Key Concepts to Assess the Readiness of Data for International Research: Data Quality, Lineage and Provenance, Extraction and Processing Errors, Traceability, and Curation

Contribution of the IMIA Primary Health Care Informatics Working Group

S. de Lusignan[1], S.-T. Liaw[2], P. Krause[3], V. Curcin[4], M. Tristan Vicente[5], G. Michalakidis[6], L. Agreus[7], P. Leysen[8], N. Shaw[9], K. Mendis[10]

[1] IMIA Primary Healthcare Working Group Co-Chair, Primary Care and Clinical Informatics, University of Surrey, UK
[2] General Practice, University of New South Wales, Australia
[3] Software Engineering, University of Surrey
[4] Imperial College London
[5] St. George's University of London
[6] Computing department, University of Surrey
[7] Center for Family and Community Medicine, Karolinska Institutet, Stockholm
[8] Faculty of Medicine, Dept. of Primary and Interdisciplinary Care, University of Antwerp
[9] ESRI Canada Health Informatics Research Chair / Scientific Director, Health Informatics Institute, Algoma University, Ontario, Canada
[10] IMIA Primary Healthcare Working Group Chair, University of Sydney, Australia

Summary

Objective: To define the key concepts which inform whether a system for collecting, aggregating and processing routine clinical data for research is fit for purpose.

Methods: Literature review and shared experiential learning from research using routinely collected data. We excluded socio-cultural issues, and privacy and security issues as our focus was to explore linking clinical data.

Results: Six key concepts describe data: (1) Data quality: the core Overarching concept – Are these data fit for purpose? (2) Data provenance: defined as how data came to be; incorporating the concepts of lineage and pedigree. Mapping this process requires metadata. New variables derived during data analysis have their own provenance. (3) Data extraction errors and (4) Data processing errors, which are the responsibility of the investigator extracting the data but need quantifying. (5) Traceability: the capability to identify the origins of any data cell within the final analysis table essential for good governance, and almost impossible without a formal system of metadata; and (6) Curation: storing data and look-up tables in a way that allows future researchers to carry out further research or review earlier findings.

Conclusion: There are common distinct steps in processing data; the quality of any metadata may be predictive of the quality of the process. Outputs based on routine data should include a review of the process from data origin to curation and publish information about their data provenance and processing method.

Keywords

Medical records systems, computerized; research design; registry; records as topic; databases genetic

Yearb Med Inform 2011:112-20

1. Introduction

Findings from International studies are more likely to be generaliseable, as the intervention will be tested across a range of cultures, ethnic groups, and health systems. The computerisation of primary care should facilitate that process [1]. However, the type of data collected, the way data are structured and coded, and language used may vary between health systems [2]; creating challenges for the primary care researcher to overcome [3].

Currently, there is no international consensus of how to describe data quality and its usability for research. The Translational Medicine and Patient Safety in Europe (TRANSFoRm) project aims to remove some of the barriers to conducting International research by developing a process which facilitates the conduct of research across European states [4]. The programme has developed two use-cases - simulated data requirements for research studies - to provide a specification against which to test the potential of existing databases for research. However, their utility is limited by a lack of consistency in many of the terms used to define the origin and quality of the data.

We carried out this review to try to define the terms which should be used to define the origin and quality of data. This is needed if primary care research is to grow from studies principally carried out in single countries in networks drawn from a single vendor to international studies which link primary care data to other data sources. These additional sources may be other health providers (e.g. hospitals and clinics); or other health data (e.g. genetic data from biobanks, or disease registry data); or social data.

2. Method

Our method was based on a literature review, workshop discussions, and developing an expert consensus. We carried out a literature review using the ISI Web of Knowledge and Pubmed Medline. We searched using the following key words: provenance, lineage,

pedigree and traceability. We combined the key word in the following ways:

"Data provenance“ OR "Provenance of data"

Provenance AND Database

Provenance AND "Service Oriented Architecture"

We repeated these searches for the other key terms. For pedigree we added "NOT genetic*" as searches for pedigree were swamped by descriptions of genetic pedigree.

We also searched for papers about data quality using:

"Data Quality" AND "Medical records systems, computerized" or using as a second search term "Computers", "Classification", or "Family practice".

We took a linear view of the overall research process: from the point of data recording to the creation of the final tables for analysis by the researcher (Fig 1 and 2). We determined that data quality was the overarching concept and that each step of the process needs to have an unambiguous quality descriptor. The role and place of each descriptor was determined by consensus.

We harnessed international informatics expertise from the IMIA (International Medical Informatics Working Group) [5] and EFMI (European Federation for Medical Informatics) [6] Primary care Informatics working groups (PCI WG); and discussed this theme at the WG workshops at the 2010 EFMI conference in Reykjavik, the 2010 IMIA MEDINFO conference in Cape Town, in subsequent email discussions, and within the TRANSFoRm working group.

We excluded broader issues relating to the quality of research data: social and cultural context; health service organisation and study specific issues as they formed part of a previous study [7]. Our investigation was orientated toward family practice, and we only included studies relevant to research which might be relevant to primary care research; albeit that such research often needs to link to other data sources to identify high risk groups or provide health outcomes data.

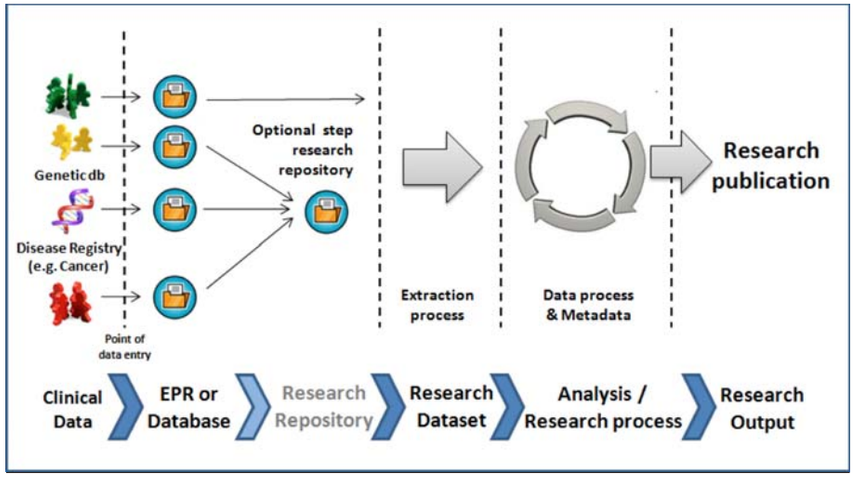

Fig. 1 Research, using routine data, as a linear process

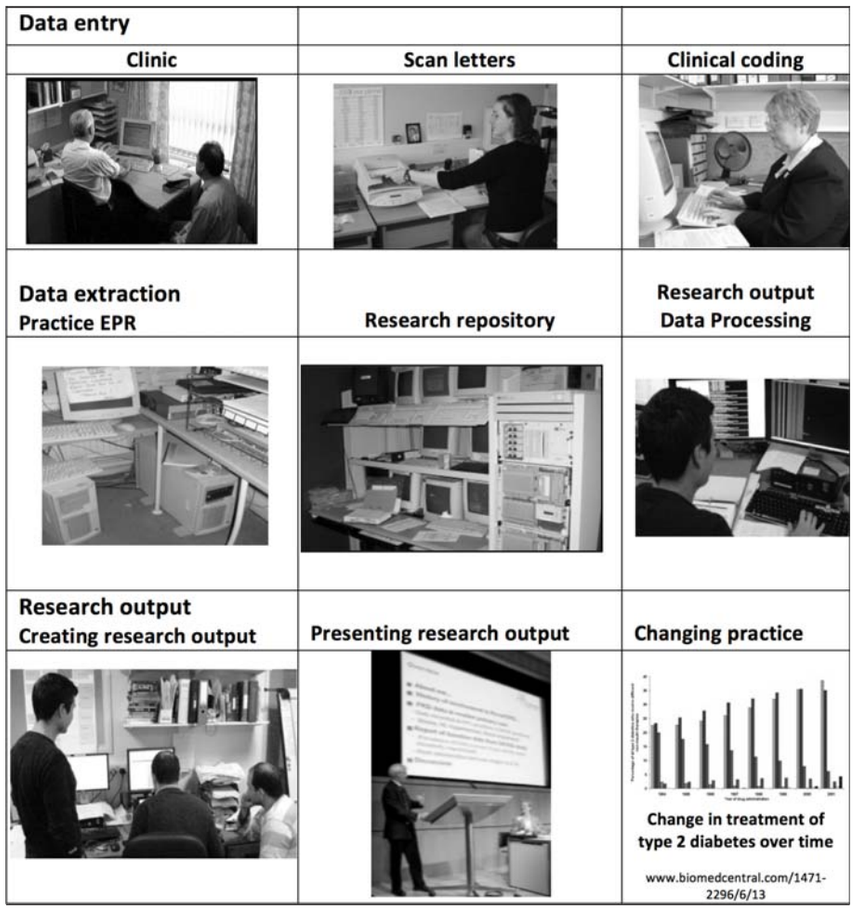

Fig. 2 The process from data recording to data analysis

Similarly, we did not consider consent, governance, privacy or data security issues; including obfuscation or other methods for masking patient identities in aggregated records [8]. Finally we excluded any data migration process between the electronic patient record system (EPR) where the data were recorded and research data repository. There are a range of methods used at this step; most are proprietary and often collect data from a single brand of computer system. There is a dearth of literature about this step, and no ready mechanism to investigate further.

We also tested our data quality model using the TRANSFoRm use-cases: One is a study of the genetics of type 2 diabetes requiring linked primary care and genetic data; the second a study to explore the relationship between gastro-oesophageal reflux disease (GORD) and its treatment in primary care with oesophageal cancer. We decided to simulate the likely very different ontologies (concepts and their relationships), informational models and semantic issues around these use-cases; and use them to test the face validity of our definitions.

3. Results

Data Quality

The International Standards Organisation (ISO) defines quality as:

> "*The totality of features and characteristics of an entity that bears on its ability to satisfy stated and implied needs*" [9]

This is echoed by the EFMI PCI WG who defined data quality as "fitness for purpose." [10]

Data quality was initially defined in quantitative terms, using measureable components that give some indication of the validity of the data; data quality was initially defined in terms of completeness and accuracy [11]. Later definitions added the concept of currency [12]: Subsequently, it was defined as the positive predictive value and sensitivity [13] or by its specificity [14].

More recently, definitions have tried to be more analytical. Aqil et al., suggest that comparisons should be made at each step: (1) Comparing what data are collected with information needs; (2) Are all data fields filled and how do they compare with expected levels of completeness; (3) Are data entry timely compared with the norm; and (4) Accuracy should be tested by comparing between records and with other data sources [15]. Arts et al., propose analysis of the whole process, using planned and systematic procedures before, during, and after data collection [16].

Improving the Quality of Data Entry

The quality of data entry for the same clinical scenario will differ between clinicians. It can reflect: the pattern of computer use of the clinician [17]; aspects of the computer interface: for example where picking lists are used for clinical coding the lists can vary between brands apparently using the same coding system [18] and cardiovascular risk scores can vary between brands [19]; or external influences, e.g. Direct transmission of pathology results into the computer system. Moves towards a more service oriented architecture where industry standard tools are used may help standardise processes. For example, semantic lookup services might standardise the way coding systems are accessed and reduce variation in coding between different brands of EPR [20], and an interoperable cardiovascular risk calculator could standardise risk calculation between different EPR systems [21].

There are three principal ways that the quality of data entry can be improved: (1) Feedback ideally through regular meetings and education[22], with or without the provision of token [23] or substantial financial incentives [24]. (2) Use of data entry forms which either facilitate or mandate the collection of a partial or complete dataset [25], or structuring the record in a way that forces linkage between problem and therapy, often referred to as problem orientation [26]. (3) Decision support which either prompts for missing data or which suggest diagnoses [27].

Data Lineage and Provenance

The terms provenance, pedigree, lineage and traceability have all been used to describe the origins of data within the informatics literature. The first references to these were: data lineage in 1991 [28], traceability in 1995 [29], pedigree in 1998 [30], and provenance in 2000 [31].

Data lineage is an output record of all the contributory inputs [32]. Traceability links the outputs to their originating inputs across a system. Data pedigree predominantly refers to the authority of the source; implying that data of good pedigree can be trusted [33].

"Data provenance" first appeared in Medline in 2004 [34]. However, in the broader scientific literature pedigree and data lineage are used as near synonyms [35]. Provenance is a type of metadata, concerned with the history of data, its origin and changes made to it, often including versioning information [36]. Definitions of data provenance include:

> "The history, lineage or provenance of a given piece of data provides understanding of how it was that the data came to be as it is. This understanding enables users to validate data by providing the means to examine the processes that produced it, for fitness for purpose, compliance to regulations, replication, validation and examination." [37]

> "The provenance of a data item includes information about the processes and source data items that lead to its creation and current representation." [38]

In medical software systems, the data (EPR and instrument data), the workflow (procedures carried out to perform extraction and analysis) and the histories (recording meaningful events in those procedures) may be distributed among several heterogeneous and autonomous

information systems. Capturing this knowledge requires a provenance framework that is separate from the individual systems, enabling the traceability of the origins of decisions and processes, the information that was available at each step, and where that information came from. In turn, this provides an integrated view of treatment processes, and enables performance analysis and procedure audit of distributed healthcare services.

Data provenance has five potential functions: (1) Understanding and ultimately Improving data quality; (2) Providing an audit trail of the data (this encompasses lineage); (3) Generating replication recipes to allow the process to be reproduced; (4) Providing attribution and ownership of the data (an important feature of pedigree): and (5) Enabling the discovery of new information about a process [39].

A provenance standard has been developed to facilitate collaboration. The Open Provenance Model [40] is designed to meet the following requirements: (1) Allow provenance information to be exchanged between systems, by means of a compatibility layer; (2) Build and share tools; (3) To define the model in a precise, technology-agnostic manner; (4) Support a digital representation of provenance for any „thing“, whether produced by computer systems or not; (5) To define a core set of rules that identify the valid inferences that can be made on provenance graphs. A provenance store can also be defined as part of a workflow based service oriented architecture. There can be single or multiple provenance stores depending on the complexity or requirements of the workflow [41].

We recommend the use of primary data provenance as the overarching term between the point of data entry and the point that the researcher extracts their data; and that that lineage and pedigree are used in data quality as subordinate terms. Lineage is primarily a study of where data comes from and pedigree a term with more emphasis on the quality and trustworthiness of data, though this is not the way pedigree is used within genetics. If the data remains unaltered through subsequent processing then this primary data provenance will apply throughout; however if in the process of extraction, analysis or curation new variables are created then this secondary data will have its own provenance. These different parts of the workflow could have a single or their own provenance store [37].

Table 1 Applications of provenance information and overlap with pedigree and lineage

Applications	How to use the information	Pedigree	Lineage
Data quality	Estimate data quality and data reliability based on the source data and transformationsProvide proof statements on data derivation	Y	
Audit trail	Trace the audit trail of data Determine resource usageDetect errors on data generation	Y	Y
Replication recipes	Allow repetition of data derivation Help maintain its currency Be a recipe for replication		Y
Attribution	Establish the copyright and ownership of dataDetermine liability in case of erroneous data	Y	
Informational	Query based on lineage metadata for data discoveryProvide a context to interpret data	Y	Y

Data Extraction Errors

Taking data from one system to another inevitably involves data loss. The migration process often includes pre-processing to anonymise or completely remove data items that identify individuals. In data extraction difficulties and errors arise because of the different architectures of the heterogeneous distributed systems, the local autonomy of these systems, problems in representational diversity of the same clinical concept, and the potential lack of precise semantic meaning [42]. We suggest the use of this error taxonomy [11] to share errors and facilitate the identification of underlying causation and enable them to be rectified (Table 2).

Data Aggregation, Linkage and Processing Errors

More than one source of data may be required to conduct a research study; for example the TRANSFoRm use-cases require primary care and genetic data to conduct one study and primary care and cancer registry data for the other (See scenarios boxes 1 and 2.)

Data aggregation is an eight step process: (1) Design (2) Data entry, (3) Extraction, (4) Migration, (5) Integration, (6) Cleaning, (7) Processing, and (8) Analysis [43]. All these steps involve making assumptions and are prone to error. Steps two and three are dealt with earlier in this paper; and not discussed further. The data processing design has to take into account whether the analysis is using complete data, just coded data, or whether there is access to free text and taking into account missing data [44, 45]. And, the conversion of extracted code into information, namely the process of creating deriving new variables. The process of creating derived variables involves, cleaning data to remove non-credible values; grouping the data into categories; using cut points or combining variables to produce a new variable (e.g. calculating cardiovascular risk) [19], relevant to the intended analysis. Any new data created must have its provenance defined; and its own metadata [46].

Data linkage between data sources is becoming more and more important in research based on routine data. There are a number of dimensions to linkage and how it can be facilitated. Registration based health systems where one individual only registers with one primary care provider; and health systems with a unique identifier can more readily link data than those without. More recently systems of private

Table 2 Types of extraction errors and consequences

Error type	Descriptor	Consequences
Patient record system	Different brands of EPR systems often vary in terms of their interface, the data entry forms etc.	What is correctly coded from the data source's point of view not necessarily moved across layers correctly: ***Metadata requirements and audit-trails/ provenance considerations.***
Data extraction tool	The extraction tool implementations may export data differently from one another, requiring different query libraries.	Renders some extracted data unusable: ***Data collector required to implement separate query libraries, and then re-visit the source.***
Coding system variations	Again, vocabularies may vary between brands. Some systems impose their own local codes.	Especially in an Inernational study: ***Dedicated team required to work on documentation for specific countries/systems. Time constraints.***
Data architecture	Clinical problems are in many cases represented in different ways. This is even more relevant in an international level, where each implementation may be unique.	Analysis (cluster etc) difficult without introducing a data pre-processing overhead: ***Data Analyst first needs to translate data often with limited feedback.***
System architecture	Variations in the software and hardware being used (or the lack of expertise) can cause problems in the data entry and data extraction layers.	Slow, often limited extraction possibilities, choice of „cheap" system over complete functionality: ***Less data on available for analysis.***

Table 3 Linking data processing activities, traceability and curation

Category	Data type	Detail	Comment
Individuals	Unique Identifier (UID)	Single registration / ghosts	Defined denominator
Coding schema	Primary	Local codes	Version at time of study
	Secondary		
	Drug dictionary	Coding may include multiple generics	Include interaction schema
Type of data	Coded	Diagnostic	
		Other categories	
	Narrative	Structured	
		Free text	Rarely available
	Encounter	Type	
		Health Care Professional	May not have equivalence
Extraction queries	Data	All	
		Selective	Understanding search syntax is vital
	Date	Collection	
		Date range of query	
EPR vendor	Brand	Version	Change in EPR may lead to data loss
Practice	Characteristics	Teaching/ Lab links	Ethnicity & deprivation are determinants of health
	Limiting factors	New practice / turnover	

record linkage are being developed which can link data, where needed in the absence of strong identifiers [47].

Traceability

Traceability is defined as the ability to retain the identity of a product and its origin [48] and can be achieved in large datasets. Traceability of each data item requires tracing the individual who contributed the data and the variable that describes that individual. Within most clinical databases a patient's name will identify them, but within research databases usually pseudonyms are used which can only be decoded by the clinicians who provided the original data. The variables describing that individual will usually be in a tuple of code-date-value (e.g. 44p is the Read code for cholesterol; 26-Sep-2009 is the date the test was conducted; and 5.5 mmol/l is the result). However to interpret this datum we need to be able to link the data to the extraction query and data about when these data were extracted; for example did the query request the latest, the minimum or the maximum cholesterol value? The Primary Care Data Quality (PCDQ) programme developed its own metadata [43] to avoid data misinterpretation and to ensure traceability.

Metadata

Metadata is "Data that describes data," the study of metadata is part of the "Semantic web" which sets out to allow data to be shared and reused across applications; the first set of work in this area was the programme of "Resource Description Frameworks," the process of developing tagged information [49]. The PCDQ metadata just relates to the data source, the content and format of the data. Typically core metadata will contain: (1) Resource, (2) Summary content; (3) Format; and (4) Security descriptors; with additional extensible layers added as needed.
The sophistication of the metadata will define what links can be made from a datum to its source. Meta-data should

help flag related concepts, elements of the data model and help ensure semantic meaning. Provenance information is also often placed into the metadata, for example as XML representations of directed acyclic graphs tracing the origin of data. Meta-data is a key enabler to the emergence of quality measures that are socially constructed from within the community of users of medical data.

Few databases publish their metadata [50]; and if we fail to do this, data risks being misinterpreted when analyzed remote from people who understand the context in which it is recorded.

Curation of Data

The term curation was defined by Lord et al [51] as an activity that manages and promotes the use of data from its point of creation, ensuring that it is available for discovery and reuse, and fit for purpose. Curation is essential to provide a substrate to access, share and reuse data collections successfully [52]. Interpretation of data may require the simultaneous archiving of the metadata schema; look up tables for clinical codes and drug dictionaries (as without these it is impossible to know the extent of coding choices available to clinicians at the time); data extraction queries; and syntax or code which describe how data were cleaned and processed to create the final analysis variables.

In provenance-enabled systems, the full trace of curation is stored for future querying and analysis, thus enabling full reproducibility and verifiability of the data transformations performed. There may also be lessons for archiving health data from the metadata standards developed for archiving reference works: Reference model for an Open Archival System (OAIS) [53].

Simulation Using the TRANSFoRm Use-cases

We used the two TRANSFoRm use-cases to simulate ontological issues, scope of the data model and semantic meaning (Tables 4 and 5). The key ontological issues shared across both the use-cases are their ontological richness – or complexity. In both use-cases risk factors are common and complex; co-morbidities are common but not readily predictive; and involve multi-disciplinary care and records. The issues with the data model issues were the complexity of relevant data and scepticism about the reliability of summary data from the specialist repositories. The semantic challenges were largely about type and severity of disease.

Table 4 Overview of challenges in the genetic study of type 2 diabetes use-case

	Primary Care Data	Genetic database
Ontological issues	Complex relationship between risk factors and diagnosis: age, gender, ethnicity, obesity, confirmatory blood tests, gap between diagnosis and therapy commencing. Family history information can support and be supported by genetic tests - but need to address situations when genetic tests contradict the family history? What are the terminology and messaging standards used? How complex will an ontology describing a comprehensive multidisciplinary approach to diabetes be? Embedding a patient-centred approach will add to the complexity?	Complex multi-factorial relationship between genes and disease still poorly understood – GWAS are difficult to interpret. Contextual information and an accepted terminology set should make it easier e.g. a positive family history as reflected in a 2-3 generation pedigree/genogram will allow some assumptions and weighting to be done.
Data model	The data model and database architecture should enable the links between clinical practice, population health and research, evaluation and quality monitoring of the care provided. The data model should enable the assessment of data quality e.g. completeness of coded data, or access to other data.	What genetic information should be included in an EHR? For instance how detailed (granular) should the information be? How useful is the information derived from Single Nucleotide Polymorphism (SNPs) or other genetic tests?
Semantics	Adequate quality of data to confirm type and severity of diabetes, enabling the choice of appropriate therapy e.g. insulin therapy or lifestyle strategies to address obesity.	How consistent are genetic terms in themselves and in relation to phenotypic terminology? Consider tailored drug prescribing and personalised medicine? Can SNOMED CT deal with the link between primary care clinical terms and genetic terminology?

Table 5 Overview of challenges in gastro-oesophageal reflux disease (GORD) and oesophageal cancer use-case

	Primary Care Data	Cancer registry
Ontological issues	The ontology should recognise that GORD is very common and commonly self-treated with OTC medications, which can lead to under-diagnosis and under-recording. Co-morbidities which may be the reason for the use of anti-indigestion drugs add complexity to the ontology along with patient symptoms which may be more dyspeptic than reflux depending on the context. Need to determine the terminology and messaging standards used to deal with alarm symptoms. Define the optimum scope for the ontology for a comprehensive multidisciplinary approach to GORD	Barrett's oesophagus (pre-cancerous condition) is common and often symptomless so may be under reported.Oesophageal cancer is very rare and biopsy of Barrett's may miss it. Ontology to link alarm symptoms and other risk factors with cancer, including sensitivity and specificity? Need to identify the relevant and valid syntactical aspects of cancer concepts and their relationships.
Data model	The data model and database architecture should enable links between the use of OTC medications, alarm symptoms, relevant co-morbidities and quality of life. Data model should also enable clinical and population health research, including audit and quality monitoring of care.	Inclusion criteria for cancer registry? Case definition used by cancer registry?How should endoscopic and biopsy results be incorporated? - Primary record, extract from hospital record.
Semantics	Adequate quality of data to confirm type and severity of GORD, enabling the choice of appropriate therapy e.g. PPI or lifestyle strategies to address risk factors such as obesity, regular indigestion medicine use, etc. For example, how many prescriptions defines regular use?	Consistency of cancer terminology with primary care terminology. National and international criteria for bias.Insured population or other differential access to health care may bias results.

This simulation suggested that omission of data in a study either through not having access to a data source or through invisible exclusion criteria may be as important as knowing the quality of the data that is included (e.g. Over the counter (OTC) pharmacy data). Metadata constructed to support the provenance model, should contain links to the comprehensiveness of the data source as well as what data fields and data types (e.g. free-text) are included.

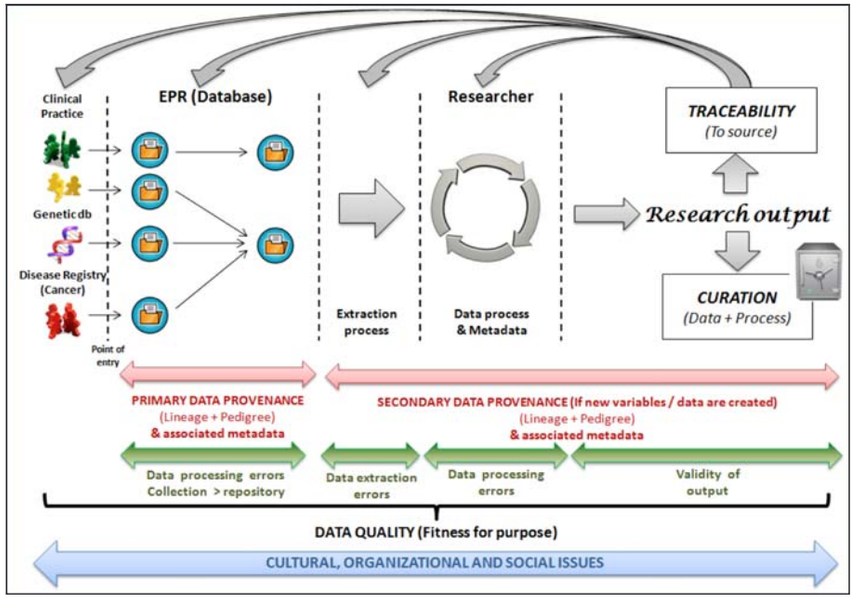

Fig. 3 An overview of the stages and linked quality concepts from data recording to extraction

4. Discussion

Principal findings:
"Black box" processing and reporting of findings based on routine data should no longer be acceptable. More explicit and structured descriptions of the origins of data, and wherever possible the use of open standards should be mandated for studies based on routine data. Auditing the whole process, from data recording to curation, is critical to ensure data quality (Fig 3) in any final published output. Validation of data is impossible unless the provenance of the data, extraction and processing errors are recorded in a structured way and each cell in the final analysis table is traceable. We have separated the primary data provenance, capturing the origin of data records, from the secondary data provenance, tracing the operations performed by the researcher, as the underlying causation is likely to be different. A shared understanding of the relevant ontology, data model and semantic issues are essential, and should be conducted on a study-by-study basis. This output feeds into the metadata schema, which should be published to demonstrate that these processes handle data in a consistent and reliable way; and also what data are not included within a study.

Implications of the findings:
A common set of descriptors of the process from data recording through to the curation of data should be adopted and published as an appendix to studies using routine data as it will improve the ability of researchers to compare data processing methods and understand where data losses may occur. This will be increasingly important for studies involving linked data. We propose that provenance, data extraction and processing errors and curation are used to describe the issues related to the processing of data to produce research outputs and its subsequent archiving respectively. The term traceability should be reserved for the retrospective audit of the data within the final research output. We have made this term the functional aspect of provenance: one of its purposes is to enable traceability. Exploring for each study its ontology, data model and semantic issues will help ensure that the metadata schema meets the needs of the project as well as describing the data quality.

Comparison with the literature:
Central to our study is a belief in an open systems approach to defining and assessing data quality, in keeping with the Toyota Production System [54]. One which is essentially socio-technical, and is efficient not overburdening the research process, or allow inconsistency.

Much pharmaceutical research uses the methods set out within the Clinical Data Interchange Standards Consortium (CDISC) including the Biomedical Research Integrated Domain Group (BRIDG)[55] - an internationally recognized standards body [56]. CDISC has also developed a relationship with health level seven (HL-7) which has adopted the BRIDG data analysis model (DAM). In clinical trials the data set is generally complete and there are less challenges in linking records and managing incomplete datasets [57,58].

Statistical process control techniques may provide better mechanisms for the

direct exploration of clinical data, with much less collection and processing overheads [59,60].

Limitations of the method:
This schema is based on experiential learning from those involved in data processing and a literature review; expert consensus of this sort forms the lowest grade of evidence [61]. Prospective studies have not yet tested whether these elements are essential.

Call for further research:
We need to test different approaches and strategies as new web-technologies unfold. The recent emergence of "tagging" in various social computing sites provides the opportunity for enriching meta-data. In the broad context of use, we see this as being used to help users provide meta-data on relevance and quality. In social computing, many of the terms are weak in information content. Allowing the medical community to tag data, or sources of data, may enable the emergence of a socially constructed quality model.

5. Conclusions

If we accept fitness for purpose as the central feature of data quality, then it is essential that our model of data quality be constructed in a way that represents consensus as to best practice amongst its community of users. Consistent description of the process will improve understanding of the validity of research findings based on routinely collected data, and this description should be formalised in the metadata schema. The process from data recording to research output is complex but can be represented as a simple linear model. Prior to embarking on research using routine data investigators should carefully map the process from data recording to curation. A schema of the research process including provenance of the data and the details of data extraction and processing should be developed as a check list and be available for all publications based on routine data.

Acknowledgements
Frank Sullivan and Mark McGilchrist for their comments on the manuscript; IMIA and EFMI for supporting their primary care informatics working groups.
TRANSFoRm is supported by the European Commission - DG INFSO (FP7 2477)

Conflict of interest
None declared.
SdeL, MTV, GM, LA, PL, FS and MMcG are investigators on the TRANSFoRm project.

References

1. Peterson K. Practice-based primary care research—translating research into practice through advanced technology. Family Practice 2006;23:149–50.
2. 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.
3. Hummers-Pradier E, Scheidt-Nave C, Martin H, Heinemann S, Kochen MM, Himmel W. Simply no time? Barriers to GPs' participation in primary health care research. Fam Pract 2008;25(2):105-12.
4. Translational Medicine and Patient Safety in Europe (TRANSFoRm). URL: http://www.transformproject.eu/
5. International Medical Informatics Association (IMIA). Primary Health Care Informatics Working Group. URL: http://www.imia-medinfo.org/new2/
6. European Federation for Medical Informatics (EFMI) Primary Care Informatics Working Group (PCI WG) URL: http://www.efmi.org/
7. de Lusignan S, Pearce C, Shaw N, Liaw ST, Michalakidis G, Vicente M, Bainbridge M. What are the barriers to conducting international research using routinely collected primary care data. Stud Health Technol Inform 2011;165:135-40. DOI 103233/978-1-60750-735-2-135
8. 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;76(4):261-8.
9. The International Standards Organization (ISO). 8402-1986 Quality Vocabulary. URL: http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=15570
10. de Lusignan S. The optimum granularity for coding diagnostic data in primary care: report of a workshop of the EFMI Primary Care Informatics Working Group at MIE 2005. Informatics in Primary Care 2006;14:133–7
11. Pringle M, Ward P, Chilvers C. Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer. Br J Gen Pract 1995;45(399):537-41.
12. Williams JG. Measuring the completeness and currency of codified clinical information. Methods Inf Med 2003;42(4):482-8.
13. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003;326(7398):1070.
14. Roten I, Marty S, Beney J. Electronic screening of medical records to detect inpatients at risk of drug-related problems. Pharm World Sci 2010;32(1):103-7
15. Aqil A, Lippeveld T, Hozumi D. PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. Health Policy and Planning 2009;24:217–228
16. Arts DGT, Keizer NF, Scheffer GJ. Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. J Am Med Inform Assoc 2002;9:600–11
17. Kushniruk A, Borycki E, Kuwata S, Kannry J. Predicting changes in workflow resulting from healthcare information systems: ensuring the safety of healthcare. Healthc Q 2006 Oct;9 Spec No:114-8.
18. Tai TW, Anandarajah S, Dhoul N, de Lusignan S. Variation in clinical coding lists in UK general practice: a barrier to consistent data entry? Inform Prim Care 2007;15(3):143-50.
19. Debar S, Kumarapeli P, Kaski JC, de Lusignan S. Addressing modifiable risk factors for coronary heart disease in primary care: an evidence-base lost in translation. Fam Pract 2010 Aug;27(4):370-8.
20. Zdun U. Semantic Lookup in Service-Oriented Architectures. Proceedings of Fourth International Workshop on Web-Oriented Software Technologies 2004: 101-10. URL: http://eprints.cs.univie.ac.at/2797/1/lookup.pdf
21. Pan J, Chen K, Hsu W. Self Risk Assessment and Monitoring for Cardiovascular Disease Patients Based on Service-Oriented Architecture. Computers in Cardiology 2008;35:637-40.
22. Turbelin C, Boëlle PY. Improving general practice based epidemiologic surveillance using desktop clients: the French Sentinel Network experience. Stud Health Technol Inform 2010;160(Pt 1):442-6.
23. de Lusignan S, Stephens PN, Adal N, Majeed A. Does feedback improve the quality of computerized medical records in primary care? J Am Med Inform Assoc 2002;9(4):395-401.
24. de Lusignan S, Mimnagh C. Breaking the first law of informatics: the Quality and Outcomes Framework (QOF) in the dock. Inform Prim Care 2006;14(3):153-6.
25. de Lusignan S, Khunti K, Belsey J, Hattersley A, van Vlymen J, Gallagher H, et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. Diabet Med 2010 Feb;27(2):203-9.
26. Carey IM, Cook DG, De Wilde S, Bremner SA, Richards N, Caine S, Strachan DP, Hilton SR.

Implications of the problem orientated medical record (POMR) for research using electronic GP databases: a comparison of the Doctors Independent Network Database (DIN) and the General Practice Research Database (GPRD). BMC Fam Pract 2003;4:14.
27. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care—a systematic review. Fam Pract 2008;25(6):400-13.
28. Lanter D. Design of a Lineage-Based Meta-Data Base for GIS. Cartography and Geographic Information Systems1991;18(4):255-61.
29. Yamamoto S. Reconstructing data-flow diagrams from structure charts based on the input and output relationship. IEICE Transactions on Information and Systems 1995;e78d(9):1118-26.
30. Moellman D, Cain J. Intelligence, mapping and geospatial exploitation system (IMAGES). Proceedings of Digitization of the Battlespace III 1998;3393:86-95.
31. Buneman P, Khanna S, Tan WC. Data provenance: Some basic issues. FST TCS 2000: Proceedings 2000;1974:87-93.
32. Cheney J, Chiticariu L, Tan WC. Provenance in Databases: Why, How and Where. Foundations and Trends in Databases 2007;1(4):379-474.
33. Chief Information Officer (CIO). Net-centric data strategy. Washington DC; Department of Defense, 2003. URL: http://cio-nii.defense.gov/docs/net-centric-data-strategy-2003-05-092.pdf
34. Beresford NA, Broadley MR, Howard BJ, Barnett CL, White PJ. Estimating radionuclide transfer to wild species—data requirements and availability for terrestrial ecosystems. J Radiol Prot 2004 Dec;24(4A):A89-103.
35. Simmhan YL, Plale B, Gannon D. A Survey of Data Provenance in e-Science. SIGMOD Record 2005;34(3):31-6.
36. Lee ES, McDonald DW, Anderson N, Tarczy-Hornoch P. Incorporating collaboratory concepts into informatics in support of translational interdisciplinary biomedical research. Int J Med Inform 2009 January;78(1):10–21.
37. Groth P, Munroe S, Miles S, Moreau L. Applying the Provenance Data Model to a Bioinformatics Case. 2008, URL: http://www.mendeley.com/profiles/paul-groth/document/861379562/#highlighted
38. Glavic B, Dittrich K. Data Provenance: A Categorization of Existing Approaches. URL: http://subs.emis.de/LNI/Proceedings/Proceedings103/giproc-103-014.pdf
39. Goble C. "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics," in Workshop on Data Derivation and Provenance, Chicago; 2002.
40. Moreau L, Clifford B, Freire J, Futrelle J, Gil Y, Groth P, et al. The open provenance model core specification (v1.1). Future Generation Computer Systems; July 2010.
41. Groth P, Luck M, Moreau L. A protocol for recording provenance in service-oriented grids. In: Ed, Higashino T. Lecture Notes in Computer Science. Proceedings of the 8th International Conference on Principles of Distributed Systems (OPODIS'04), Grenoble, France. Springer-Verlag; Berlin,2005;3544: 124-39. DOI: 10.1007/b138689
42. 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 and quality improvement. Stud Health Technol Inform 2010; 160(Pt 1):724-8.
43. 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;116:1010–5.
44. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidemiol Drug Saf 2010;19(6):618-26.
45. Davis P, Jenkin G, Coope P, Blakely T, Sporle A, Kiro C. The New Zealand Socio-economic Index of Occupational Status: methodological revision and imputation for missing data. Aust N Z J Public Health 2004;28(2):113-9.
46. 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. Informatics in Primary Care 2005;13:281-91.
47. Durham E, Xue Y, Kantarcioglu M, Malin B. Private medical record linkage with approximate matching. AMIA Annu Symp Proc. 2010 Nov 13; 2010:182-6.
48. Khabbazi MR, Yusof Ismail MD, Ismail N, Mousavi AS. Modeling of Traceability Information System for Material Flow Control Data. Australian Journal of Basic and Applied Sciences 2010;4(2):208-16.
49. World Wide Web Consortium (W3C). Technology and Science Domain: Metadata and Resource Description. URL: http://www.w3.org/Metadata/
50. 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. Informatics in Primary Care 2005;13:281-91.
51. Lord P, Macdonald A, Lyon L, Giarretta D. From Data Deluge to Data Curation. In: Proceedings of the UK e-science All Hands meeting 2004: 371–5.
52. Karasti H, Baker KS, Halkola E. Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. Computer Supported Cooperative Work 2006,15:321–58.
53. Lavoi F. The Open Archival Information System Reference Model: Introductory Guide Microform and Imaging Review. Spring 2004;33(2):68–81 DOI: 10.1515/MFIR.2004.68,
54. Seddon, J. Systems Thinking in the Public Sector. Triarchy Press; 2008.
55. Biomedical Research Integrated Domain Group (BRIDG). URL: http://www.cdisc.org/bridg
56. Clinical Data Interchange Standards Consortium (CDISC) URL: http://www.cdisc.org
57. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, Brand CA. Data Linkage: A powerful research tool with potential problems. BMC Health Serv Res 2010;10:346.
58. Nur U, Shack LG, Rachet B, Carpenter JR, Coleman MP. Modelling relative survival in the presence of incomplete data: a tutorial. Int J Epidemiol 2010;39(1):118-28.
59. Abdel Wahab MM, Nofal LM, Guirguis WW, Mahdy NH. Statistical process control for referrals by general practitioner at Health Insurance Organization clinics in Alexandria. J Egypt Public Health Assoc 2004;79(5-6):415-48.
60. Aylin P, Best N, Bottle A, Marshall C. Following Shipman: a pilot system for monitoring mortality rates in primary care. Lancet 2003; 362(9382):485-91.
61. University of Oxford. Centre for Evidence Based Medicine - Levels of Evidence. 2009. URL: http://www.cebm.net/index.aspx?o=1025

Correspondence to:
Simon de Lusignan
Clinical Informatics
Department of Health Care Management and Policay
School of Management,
University of Surrey
GUILDFORD GU2 7XH
Tel: +44 1483 683089
Fax: +44 1483 301132
E-mail: s.lusignan@surrey.ac.uk
Web: www.clininf.eu