a direct and significant impact on health data reuse – it stifles it. The problem is that disclosure risk is considered separately from the risk management of information supply and quality. A holistic risk assessment of the healthcare system’s operations is required. The technical and operational procedures for mitigating disclosure risk and well researched and widely available and so are available for organisations to apply. [25] Assessing disclosure risk in isolation results in under-information that in turn reduces quality and efficiency. It is essential to assess disclosure, quality and efficiency together.

4.2.3 Informatics Workforce: Capacity and Capability Gaps

Even when high quality data are made available for reuse health systems face a further problem. Where are the people to prepare, analyse and interpret these complex data? The nascent health informatics profession is under capacity and there are few education and training programmes that can deliver skilled professionals able to prepare low quality data, perform an analysis and interpret the results. [26–28] Effective reuse of healthcare data requires an understanding of how the data were created, to enable effective interpretation. This includes understanding the organisational processes in which the data were captured, for example was the patient’s weight recorded because they seem overweight or as part of a systematic review? Was the clinical code assigned by a clinician during the clinical encounter or by a coding clerk? Have records been searched for missing,
remunerated diagnoses by inferring missing diagnoses from medication lists? Increasing the denominator size will not compensate for unexplained heterogeneity in datasets when trying to make robust inferences with biostatistical and epidemiological methods. [29, 30] Health systems need a workforce that can combine informatics, statistical and (clinical) epidemiological understanding on demand.

4.3 Technologies for Innovation in Health Data Reuse

Over the past three decades the information revolution has fundamentally changed economies and society. This revolution has been enabled by information technology. The Internet provides a global data communications infrastructure that enables us to exchange information with almost anyone without restriction. The World Wide Web provides the means to connect and search for information. More recently, social media have removed the boundaries between producers and consumers of information. Industry and commerce have exploited this infrastructure in activities that are now core to most economies. [31] This rapid growth has been fuelled by innovative ways of using the vast quantities of connected data at the level of the individual that has been an emergent feature of this information revolution. This innovation has been characterised by four paradigms: the Network Effect; [32] Crowd Sourcing; [33] the Long Tail; [34] and (Big) Data Analytics. [35]

i. The Network Effect is most visible in social networking sites whereby the value of the products and services increases in proportion to the number of users. On a social networking site where data on individuals preferences and behaviour are routinely provided by users either directly or indirectly, this can be utilised by the collectors of the data to target typically the marketing of products and services. For example, the matching of one consumer with another to provide recommendations of products based on past behaviour.

ii. The Long Tail in retail refers to the paradigm of “selling less of more” [36]; what this means in practice is serving a wider more diverse consumer base with the products they want. The analogous situation in healthcare would be treating less of more. What this means in practice is being able to identify individual needs and being able to provide timely, targeted care. It can also be seen as another expression of Rose’s primacy of sick populations over sick individuals. [37]

iii. Crowd Sourcing is where motivated individuals contribute on-line to a call for participation in tasks, such as raising capital, data collection, [33] and problem solving.

iv. Data Analytics, often referred to as “Big Data”, refers to the techniques of discovering hidden structure in data that can then be used to model individual behaviours and make predictions based on those models. The techniques include cluster analysis, graph databases and machine learning. It is widely used in social networking, [38] recommendation systems, [39] and search engine advertising. [40]

Other industries/sectors have gone through a period of rapid innovation centred on using the information available to them to identify and meet the needs of their consumers/customers. Healthcare needs to learn these lessons. The three paradigms above could help healthcare to scale up, titrating its scarce resources to the needs of better characterised populations. PatientsLikeMe (www.patientslikeme.com) provides a compelling example, leveraging each of the three paradigms, and even stimulating new research to address weaknesses in the medical evidence base. [41]

The current technology momentum is centred on ubiquitous computing and pervasive connectivity, now widely available via smartphones and an explosion of personal monitoring and tracking devices – the impacts are likely to be profound. [42] The Quantified Self movement reflects the self-motivation to track and monitor personal wellbeing, which is moving from a fitness niche into the mainstream. [43, 44]

We are on the cusp of a data deluge, where individual data from bio-medical sensors, GPS, and accelerometers can be combined to with population and environmental data to provide a rich longitudinal picture of our health.

Self-quantification is laying cultural and technical foundations for self-experimentation, which can translate to routinely personalised healthcare employing ‘n-of-1’ trials. [45] Health systems need to prepare to move beyond the prescription pad, to delivering packages of care that are tuned to patient experiences and outcomes reported between clinical encounters. Are we about to encounter the health data paradox – struggling to make sense of limited clinical data to optimize health systems when self-tracking, personal health monitoring and ubiquitous computing and connectivity are about to provide a deluge of high resolution data? Health systems need to embrace the concept of coproduction, [19] where data generated by patients are used to optimize care. This requires the interaction of patients, clinicians and machine algorithms to extract the signal from the data and to transform it into knowledge for shared decision-making, as shown in Figure 1.

Thus health information systems are set to move from separate requirements for machine-clinician and machine-patient interaction to the triangle of digital health illustrated in Figure 1. Communication within this triangle will need the union of records and models that represent an individual’s health to be conveyed with context-aware ‘personalities’, which is more avatar than filing cabinet (the stale paradigm of medical records). [46, 47]

For health systems, the interaction of individuals over shared health information is critical. For example citizens influencing each other to be more physically active and reducing the burden of type 2 diabetes. And in the clinical context, patients, carers and practitioners being armed with early warning of conditions deteriorating in community settings, able to take actions that avoid hospital admissions. So future health information systems must be able to support networked conversations over ‘health avatars’. The evolving concept of social machines holds promise that this will be tractable in the next decade. [48]
4.4 Architectures for Health System Learning

4.4.1 Bidirectional Evidence Pipelines

The assumptions of time invariance, isolation, and homogeneity identified above, can be eliminated by ensuring that timely evidence is available at the point of need, by creating new pipelines of evidence that connect the generators of data with the consumers of the data. We propose a conceptual framework for pipelines of evidence as illustrate in Figure 2. Each evidence pipeline essentially represents a bond of trust between the generator of the data and their consumer. A pipeline has a custodian who is the representative of the patient/citizen and has oversight of the establishment, and operation of the pipeline. A pipeline is licensed to operate for a defined period of time, and at the end of the licence period the custodian will have the power to revoke or renew the licence. A large healthcare organisation such as a hospital may manage a data node that has all of these roles, whereas a small information intermediary might have just one. One organisation might act on behalf of others in respect of a role, for example transforming data to maintain a mortality risk model for percutaneous coronary intervention, or consuming data to lead on a suprard district audit of the same intervention.

Each pipeline may have zero or more transformers, the role of which is to process the raw data to produce information or knowledge. The consumer may use the resulting evidence for the purpose defined by the custodian. The pipeline of evidence conceptual framework ensures that the interests of all parties are represented, the reasons for and the usage of the pipeline is well defined and that governance rests with the representative of the patient. Wherever possible the pipeline should incorporate a feedback loop delivering information and/or knowledge derived from the data. This feedback loop enables the generator to adapt and optimise.

To illustrate this with the EuroSCORE example given above: the data generators are cardiac surgeons and cardiologists completing audit forms; the custodian is the National Institute for Cardiac Outcomes Research in the UK; the transformers are researchers modelling the risk of patients dying during, or within 30 days of, procedures to unblock coronary arteries; and the consumers are providers and payers monitoring the quality of services between provider organisations and individual practitioners. Currently, there is no systematic feedback, which means that poor quality data continue to be collected, including a binary flag for renal function instead of a more informative creatinine value. Systematic feedback to the practitioners whose performance is being measured may improve data quality, and feedback to researchers on the calibration drift of their published model may trigger timely recalibration.

Under this framework health systems would transition from the current unidirectional model of ‘evidence into practice’ to a bidirectional model of evidence pipelines with feedback.

4.4.2 Distributed Sense Making

In order to eliminate the barriers to health data reuse described above we developed the population health e-Lab model. [10, 49] Under this model, communities may integrate pipelines of evidence from primary care, secondary care, specialist care sources such as disease registries, and administrative sources such deaths and demo-
Fig. 3  Connected Health City model for health system learning

graphic data – this would be natural for the formation of an integrated health intelligence system for that community. Record linkage, across multiple pipelines, is performed for the purposes of: informing individual patient care; managing the quality of local care services; commissioning/planning care services; public health intelligence; and research. This linkage takes place within the secure network of the local healthcare agencies before removing data that would unnecessarily identify individuals to analysts. Deep analysis of the data can then be performed within the same information governance environment used to protect identifiable records. Thus the risk of deductive disclosure of individual identities is managed.

In the UK, laws permit large-scale linkage of health data under an opt-out consent model for multiple purposes, including research. [7] This has been difficult to apply across large populations such as England’s 53m but it is happening at the level of natural health systems (populations of around 1–6m and sub-districts thereof). [50] Such systems map to geographically-defined populations with cohesive cultural identities such as small countries, metropolitan districts and civic regions. Civic-centred e-Labs for handling large-scale health data afford tangible public oversight of data uses. So research with opt-out consented data has a different ethical context to more remote uses of the data, with greater clarity over autonomy, beneficence, non-maleficence and justice in determining data uses. [51]

There are two important situations where different health systems may need to borrow analytic strength from each other without sharing potentially disclosive data: 1) where an expert analyst is external to custodian of the data; and 2) where research findings need to be replicated across heterogeneous populations/settings. Here we have developed the Research Object (or Work Object [10]) model whereby computable representations of a research protocol can travel between e-Labs gathering results while leaving potentially disclosive data within the custodian’s firewall.

Coordinated research across a global network of population health e-Labs could effectively eliminate the problem of non-reproducible research that plagues the medical literature. [29, 52]

4.4.3 Connected Decision-making

So when does a health system have enough connectedness of data, methods and expertise to start learning in computable ways? In addition to bidirectional evidence pipelines and distributed sense-making we consider the connection between evidence and person-accountable decision-making to be the other missing link for igniting health system learning. The evidence base on which most healthcare decisions are taken is currently not recorded. This applies to population/system level planning decisions, clinical and self-care decisions alike.

Recognising this problem, the North of England (population 15m) is building four pilot “Connected Health Cities” that will use care pathway optimisation as the organising principle for seeding learning health systems. In this model, a critical mass of data, methods/tools and expertise is brought together in one physical and digital location under strict healthcare information governance. We call the location an Ark to reflect the dual metaphor of a covenant of trust with the public over the uses of their personal data, and protection from the data flood that might otherwise impeded system learning. The information/action flows for targeting care along pathways is illustrated in Fig. 3.

The Ark is an extension of the e-Lab model to defined populations and comput-
able information/action loops. It works not only for the population served but also offers external researchers and enterprises ways to interact with the Ark’s large-scale linked data in ethical, efficient ways. For example, where a company needs clinical prediction algorithms to be developed/calibrated they can ‘spin-in’ to the Ark using the object model described above and be validated in a trusted-third-party manner, employing local contextual knowledge that would not exist in the company. Similarly, a network of Arks using similar operating procedures can provide replication facilities for research or algorithm development across heterogeneous populations/settings.

The creation of the Ark critical mass should also save the health system money by reducing the current duplication of data processing where disparate teams of analysts run separate data warehouses and overlapping analysis pipelines as shown in Figure 4.

5. Discussion

The transition from data-sparse to data-intensive healthcare cannot be achieved through technologies and system management alone, as there is a major culture change involved. We note a train of relevant culture change in healthcare from the introduction of integrated care pathways in the 90s, [53] through health systems integration in the 00s to current aspirations of personalised and precision medicine. [54] These are not fashions they are phases in moving from the expected information core being the clinician to it being the health system. [47]

Clinician performance is already measured using quality standards derived from the clinical practice guidelines that underpin care pathways. In most parts of the world these metrics emerged in the 80s and 90s from clinical audit/governance/quality management initiatives. The ideal was the “close the audit loop”, which is in effect a ‘learning health sub-system’. The challenge of quality management is greater in primary care than more specialist areas: it covers more clinical processes & outcomes so is more data-hungry. The UK was able to establish a Quality and Outcomes Framework on the back of its established capture of structured clinical data by general practitioners. [55] Work has undertaken to integrate this ‘quality conversation’ between primary and secondary care by analysing linked care records to follow patient journeys along care pathways between provider organisations, comparing observed with guideline-suggested care, as illustrated in Figure 5. [56, 57] It is our belief that cross-provider pathway optimisation is a potential ignition point for system-wide learning, particularly when coupled with service planning/commissioning.

We posit data-action latency (DAL) as the key metric of health system learning: this is the time lag between data being

© Schattauer 2015
available that an action should be taken and the action being taken. Minimising DAL requires concerted data capture, aggregation and analysis followed swiftly by interpretation of results, assignment of responsibility for any actions, and recording of actions. Bidirectional evidence pipelines will enable the measurement of DAL.

6. Conclusions

Current health data reuse is too little, too late, at too high a cost of information management. Pressures on health services from obesity, diabetes, and aging populations require more to be done with less. There is a pressing need for health systems that: can adapt to the needs of the individuals and population they serve; and can optimise the usage of resources to provide the maximum health gain to the maximum number. We have described three legacy assumptions (structural, organisational and technical) of health systems and argued the case that they prevent health systems from delivering optimal care, but can be fixed by building pipelines of evidence that enable health systems to adapt to their communities, through the integration of multi-source healthcare data. Our analysis of existing examples of data re-use, reveals some fundamental issues that need to be resolved. As health informaticians we have much to learn from the socioeconomic information revolution and the way innovation has been driven by the Long Tail, Crowd Sourcing, the Network Effect and (Big) Data Analytics. Similarly we must fully grasp the opportunity that ubiquitous computing and pervasive connectivity will provide to understand individual patients in unprecedented detail. We will all increasingly be producers and consumers of health information and need to establish pipelines of evidence between patients, algorithms and clinicians to enable the co-production of care.

Acknowledgments

This work was supported by funding from the MRC Health eResearch Centre grant MR/K006665/1 and NIHR Greater Manchester Primary Care Translational Research Centre grant.

References


Appendix B - Health Informatics Design Patterns

Term Selection from Clinical Code Sets

When working with data users often need to choose terms from a clinical code set. These sets tend to be very large (often within excess of 100000 terms), too large to be known fully by end users. In fact end users often do not know about clinical terms at all, but they understand diseases, medication, operation, procedures and tests. Clinical code sets are often organised as taxonomies and may also permit synonyms. The problem is how to help users find the clinical terms they need. The solution is a faceted search engine with partial keyword look-ahead matching, and hierarchical term expansion. The faceted search scopes the range of the keyword search of the terminology according to its main subdivisions. The look-ahead matching capability dynamically displays to the user the possible completions based on the string they have typed. This enables users to navigate efficiently a large terminology set.

Graphical Query Builder

When working with data representing individuals in a population users often need to select a sub-population based on a complex expression of characteristics. For example, a user may need to select the set of patients who are male, over 45 years in age, have never been prescribed statins and have two HBa1C tests in the past 12 months. The problem is how to provide non-technical users with a means of constructing complex queries over temporal health record datasets. The
solution is a query builder that provides the user with the ability to construct arbitrarily complex nested queries using logical operators (AND, OR, NOT) and temporal operators (before, after) using a simple graphical user interface. The user interface allows the user to add conditional statements incrementally build up complex queries. A further extension is to add a natural language description of the query which we have found to be an effective tool to aid understanding.

The Term Selection pattern can be used in conjunction with the Graphical Query Builder pattern.

Multiple Terminologies

There is a wide range of terminology sets for coding Electronic Healthcare Records. This is typically dependent on the setting in which they are used e.g. READ in primary care, LOINC in pathology laboratories or ICD in secondary care. The problem is how to design a system that enable the user to write a query once and then execute it over heterogeneously coded datasets. The solution is to use a canonical terminology for driving the user interface and query building and then use the mapping files to translate to alternative terminologies for actual datasets. This allows new terminology sets to be plugged-in without requiring changes to the main code base. We have successfully used SNOMED-CT as the canonical terminology which has pre-existing mapping files to READ, OPCS and ICD.

Just In Time Terminology Mapping

In systems where queries are federated across multiple data sources we often have heterogeneous terminologies and must use the Multiple Terminologies pattern. The problem is how to design the system such that terminology mapping can
be performed independently of the core query tool. The solution is to perform terminology mapping just in time at the data provider endpoint. This enables it to be managed by the data provider and can be maintained independently of the other components in the federation.

The Multiple Terminologies pattern can be used in conjunction with the Just In Time Terminology Mapping pattern.

Use flat structure data models and NOSQL databases

The patient record is a chronological sequence of coded events. The problem is how best to structure the data model to accommodate different sources of data and to facilitate secondary uses. The solution is to use a flat data model and a NOSQL time series database.