

Business Process Modelling is an Essential Part of a Requirements Analysis

Contribution of EFMI Primary Care Working Group

S. de Lusignan^{1,3}, P. Krause², G. Michalakidis², M. Tristan Vicente³, S. Thompson¹, M. McGilchrist⁴, F. Sullivan⁴, P. van Royen⁵, L. Agreus⁶, T. Desombre¹, A. Taweel⁷, B. Delaney⁸

¹ Department of Health Care Management and Policy, University of Surrey, Guildford, Surrey, UK

² Department of Computing, University of Surrey, Guildford, Surrey, UK

³ Primary Care Informatics, Division of Population Health Sciences and Education, St. George's – University of London, London, UK

⁴ Division of Clinical & Population Sciences and Education, The Mackenzie Building, Dundee, Scotland

⁵ Dept of Primary and interdisciplinary care, University of Antwerp, Antwerpen (Wilrijk), Belgium

⁶ Dept of Neurobiology, Care Sciences and Society, Karolinska Institutet, Huddinge, Stockholm, Sweden

⁷ Department of Informatics and Public Health, King's College London, London, UK

⁸ Department of Primary Care and Public Health Sciences, London, UK

Summary

Objectives: To perform a requirements analysis of the barriers to conducting research linking of primary care, genetic and cancer data.

Methods: We extended our initial data-centric approach to include socio-cultural and business requirements. We created reference models of core data requirements common to most studies using unified modelling language (UML), dataflow diagrams (DFD) and business process modelling notation (BPMN). We conducted a stakeholder analysis and constructed DFD and UML diagrams for use cases based on simulated research studies. We used research output as a sensitivity analysis.

Results: Differences between the reference model and use cases identified study specific data requirements. The stakeholder analysis identified: tensions, changes in specification, some indifference from data providers and enthusiastic informaticians urging inclusion of socio-cultural context. We identified requirements to collect information at three levels: micro- data items, which need to be semantically interoperable, meso- the medical record and data extraction, and macro- the health system and socio-cultural issues. BPMN clarified complex business requirements among data providers and vendors; and additional geographical requirements for patients to be represented in both linked datasets. High quality research output was the norm for most repositories.

Conclusions: Reference models provide high-level schemata of the core data requirements. However, business requirements' modelling identifies stakeholder issues and identifies what needs to be addressed to enable participation.

Keywords

General practice; medical records systems, computerized; research design; registry; records as topic

Yearb Med Inform 2012;34:43

Introduction

Requirements analyses are part of software engineering with the purpose of reducing the chances of failure or the need for frequent expensive changes further along the development pathway. A requirements analysis is the process of determining what a system must do, acknowledging the at times conflicting needs of the various stakeholders [1, 2]. Conventionally conducting a requirements analysis was an early step in the waterfall approach to software development [3, 4, 5]. However, "agile" approaches which include user and stakeholder involvement throughout development [6] and recognize the need for flexibility and change [7] have been used more and also combined with the waterfall approach.

Medical research is a complex process and although there are many computerised data repositories, they are not commonly linked in research studies [8]. These data repositories include: genetic databases, disease registries, and, routinely collected primary care data. Genetic data has been collected into biobanks since the completion of the mapping of the human genome [9, 10], though biological specimens have been used in forensic science for longer

[11, 12]. Cancer registries are also well established [13, 14] and more recently have become international and used to study of a wide range of conditions [15], and the effectiveness of treatments [16]. Primary care data are also widely used in research. These data are rich, collected longitudinally and in countries with registration systems provide a population denominator [17, 18].

Research networks have been developed within primary care, at regional and national level and have improved the capability and capacity for research in primary care; including recruitment into trials [19, 20]. There are many models, sometimes they are part of National or regional sentinel networks [21], or of single vendor linked data collection schemes [15].

There have been a number of attempts to link these different types of data for research [22, 23, 24, 25], but the successes have largely been limited to the use of specific methods in bioinformatics [26, 27].

Objective

We carried out this requirements analysis to identify how to maximise the use

of IT in conducting complex research projects linking aggregated primary care data to a disease registry or a genetic data repository.

Methods

1 Overview

We adopted an agile approach to allow for any alterations in the requirements as other parts of the project developed [28]. We simultaneously conducted a literature review and held expert consensus workshops [29, 30]. We developed a reference model and models for data flow and of the business process involved. We conducted a stakeholder analysis that included interacting with potential data providers and electronic health record (EHR) system vendors.

2 TRANSFoRm Research Programme

This investigation was part of the TRANSFoRm (Translational Research and Patient Safety In Europe) research programme [31]. TRANSFoRm is designed to reduce the barriers to conducting research using routine health data and promote interoperability between different health computer systems so that international data can be collected about the quality of care and for research [32]. One element of TRANSFoRm was a requirements analysis to assess the feasibility of conducting research studies set out in the form of use cases.

3 Creating a Reference Model for the Research Process and Studies

We separately modelled the use case defined research studies, using the Unified Modelling Language (UML) [33]. We excluded consent and data privacy issues as they were being dealt

with elsewhere within the project [34]. We developed a schema of the processes associated with a linked data research project as a reference model (Figure 1) [35] based on our experience of extracting, aggregating and processing routinely collected data [36, 37]. Additionally we recognised that linked research required the same individuals to be represented in both linked databases; as cases present in only one database could not be used for linked research. We called this link a "geographical requirement" (Figure 2) and included it within our business requirements.

We created UML and data flow diagrams (DFD) for the reference model at a high level of abstraction, so we could contrast them with individual study use cases [38, 39, 40].

4 Stakeholder Analysis

We identified the following stakeholder groups and summarised our findings in a mind map and stakeholder analysis grid (Figure 3, Table 2):

- (1) End users of the linked datasets; specifically those looking to run the use case defined linked-data studies.
- (2) Direct and indirect providers of these data. Direct suppliers include database owners or other direct data providers to a study. We spoke in depth to ten. Indirect data providers include those recording or aggregating data earlier in the process.
- (3) The TRANSFoRm study team and their funders, and
- (4) External experts, principally from the International Medical Informatics Association (IMIA) and European Federation for Medical Informatics (EFMI) primary care informatics working groups.

5 Use Case Testing

We analysed the TRANSFoRm use cases to define the data requirements and the integrity of the use cases, constructing a DFD and UML models for each. The use cases were in two clinical domains, type-2 diabetes and acid-reflux related oesophageal disease. The type-2 diabe-

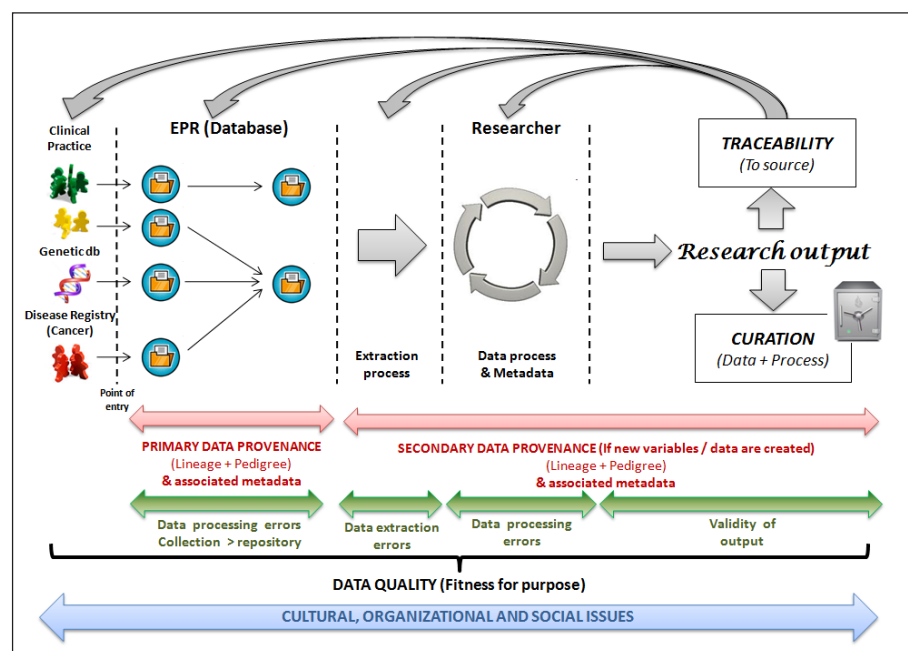


Fig. 1 Schema of the process involved in conducting a linked data research study

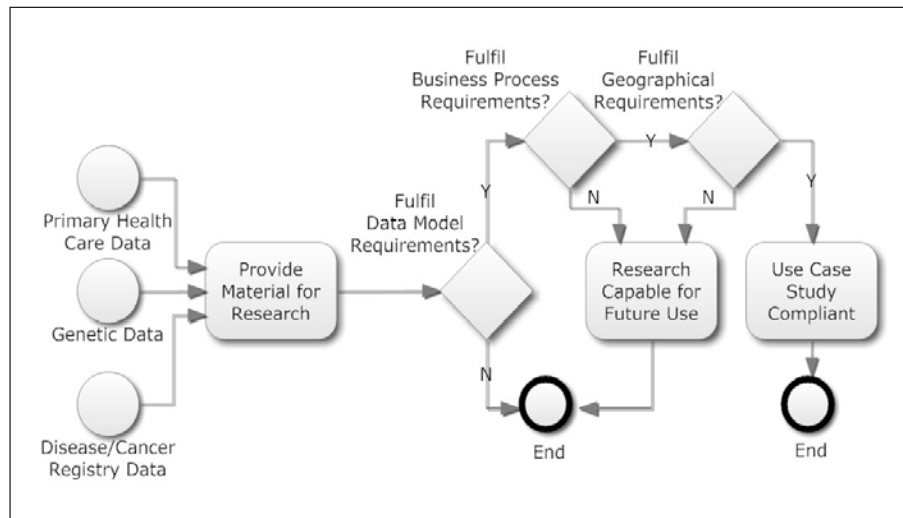


Fig. 2 BPMN schema of the key business decisions which inform whether a data repository is likely to participate in a TRANSFoRm research study

tes cohort study explored the risk of complications and response to oral medication, a cohort study linking primary care and genetic data. The second domain, gastro-oesophageal reflux disease (GORD), included two use cases linking primary care and cancer registry data. One was a randomised controlled trial of on-demand compared with continuous use of proton pump inhibitors (PPIs) a class of anti-indigestion medicine. The other was a nested case-control study to explore the association of GORD symptoms and the use of PPIs with developing Barrett's disease and cancer of the oesophagus.

6 Defining the Scope of the Information Needed from Candidate Databases

We defined the scope of the information we would need to collect from candidate data providers; though asking stakeholders to define what they considered to be essential requirements.

7 Business Process Modelling

Late on in our process we introduced Business Process Modelling Notation (BPMN) to define business requirements.

The stakeholder analysis informed us these were important and they were neglected in our initial programme design.

8 Sensitivity Analysis

We collected data about peer review publications produced by data repositories being considered for inclusion in the TRANSFoRm research studies. We did this because that track record may be predictive of future success. Finally, we conducted a review of what elements of our requirements analysis contributed to our final specifications.

9 Ethical Considerations

This investigation did not involve any direct contacts with patients or access to medical records. No specific ethical consent was deemed necessary.

Results

1 Stakeholder Analysis

We identified key stakeholders and used a mind map to capture key concepts (Figure 3) and summarised their views in a stakeholder analysis grid

(Table 2). There were challenges in getting direct data providers to engage with the process. We made 316 contacts with data providers we thought might participate and only 56 (17.7%) supplied the comprehensive information defined by the requirements analysis. The proportion of valid responses from primary care data providers was 26.4% (29/110); from cancer registries 16.9% (15/89); and from genetic data bases 10.3% (12/117). Many of the data repositories were regional not national and already sustained by a viable programme of research; most were producing high grade research output. Their data processing systems had grown organically as EHR systems had become more sophisticated. They were generally proprietary, not using standard metadata, and functioning as "black boxes" between data entry and export for research. However, all could output data in a range of standard formats.

We identified and contacted 17 EHR vendors identified through our initial searches and a further nine flagged by stakeholders. EHR vendors were reluctant to engage unless there were some chance of benefits realisation including a priori publication of criteria for inclusion in future TRANSFoRm studies. One primary care EHR system reported it has an open application programme interface (API). The lack of open APIs makes this approach to interaction with EHRs more challenging as it might require separate licenced APIs to be developed with each vendor.

TRANSFoRm investigators had different perspectives reflecting their varying backgrounds; they were largely clinical or information systems researchers. These differences were nuanced, but the former generally described the project as a complex clinical process, while the latter looked to rationalise things into machine processable activities. The TRANSFoRm senior investigators were trying to co-ordinate the "big

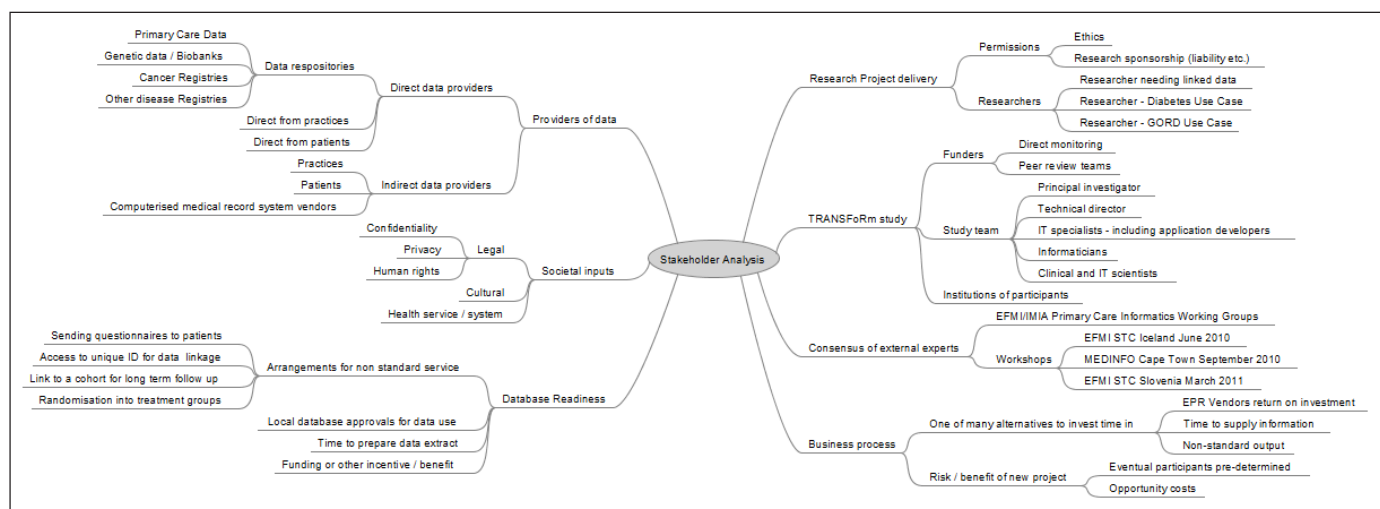


Fig. 3 Mind map of stakeholders and issues

Table 1 Summary stakeholder analysis

	Project Stakeholder	Specific Information Needs	Best Source of Information Needed	Planned Method of Delivery	Timing Considerations
1	End users / researchers	Will data deliver study? 1. Unique linked patient records; 2. Case definition (Coding); 3. Outcome variables; 4. Inclusion & exclusion criteria 5. Study specific - controls, randomisation, patient link	1. Own study protocol 2. Relevant use case 3. Track record of provider	Require a third party to link data	Want to avoid any delays in research process
2	Direct Data providers (Data repositories and networks) Indirect Data providers (Citizens, Patients, GPs)	1. Return on investment of time - studies, funding 2. Strategic interest 1. About use of their data 2. Maintenance of confidentiality/privacy	Who will get a funded study? Track record of researcher Network/ repository owner	Engage once benefits Largely via network/ repository	Have standard processable outputs readily available Challenging to change arrangement
3	Study team & funders	1. Progress towards achieving milestones in project 2. Requirements of next steps in project 3. Meeting own institutions objectives	1. TRANSFoRm study protocol 2. Work package outputs	1. Connecting data via TRANSFoRm engine 2. Provenance engine	Critical nature of dependencies from previous work packages
4	External experts	Academic interest in how and whether this can be done	Simulations & studies - proof of concept	Informal input / advice re: projects	Variable

picture" and focused on project milestones and reports to funders. There were also inevitable changes in project scope. Important changes for the requirements analysis were the inclusion of cancer registries and adoption of *contextual provenance* with the requirement to describe "How that data came to be." This change

in scope of provenance extended beyond what is defined within the use cases as it included influences prior to recording.

The Informatics working group members also broadened the scope of the requirements. They argued that human factors and the wider context of the health system, legal

and socio-cultural factors all influence the process and should be included in the requirements analysis [23, 24]. Their view was that contextual provenance whilst important could only apply to data that were recorded and that a broader contextual overview might better explain why data were present or not.

2 Use Case Analysis

We compared the use cases developed for the TRANFoRM project with the reference model schema we constructed for a generic research study to identify specific data requirements.

The diabetes use case varied little from the reference model schema other than we needed to add a prescribing database. The UML model reflected that this was a cohort study where a subset of the population is identified (Figure 4(a)). The GORD use cases were more complex, also reflecting the study type (Figures 4(b) and 4(c)).

4(a) In this use case genetic and primary care data about type-2 diabetes (T2D) are linked; <<uses>> designates that the information is required at least once. 4(b) Case-control study to exploring the associate of GORD, use of PPIs, and oesophageal adenocarcinoma. 4(c) GORD RCT comparing on-demand and regular PPI

The diabetes study has an almost identical DFD to that in the reference model schema (Figure 5(a)). The GORD case-control study DFD (Figure 5(b)) only differed from the diabetes study DFD as a result of the addition of randomisation. The data flows in the RCT were much more complex and reflected a different approach to collecting data (Figure 5(c)). In this use case, the GP or the researcher mined their EHR data to identify and consent patients to participate at the beginning of the study, whereas in the other studies there is no GP-patient interaction.

3. Scope of the Information Required to Link Candidate Databases:

We identified three levels of granularity of information required to assess if a data source could be used to participate in a research project (Table 2):

- (1) Micro-level requirements were a detailed description of the type data source, the data itself, the potential for linkage and achieving semantic interoperability between data sources [42];

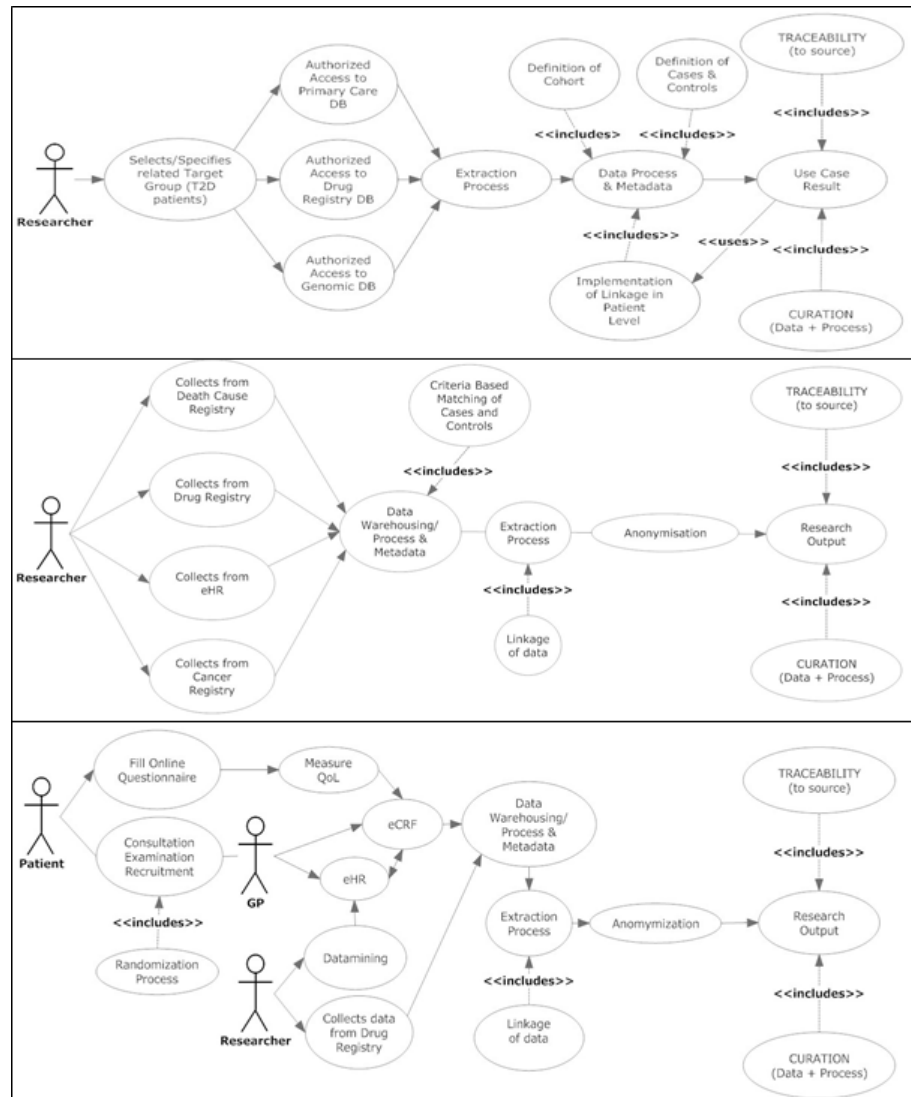


Fig. 4 UML use case model giving an overview research process for: (a) Diabetes (cohort study), (b) GORD (case control) and (c) GORD (randomised controlled trial) use cases

- (2) Meso-level requirements included reporting methods of data extraction [43, 44], details of the EHR systems used and their architecture [45], audit trails that might provide information about data provenance, and the size of the database;
- (3) Macro-level issues related to the nature of the health system and socio-cultural issues.

Study specific requirements were defined by differences between the reference model and use cases.

4 Scope of the Business Requirements

We identified and integrated into our requirements analysis a number of business requirements that went beyond the data requirements for conducting a study (Figure 6). Most of what we framed as business requirements fit within the scope of a research network. The business process modelling was critical in developing an understanding of the lack of engagement, and how concentration on the data model ignored the need for a business case.

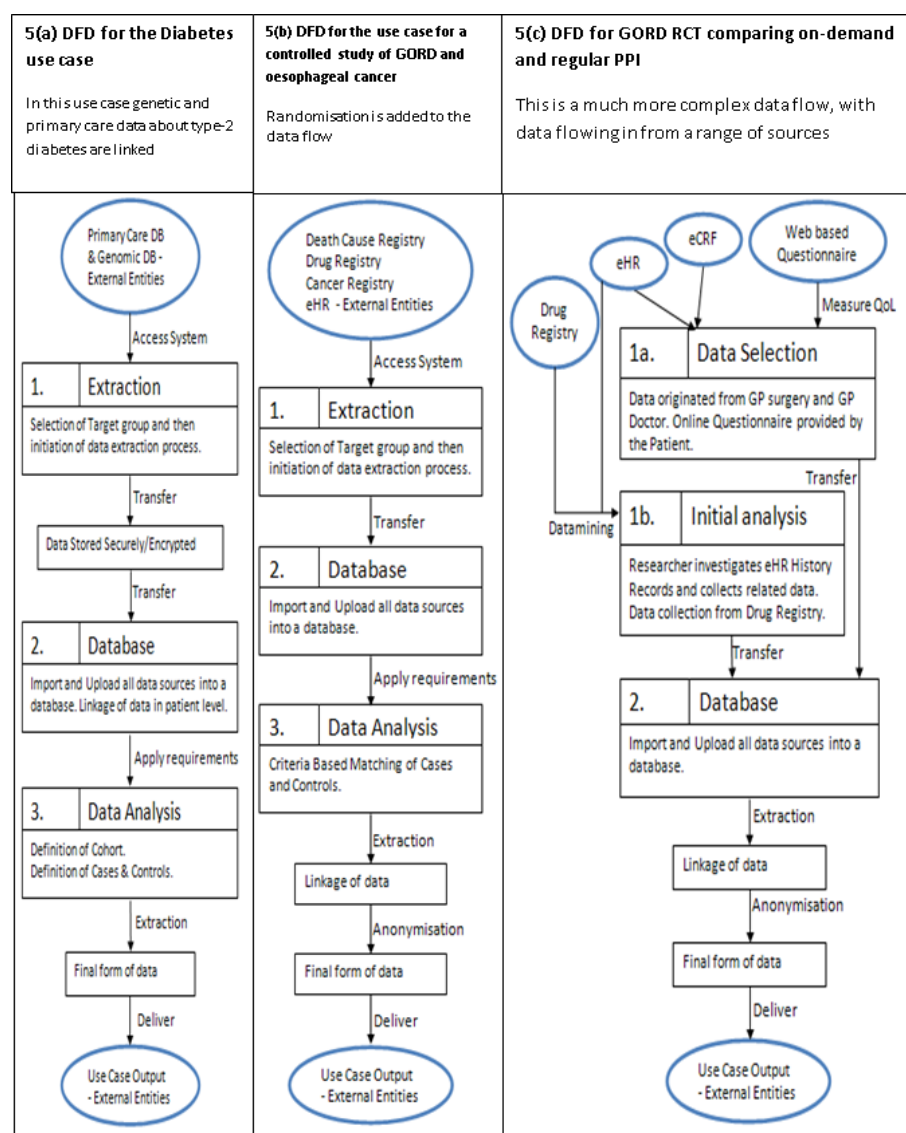


Fig. 5 DFD for the TRANSFoRm use cases

5 Sensitivity Analysis

We used publication of peer-reviewed studies as an index of a functional business process. Eight of the ten data providing stakeholders who provided full information could provide evidence of work produced from their data and indexed in the bibliographic database Medline. Four data providers estimated between 30 and 100 peer-reviewed publications had been published in the last five years. Of the re-

maining four, two estimated 11 to 20 and two estimated 2 to 5 publications.

6 Contributions to the Requirements Analysis

Finally, we report how the different inputs contributed to the final requirements analysis (Table 2). With the exception of socio-cultural issues, all the requirements were identified by more than one source.

Discussion

1 Principal Findings

The requirements for linking complex heterogeneous databases are extensive, and failure to foresee the business requirements resulted in stakeholder disengagement. In addition to the anticipated core data requirements we have identified geographical and socio-cultural requirements. The use case, data driven model, helped define whether a database was suitable to conduct a particular research study. However, we also required information about the geographical overlap between data repositories and readiness to participate. We used a reference model to identify generic as well as study specific requirements. Local language and the regional nature of some data repositories are additional challenges to conducting international research. Many of the successful data collection systems were proprietary and saw no reason to move to more open methods and systems; they believed this entailed risk and expense but with no obvious benefit. Only after we included BMPN in our approach did we manage to rationalise and frame these requirements.

2 Implications of the Findings

The data requirements from the use case analysis identified most data needed to conduct research, but may miss important business issues which may constrain involvement. Whilst generic reference models for research studies make sense of the processes and data flows; the business requirements of the data repository owners and EHR vendors also need to be modelled and met. Researchers who design data projects that involve linking databases need to make sure that they include incentives for participation otherwise they risk low levels of engagement.

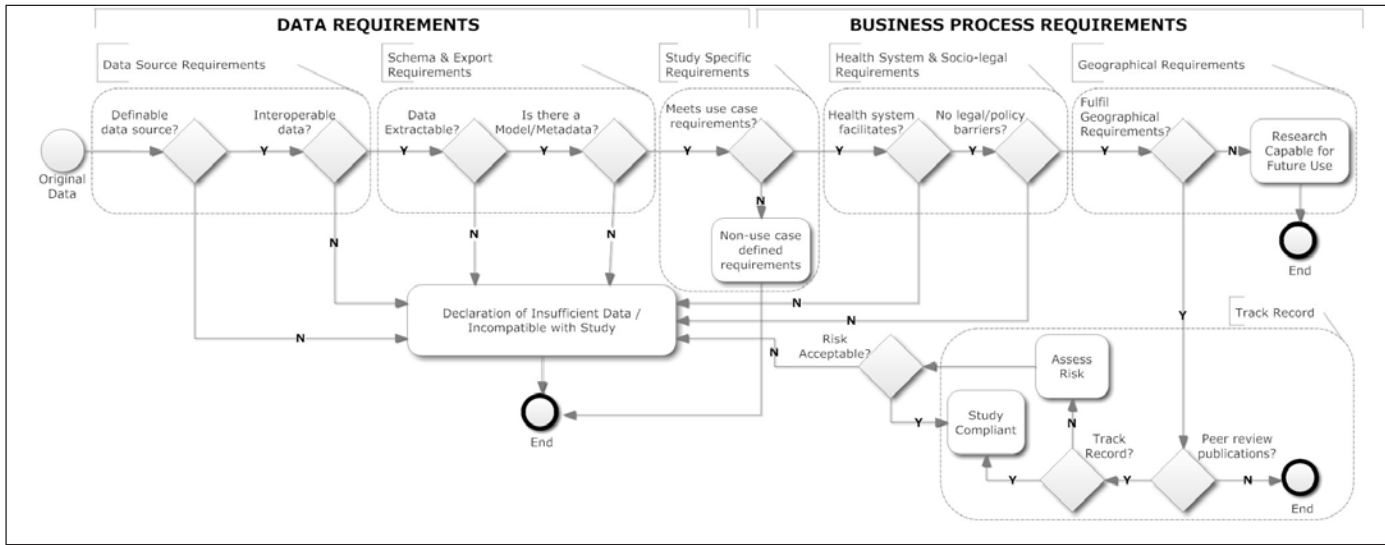


Fig. 6 Detailed BPMN diagram to demonstrate the steps in ascertaining if a data source meets the business requirements

Table 2 Scope of data collection defined by the requirements analysis

Micro-level		Meso-level		Macro-level		Study specific
Data source/ Data type	Data Interoperability level	Method & control of data extraction	Record system & information model	Organisational level	Socio-cultural factors	Use-Case compatibility
Genetic data	HL7 Health Level Seven - www.hl7.org	Data extracted Demographics ID Coded Data Free-text Data quality	Use case/protocol and EPR architecture	Primary Health Care System	Wider social influences	Diabetes study Primary care & genetic data
Primary Care/ Cancer Registry data	CDISC Clinical Data Exchange Consortium - www.cdisc.org	Taxonomy of errors associated with data extraction	PCROM Primary Care Research Model	Fragmentation of care Separate providers	On data availability / consent & access to data	
All types of data	BRIDG Biomedical Research Integrated Domain Group - www.bridg.org	ISO 11179 Metadata http://metadata-standards.org/11179	OpenEHR Archetypes www.openehr.org	Coverage Univesal/limited	Socio-cultural dimensio to what is disease & presented to primary care	Barrets oesophagus & medicatoin
				Management of Ambulatory care sensitive conditions		

One strategy to overcome this might be creating a research network infrastructure to register researchers and data providers wishing to conduct linked data studies that might provide a forum to facilitate and broker solutions.

3 Comparison with the Literature

Linking heterogeneous data is challenging [46]. Use cases and UML have used in research reporting adverse events in clinical trials [47] and medical image analysis pilots [48]. The OntoRAT toolkit

is a working example of the use of ontologically driven systems [49]. Such ontologically driven systems might best be developed using object-orientated approaches as has been used in specialist cancer management [50] and for checking healthcare domain models [51].

Although widely used in software engineering, the modelling tools we propose to be used have been little used in the health domain; despite calls that they should be made more accessible to health researchers [52]. DFD have been used little in health care though they have been proposed as a component of object orientated health information management system [53]. UML has been used the most. It has been proposed as a tool for avoiding integration problems [54] and for modelling research studies [55], but thus far with little evidence of benefit [56]. BPMN is now considered an appropriate tool for modelling digital pathology and care pathways but has not been proposed for modelling health research [57, 58].

4 Limitations of the Method

We could have just used UML. The UML alternatives are either to develop a domain model or use UML state-machine or finite state machine diagrams [26-8]. Business architecture models (BAM) have been applied

to the life sciences to help generate UML models [59]; and other tailored packages have been used to generate UML diagrams for bioinformatics research [60] and to generate genotype-phenotype maps [61].

Our requirements analysis produced a non-executable set of models and methods for sorting candidate databases into those potentially useable in linked data research. We did not define these steps in sufficient detail to be able to set them out using business process executable language (BPEL) specifications [62], BPEL has been used within health-care [63], and might help reduce barriers to potential researchers and data providers.

5 Call for Further Research

The user requirements developed need to be tested prospectively. Where the required activities can be published with clearly defined interfaces it will be possible to produce an executable model such as BPEL.

Conclusions

We defined generic user requirements for research studies linking health data based on a reference model, which can be extended for specific studies; using UML use case diagrams and DFD. Additionally we have identified geographical requirements and the importance of socio-cultural context. However, the key learning from this exercise is the importance of modelling the business process and including this within a requirements analysis. Failure to include and recognise the importance of modelling the business process as part of requirements analysis risks lack of engagement by intended participants and increases the risk of project failure.

Acknowledgements

IMIA and EFMI for supporting their primary care informatics working groups. Elena Crecan for her contribution to the research. TRANSFoRM is part-financed by the European Commission - DG INFSO (FP7 2477). Antonis Ntasioudis for assistance with the diagrams and modelling.

Data sharing agreement

The limited quantitative data collected as part of this study have been placed on a database at University of Dundee with the Work Package lead (FS); initially for further analysis within the TRANSFoRM project. The generic models developed as part of this study are freely available for use from this paper or from the clinical informatics website: <http://www.clininf.eu/refmodel/> - users are requested to cite this paper.

References

1. Cadle J, Paul D, Turner P. Business analysis techniques, 72 Essential Tools for Success. Swindon: BCS; 2010.
2. Alexander I, Stevens R. Writing better requirements. London: Addison-Wesley; 2002.
3. Krause P, de Lusignan S. Procuring interoperability at the expense of usability: a case study of UK

Table 3 Contribution of use cases and stakeholder analysis to final requirements

Data level / type	Micro		Meso		Macro		Study Specific		Sensitivity analysis
	Patient level data	Coding system & Interoperability	Metadata & Extraction	Record system & information model	Health service organisational	Socio-cultural & legal	Diabetes & GORD studies	Non-use case	Track record peer review publications
Use cases	x	xx		x			xx		
Stakeholder analysis:									
End users	xx	xx	x	x			xx	xx	xx
Data Providers	xx		xx				xx		x
Study Team	xx		x	x	x			x	x
External experts	x	x	xx	xx	xx	xx			xx

Key: x=contribution to requirements; xx= major contribution to requirements

- National Programme for IT assurance process. *Stud Health Technol Inform* 2010;155:143-9.
4. Hood C, Wiedemann P, Fichtinger S, Pautz U. The interface between requirements development and all other systems engineering processes. Berlin; Springer-Verlag; 2010.
 5. Reddy M, Pratt W, Dourish P, Shabot M. Sociotechnical requirements analysis for clinical systems. *Methods Inf Med* 2003;42(4):437-44.
 6. Montabert C, McCrickard D, Winchester W, Perez-Quinones M. An integrative approach to requirements analysis: How task models support requirements reuse in a user-centric design framework. *Interact Comput* Aug 2009;21(4):304-15. DOI: 10.1016/j.intcom.2009.06.003
 7. Surendra N. Using an ethnographic process to conduct requirements analysis for agile systems development. *Inf Technol Manag* 2008;9(1):55-68. DOI 10.1007/s10799-007-0026-6
 8. Matteson S, Paulauskis J, Foisy S, Hall S, Duval M. Opening the gate for genomics data into clinical research: a use case in managing patients' DNA samples from the bench to drug development. *Pharmacogenomics* 2010;11(11):1603-12.
 9. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001;409(6822):860-921
 10. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;431(7011):931-45.
 11. Hirtzlin I, Dubreuil C, Préaubert N, Duchier J, Jansen B, Simon J, et al.; EUROGENBANK Consortium. An empirical survey on biobanking of human genetic material and data in six EU countries. *Eur J Hum Genet* 2003;11(6):475-88.
 12. McEwen J. Forensic DNA data banking by state crime laboratories. *Am J Hum Genet* 1995;56(6):1487-92.
 13. Schneider B. Progress in cancer control through cancer registries. *CA Cancer J Clin* 1958;8(6):207-10
 14. Navarro C, Martos C, Ardanaz E, Galceran J, Izarzugaza I, Peris-Bonet R, et al.; Spanish Cancer Registries Working Group. Population-based cancer registries in Spain and their role in cancer control. *Ann Oncol* 2010;21 Suppl 3:iii3-13.
 15. Gatta G, Capocaccia R, Trama A, Martínez-García C; RARECARE Working Group. The burden of rare cancers in Europe. *Adv Exp Med Biol* 2010;686:285-303.
 16. Francisci S, Capocaccia R, Grande E, Santaquilani M, Simonetti A, Allemani C, et al.; EUROCCARE Working Group. The cure of cancer: a European perspective. *Eur J Cancer* 2009 Apr; 45(6):1067-79.
 17. de Lusignan S, Chan T. The development of primary care information technology in the United Kingdom. *J Ambul Care Manag* 2008 Jul-Sep;31(3):201-10.
 18. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Fam Pract* 2006;23(2):253-63.
 19. Clement S, Pickering A, Rowlands G, Thiru K, Candy B, de Lusignan S. Towards a conceptual framework for evaluating primary care research networks. *Br J Gen Pract* 2000;50(457):651-2.
 20. Peterson KA, Fontaine P, Speedie S. The Electronic Primary Care Research Network (ePCRN): a new era in practice-based research. *J Am Board Fam Med* 2006;19(1):93-7.
 21. Boffin N, Bossuyt N, Vanthomme K, Van Casteren V. Readiness of the Belgian network of sentinel general practitioners to deliver electronic health record data for surveillance purposes: results of survey study. *BMC Fam Pract* 2010;11(1):50.
 22. Friedman C, Hripcsak G, Johnson SB, Cimino JJ, Clayton PD. A generalized relational schema for an integrated clinical patient database. *Proceedings 14th Annual Symposium on Computer Applications in Medical Care*. Los Alamitos, CA: IEEE Computer Society Press; 1990. p. 335-9.
 23. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods Inf Med* 2009;48:38-44.
 24. Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. *Yearb Med Inform* 2008;91-101.
 25. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. *Interoperability and integration*. *Methods Inf Med* 2009;48(1):45-54.
 26. Mirhaji P, Zhu M, Vagnoni M, Bernstam EV, Zhang J, Smith JW. Ontology driven integration platform for clinical and translational research. *BMC Bioinformatics* 2009;10 Suppl 2:S2.
 27. Wang X, Liu L, Fackenthal J, Cummings S, Olopade OI, Hope K, et al. Translational integrity and continuity: personalized biomedical data integration. *J Biomed Inform* 2009;42(1):100-12.
 28. Huo M, Verner J, Zhu L, Ali Barbar M. Software quality and agile methods. *Proceedings of 28th Annual International Computer Software and Applications Conference (COMPSAC'04) 2004;Vol 1: 520-5*. <http://doi.ieeecomputersociety.org/10.1109/CMPSAC.2004.1342889>
 29. de Lusignan S, Peace C, Shaw N, Liaw S-T, Michalakidis G, Vincente M, et al. What are the barriers to conducting international research using routinely collected primary care data? *Stud Health Technol Inform* 2011;165:135-40. DOI: 10.3233/978-1-60750-735-2-135
 30. de Lusignan S, Liaw S-T, Krause P, Curcin V, Vincente M, Michalakidis G, et al. Key concepts to assess the readiness of data for international research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation. *Yearb Med Inform* 2011;6(1):112-20.
 31. TRANSFoRM (Translational Research and Patients safety In Europe). URL: <http://www.transformproject.eu>
 32. European Union, Europe's Information Society. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions - e-Health - making healthcare better for European citizens: an action plan for a European e-Health Area {SEC(2004)539}; 2004. URL: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52004DC0356:EN:NOT>
 33. Kumarapeli P, de Lusignan S, Ellis T, Jones B. Using Unified Modelling Language (UML) as a process-modelling technique for clinical-research process improvement. *Med Inform Internet Med* 2007 Mar;32(1):51-64.
 34. de Lusignan S, Chan T, Theadom A, Dhoul N. The roles of policy and professionalism in the protection of processed clinical data: a literature review. *Int J Med Inform* 2007 Apr;76(4):261-8.
 35. Blobel B. Ontologies, knowledge representation, artificial intelligence – hype or prerequisites for International pHealth interoperability. *Stud Health Technol Inform* 2011;165:11-20. DOI: 10.3233/978-1-60750-735-2-11
 36. van Vlymen J, de Lusignan S, Hague N, Chan T, Dzregah B. Ensuring the quality of aggregated general practice data: lessons from the Primary Care Data Quality Programme (PCDQ). *Stud Health Technol Inform* 2005;116:1010-15.
 37. van Vlymen J, de Lusignan S. A system of metadata to control the process of query, aggregating, cleaning and analysing large datasets of primary care data. *Inform Prim Care* 2005;13(4):281-91.
 38. Podeswa H. UML™ for the IT Business Analyst: A Practical Guide to Object-Oriented Requirements Gathering. Boston, MA: Thompson Course technology PTR; 2005.
 39. Fakhroudinov K. UML, Use Case Diagrams. URL: <http://www.uml-diagrams.org/use-case-diagrams.html>
 40. Ambler S. UML2 Use Case Diagrams. URL: <http://www.agilemodeling.com/artifacts/useCaseDiagram.htm>
 41. Object Management Group (OMG) business management initiative. Business process modelling notation (BPMN), version 2. URL: <http://www.bpmn.org/>
 42. Dolin RH, Alschuler L. Approaching semantic interoperability in Health Level Seven. *J Am Med Inform Assoc* 2011 Jan 1;18(1):99-103.
 43. Michalakidis G, Kumarapeli P, Ring A, van Vlymen J, Krause P, de Lusignan S. A system for solution-orientated reporting of errors associated with the extraction of routinely collected clinical data for research. *Stu Health Technol Inform* 2010;160:724-8.
 44. Nadkarni, P.M., Brandt, C. Data Extraction and Ad Hoc Query of an Entity-Attribute-Value Database. *J Am Med Inform Assoc* 1998; 5:511-7
 45. Santos MR, Bax MP, Kalra D. Building a logical EHR architecture based on ISO 13606 standard and semantic web technologies. *Stud Health Technol Inform* 2010;160(Pt 1):161-5.
 46. Castano S, De Antonellis V. Global viewing of heterogeneous data sources. *IEEE Transactions on Knowledge and Data Engineering* 2001;13(2):277-97. DOI: <http://dx.doi.org/10.1109/69.917566>
 47. London JW, Smalley KJ, Conner K, Smith JB. The automation of clinical trial serious adverse event reporting workflow. *Clin Trials* 2009; 6(5):446-54.
 48. Estrella F, Hauer T, McClatchey R, Odeh M, Rogulin D, Solomonides T. Experiences of engineering Grid-based medical software. *Int J Med Inform* 2007;76(8):621-32.
 49. Al-Hroub Y, Kossmann M, Odeh M. Developing

- an Ontology-driven Requirements Analysis Tool (OntoRAT): A Use-case-driven Approach. Proceedings of second international conference on the applications of digital information and web technologies (ICADIWT 2009) 2009:130-8.
50. Weber R, Knaup P, Knietig R, Haux R, Merzweiler A, Mludek V, et al. Object-oriented business process analysis of the cooperative soft tissue sarcoma trial of the german society for paediatric oncology and haematology (GPOH). *Stud Health Technol Inform* 2001;84(Pt 1):58-62.
 51. Baksi D. Model checking of healthcare domain models. *Comput Methods Programs Biomed* 2009;96(3):217-25.
 52. Jun GT, Ward J, Morris Z, Clarkson J. Health care process modelling: which method when? *Int J Qual Health Care* 2009;21(3):214-24.
 53. Krol M, Reich DL. Object-oriented analysis and design of a health care management information system. *J Med Syst* 1999;23(2):145-58.
 54. Benson T. Prevention of errors and user alienation in healthcare IT integration programmes. *Inform Prim Care* 2007;15(1):1-7.
 55. Kumarapeli P, De Lusignan S, Ellis T, Jones B. Using Unified Modelling Language (UML) as a process-modelling technique for clinical-research process improvement. *Med Inform Internet Med* 2007;32(1):51-64.
 56. Vasilakis C, Leczarowicz D, Lee C. Application of Unified Modelling Language (UML) to the Modelling of Health Care Systems: An Introduction and Literature Survey. In Ed. Tan J. *Developments in Health Information Systems and Technologies: Models and methods*. Vancouver, BC: IGI-global; 2010,275-87. DOI: 10.4018/978-1-61692-002-9.ch019.
 57. Scheuerlein H, Rauchfuss F, Dittmar Y, Molle R, Lehmann T, Pienkos N, et al. New methods for clinical pathways-Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM). *Langenbecks Arch Surg* 2012 Feb 24. [Epub ahead of print]
 58. Rojo MG, Daniel C, Schrader T. Standardization efforts of digital pathology in Europe. *Anal Cell Pathol (Amst)* 2012;35(1):19-23.
 59. Becnel Boyd L, Hunicke-Smith SP, Stafford GA, Freund ET, Ehlman M, Chandran U, et al. The caBIG(R) Life Science Business Architecture Model. *Bioinformatics* 2011 Mar 29. [Epub ahead of print]
 60. Yan Q. Bioinformatics for transporter pharmacogenomics and systems biology: data integration and modeling with UML. *Methods Mol Biol* 2010;637:23-45.
 61. Baker EJ, Jay JJ, Philip VM, Zhang Y, Li Z, Kirova R, et al. Ontological Discovery Environment: a system for integrating gene-phenotype associations *Genomics* 2009;94(6):377-87.
 62. Zheng Y, Zhou J, Krause P. An Automatic Test Case Generation Framework for Web Services. *Journal of Software* 2007;2(3):64-77. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.6186&rep=rep1&type=pdf>
 63. Tan W, Missier P, Foster I, Madduri R, Goble C. A Comparison of Using Taverna and BPEL in Building Scientific Workflows: the case of caGrid. *Concurr Comput* 2010;22(9):1098-1117.

Correspondence to:

Simon de Lusignan
 Department of Health Care Management and Policy
 University of Surrey
 Guildford, Surrey, UK
 Tel +44 (0) 1483 683 089
 E-mail: s.lusignan@surrey.ac.uk