MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010
© 2010 IMIA and SAHIA. All rights reserved.
doi:10.3233/978-1-60750-588-4-724

A system for solution-orientated reporting of errors associated with the extraction of routinely collected clinical data for research and quality improvement

Georgios Michalakidis[a,b], Pushpa Kumarapeli[b], Andre Ring[b], Jeremy van Vlymen[b], Paul Krause[a], Simon de Lusignan[b]

[a] *Department of Computing, University of Surrey, Guildford, UK*
[b] *Primary Care Informatics, Division of Population Health Sciences and Education, St. George's - University of London, London, UK*

Abstract

Background: We have used routinely collected clinical data in epidemiological and quality improvement research for over 10 years. We extract, pseudonymise and link data from heterogeneous distributed databases; inevitably encountering errors and problems. Objective: To develop a solution-orientated system of error reporting which enables appropriate corrective action. Method: Review of the 94 errors, which occurred in 2008/9. Previously we had described failures in terms of the data missing from our response files; however this provided little information about causation. We therefore developed a taxonomy based on the IT component limiting data extraction. Results: Our final taxonomy categorised errors as: (A) Data extraction Method and Process; (B) Translation Layer and Proxy Specification; (C) Shape and Complexity of the Original Schema; (D) Communication and System (mainly Software-based) Faults; (E) Hardware and Infrastructure; (F) Generic/Uncategorised and/or Human Errors. We found 79 distinct errors among the 94 reported; and the categories were generally predictive of the time needed to develop fixes. Conclusions: A systematic approach to errors and linking them to problem solving has improved project efficiency and enabled us to better predict any associated delays.

Keywords:

Computerized medical records system, Data quality, Databases, Semantics, Controlled vocabulary, Computers.

Introduction

Internationally, routinely collected clinical data from primary care electronic patient record systems (EPR) is used for research and quality improvement [1]. However, many of the research databases only draw their data from a single vendor; thereby circumventing many of the difficulties due to variation in the way that national and international standards are implemented in different brands of EPR system. By way of contrast the Primary Care Data Quality (PCDQ) programme has worked with large datasets drawn from different vendors some using different classification systems and generating their own local codes to supplement standard taxonomies: truly heterogeneous distributed databases [2]. The largest PCDQ study drew data from 2.4 million patient records [3]; current studies work with databases of just under 1 million records drawn from six different brands of EPR system but extracting several hundred variables [4].

The difficulties in extracting data from heterogeneous distributed sources are well known [5] and standard methods and toolkits for measuring the validity and utility of electronic patient record systems have been proposed [6]. The difficulties 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. None of the classification systems within the UK have definitions, making it possible for meaning to vary between professional groups and over time [7]. There may be a trade-off for vendors between achieving functionality and strict adherence to guidance on requirements of EPR systems. Difficulties can then arise because standards are not strictly implemented.

We frequently encountered data extraction problems, which we historically reported in terms of what data were missing from our response files rather linking our problem to the underlying cause. We carried out this study to see if we could develop a system of solution-orientated error reporting, which improved our problem solving and predicted likely time to resolution.

Method

We carried out a literature review on the standard bibliographic databases to identify structured approaches to data extraction techniques from heterogeneous distributed databases. We narrowed our search to clinical data and databases. We tried to identify any current approaches to identifying and resolving data extraction errors apply prospective and retrospective statistics with dynamic validation as data are being extracted [8].

Our literature review initially focussed on the UK National Health Service methods of date extraction. We also looked for proprietary tools for data extraction provided by EPR vendors. We found two generic approaches to enable enquirers to execute queries and extract data from different types of general practice computer systems using a common query language. The two alternatives are the MIQUEST (Morbidity Information and Export Syntax) HQL (Health Query Language) data extraction tool and the proprietary Apollo SQL interface. PCDQ uses the MIQUEST data extraction tool.

We then collected information about the experiences of five data collectors during 2008 and 2009. Interviews with the collectors, observations, documentation reviews and test data extraction queries were used as information gathering techniques. The study exports included detailed descriptions of errors, frequencies of occurrence and comments on specific issues, system installations and system versions. A post-hoc exploration process was carried out.

We constructed an initial list of the errors encountered and classified the errors according to their effect on the data collection process. This included; information about whether the query would execute at; the stage each query would run to, and whether it produced partial or no results at all. We then reviewed and categorised the observed data collection issues based on their impact.

We subsequently redefined the data collection issues to create a briefer but nevertheless descriptive functional IT component approach to facilitate problem solving.

Finally, we reclassified the improved list of errors based on our need for a system which enabled understanding of whether certain groups of errors could be resolved by the collection team or not; and finally whether they were vendor specific. This led to the creation of a taxonomy of errors for our data collection problems.

We then created an on-line resource for the PCDQ data collectors. The on-line problem reporting form was designed to capture the key information needed to diagnose the IT component which was responsible for any errors.

The studies carried out during the development of our taxonomy were ethically approved, and only used pseudonymised data.

Results

Errors during the Data Collection Process

We identified 94 problems with the data extraction from the four major UK GP electronic patient record (EPR) suppliers: EMIS PCS, EMIS LV, INPS Vision and iSOFT Synergy. These four EPR systems account for over 90% of GP EPR market for England [9].

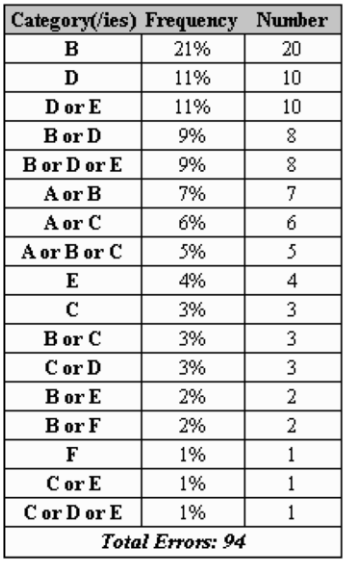
Problems of different levels of severity and impact were identified and initially mapped to a set of groups in a purely clinical use driven approach. The frequency of the type of problems encountered is summarised in Table 1. Errors were reported in the following categories: (A) MIQUEST – the data extraction tool did not work, or the query code failed at some point, (B) The MIQUEST specification was differently implemented on one of the brands of EPR systems. For example the word "CHOSEN" returns different response files, (C) Clinical System and database would not return information, (D) Supporting Software and operating system, (E) Hardware and Infrastructure, (F) Generic/Uncategorised and/or Human Errors.

We documented the cumulative frequencies for each issue based on each individual collector's recordings and assigned them to any category they applied to.

Multiple mappings

This type of direct mapping did not allow for an optimised approach to error solving. We found that 56 of the 94 errors (60%) could be assigned to multiple categories as shown in Figure 1; a limitation of our initial categorised reporting.

Table 1 - Stage Process Categorisation Frequencies



Category(/ies)	Frequency	Number
B	21%	20
D	11%	10
D or E	11%	10
B or D	9%	8
B or D or E	9%	8
A or B	7%	7
A or C	6%	6
A or B or C	5%	5
E	4%	4
C	3%	3
B or C	3%	3
C or D	3%	3
B or E	2%	2
B or F	2%	2
F	1%	1
C or E	1%	1
C or D or E	1%	1
Total Errors: 94		

Post processing of the error list

We initiated another round of result interpretation to identify the generic functional IT standings of each specific issue and generate a list with strictly targeted errors that could be classified to one of our newly defined categories.

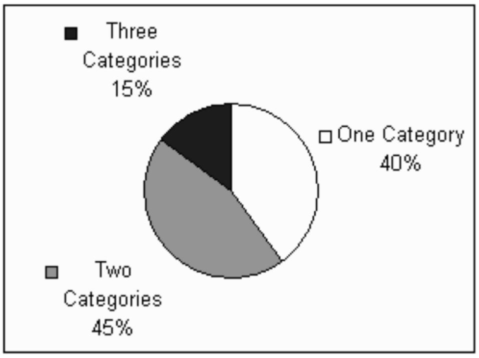


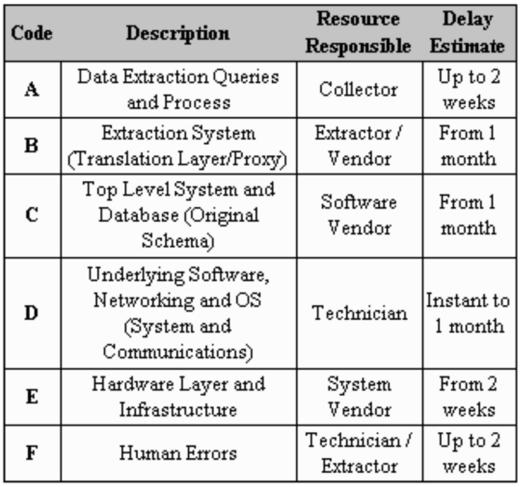
Figure 1 - Frequencies for Multiple Categories

Taxonomy of errors

We then investigated whether the errors identified could be mapped to a single root cause – recognising that this may involve careful analysis as to which IT component was responsible for the error encountered. We connected each of the redefined issues to the respective group described in Table 2. The new definition allowed for identifying the point of contact as well as an estimate on the delay for solving individual cases.

The list was structured from the data collectors' experiences in the field, response time average and contact points.

Table 2 - Component Oriented Approach



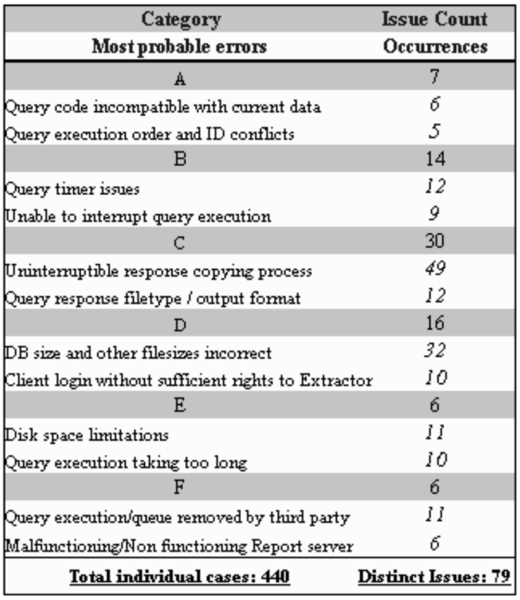
Code	Description	Resource Responsible	Delay Estimate
A	Data Extraction Queries and Process	Collector	Up to 2 weeks
B	Extraction System (Translation Layer/Proxy)	Extractor / Vendor	From 1 month
C	Top Level System and Database (Original Schema)	Software Vendor	From 1 month
D	Underlying Software, Networking and OS (System and Communications)	Technician	Instant to 1 month
E	Hardware Layer and Infrastructure	System Vendor	From 2 weeks
F	Human Errors	Technician / Extractor	Up to 2 weeks

We completed our direct mapping of each individual case to one category through several iterations; based on comments and feedback from collectors, IT resources and documented processes. This enabled us to summarise all the errors into the categories and export metrics as in Table 3.

This process, involved merging a total of 15 error descriptions with closely related ones. This created an accurate assignment to the various categories of secondary level processing.

Of the 79 final individual error types, category C (Clinical EPR System) was most frequent and found in 30 of these 79 problem categories – requiring input from the vendor. This was followed by D (Network and Operating System) and B (Data Extractor Specification) with 16 and 14 occurrences respectively. Categories A (Query Content), E (Hardware) and F (Other Human Errors) had 6 and 7 error types recorded for each. Around 40% (30/79) of the problems could be solved by phoning software support for the particular brand of EPR system, 20% by actions from the data collector, and 10% by senior team member. 30% could not be resolved by the team in the expected timeframe and required some external input. The majority involved external resources and required changes to the underlying hardware or technical intervention with the on-site systems; albeit usually delivered remotely by the vendor.

Table 3 - Taxonomy of Errors and Frequent Cases



Category	Issue Count
Most probable errors	**Occurrences**
A	7
Query code incompatible with current data	6
Query execution order and ID conflicts	5
B	14
Query timer issues	12
Unable to interrupt query execution	9
C	30
Uninterruptible response copying process	49
Query response filetype / output format	12
D	16
DB size and other filesizes incorrect	32
Client login without sufficient rights to Extractor	10
E	6
Disk space limitations	11
Query execution taking too long	10
F	6
Query execution/queue removed by third party	11
Malfunctioning/Non functioning Report server	6
Total individual cases: 440	**Distinct Issues: 79**

An online structured system for the initial reporting of errors was implemented. We used a project management platform and divided the tracking mechanism into several sub-systems for different projects (QICKD [4], IAPT (an evaluation of Improved Access to Psychological Therapies), Osteoporosis, and Diabetes studies) but with the same underlying format for cross reporting. A set of optional and required fields, allowed for immediate connection to a predefined taxonomy.

Practical examples of errors and solution orientated reporting

We report exemplar errors in each of the seven areas and the action taken and time taken to solve them:

(a) Data extraction queries and process problems: One brand of EPR vendor creates its own local codes to plug what it perceives as gaps in the standard (Read code) hierarchy. Unless queries collect these local codes this data area is deficient. Smoking data provides a good example - where sometimes local codes have accounted for >10% of the data. The solution took under two weeks, involving query re-writing and exploring with clinic doctors how these patients were represented in the EPR system.

(b) Extraction system errors: The INPS Vision and EMIS PCS systems can take several minutes to extract large response files during which the session cannot be interrupted. We prevented this by logging in another session on the same workstation on the proviso that login details are held. We asked that these details be provided so that restarting the extraction would require less time. Interference from support staff would often take hours or require a revisit.

(c) Top level system and database errors: Systems are not designed to handle large-size result files. EMIS LV in particular, is negligent for not having sufficient free database and disk space which is an essential prerequisite for data extraction. This often resulted in system crashes if it was utilised for clinical purposes at the same time. We made sure at least 100MBytes are free, to avoid interruptions and system restarts that would take 30 minutes or more.

(d) System and communication errors: Due to insufficient user privileges, execution of queries is sometimes impossible. We used accounts with >Level 4 access to MIQUEST. Also, the interface would sometimes incorrectly indicate the current load. We were proactively ensuring that we had enough resources than the stated, to prevent having the support staff sort it for us with a 15 minute to 1 hour addition to the collection.

(e) Hardware and infrastructure errors: iSOFT Synergy and Premiere are both vulnerable to complex query sets when an external reporting server is not being used. We found that even in cases where a reporting server was installed, we had to use the live clinical system because of poor maintenance and data that were not up-to-date or incorrectly linked.

(f) Human errors: Experience has shown that human errors were also commonplace even in some instances where communication had been successful. On occasion, queries had been removed from the system by practice staff unknowingly, where researchers had scheduled set execution times or where practice staff needed to execute internal queries or maintenance tasks (the data extraction tasks get lower priority than the system maintenance processes).

Discussion

Principal findings

Data extraction techniques are widely used to answer research questions from routinely collected clinical data. Problems are faced during data extraction. Most appear to be associated with the way specific data extraction engines have been implemented by the different EPR system vendors. The adoption of this error taxonomy would enable consistent reporting to system manufacturers and potentially improved efficiency in data collecting.

Implications of findings

Our system imposed a much more analytical approach to error reporting and handling from data collectors. Errors were assigned to all categories. Although the incidents were not equally distributed, they were correctly proportioned based on the influence of each individual category (and system element) to the overall extraction and anonymised patient record collection process. The process allowed us to be more accurate, fast and proactive. For example, the common error of queries being randomly interrupted was originally handled as a query writing issue. Statistically, and based on our classification and findings, there are less chances of this error feedback because of a syntax or vocabulary error and the failure output is almost directly connected to automated backup processes on the reporting server or other maintenance issues. On another example, the traditional way of handling long-running queries was to generate subsets even in cases where the practice systems had a specific (in most cases easy to resolve) issue. With our process, we would classify the error, check the most probable causation and proceed with the collection without rescheduling a visit which involved writing vendor-specific subset queries in-between, using valuable study time and resources. Also, the taxonomy helped us provide feedback to the EPR vendors (and MIQUEST) as in a recent example, new data together with the inclusion of post-collection validation error information pointed to a vendor bug with a false return of text data type ACR values instead of numeric ones, resulting in gaps inside the collected data. This was flagged, the developers were notified and a fix was introduced prior to our next collection. We found that the duration of this process was less than our estimate for rewriting the queries, deploying and executing them as well as any changes to our analysed flat-file generation mechanisms for converting the inconsistent data type.

We followed a set of principles for our own research projects and found that early steps and precautions also allowed for minimised error frequency. For example, the use of mechanical processes for query writing, code execution and system maintenance minimised errors in categories A and F (we followed the commonly used reusable code principle via component based engineering for our processes), whereas documented solutions to usual problems on software-level (for categories B, C and D) allowed for error solving by the collectors themselves. Finally, articulating the least acceptable hardware features and specifications before a data provider joined a study minimised the impact of any issues with the infrastructure (category E).

The error classification through the taxonomy we implemented makes error reporting a solution-orientated approach. It allows for the flagging of the nature of any obstacles combining precautions as well as immediate action thenceforth. Problems which were frequent 12 months ago no longer feature on our error list, these include but are not limited to: Some human errors, disk space issues where feedback would be inaccurate, crashes on the server because of lengthy or problematic queries and shared folder issues where the mapping of local drives needs to be set up by staff with sufficient privileges.

This process provided team members the confidence to approach vendors or explain the limitations of hardware or software to healthcare providers participating in research.

Comparison with literature

We identified a dearth of literature on error reporting. Though much is written about how data from EPR systems are expected to have a central role within healthcare commissioning, and quality improvement [10].

There are generic IT approaches to problems with data extraction: namely the resolution of possible data conflicts occurring in the database integration process; incompatibilities between databases, differences in data types; and copies of the same information stored in different databases [11]. There are examples of logging mechanisms able to identify errors either in extraction itself or the underlying EPR system data (for example, the miscoding of family history as heart disease resulting in apparently 25% of practice population as having this diagnosis) [12]. The above reveal the need for a structured error handling process. The literature recognises the problems with heterogeneous distributed databases as well as the cost and effort for overcoming them without a standard well-defined and designed process [5,11].

The UK national data quality programme PRIMIS+ discussion board illustrates how data extraction problems extend widely; but does not incorporate any sort of error taxonomy [13].

Limitations of method

We found that on rare occasions the translation of an issue that can be connected to multiple categories has a slight change on its meaning (not on its effect, however) when redefined for our functional approach.

Also, for a number of problems we had to analyse feedback from the data collectors several times, in order to define the most appropriate category definition and the course of action that required the least effort (in terms of man-hours or external resources involved).

Based on our need for immediate logging of the problems and comments about them, we had to update the list as the extractions progressed and collections were rescheduled.

The design of the online system for error reporting allowed us to propagate knowledge on the effect and effort in resolving issues across several projects by setting principles and user-access rights for different teams in several locations connected to the same central repository. An approach that can widely be used for cross-platform access with direct assignees, minimising the time spend for both administering and providing solutions in timely manner.

Call for further research

Based on our findings, the implementation of a shared reporting system adopted by the individual EPR system vendors can help the secondary use of routinely collected data by allowing for faster solutions and therefore interpretation and use of the output data.

Conclusion

This approach has enabled us to achieve higher levels of successful problem reporting and solving. Its method could be replicated in other projects. We recommend the use of this taxonomy for error reporting for any type of study whether using primary or secondary care data.

System vendors should be more aware of the potential impact of non-standard interfaces. Better structured systems, which more strictly implemented standards, would reduce the time spent in crisis managing problems when extracting data. A national or international system of error reporting would allow sharing of workarounds is urgently needed. Adoption of a standardised method of solution-orientated error reporting would help EPR vendors identify and address errors in data extraction from their systems.

Acknowledgments

Biomedical Informatics students Saqba Arif and Shaista Somani for their experiences during the data collection process.

References

[1] 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.

[2] de Lusignan S, Hague N, van Vlymen J, Kumarapeli P. Routinely-collected general practice data are complex, but with systematic processing can be used for quality improvement and research. Inf Prim Care. 2006;14(1):59-66.

[3] de Lusignan S. An educational intervention, involving feedback of routinely collected computer data, to improve cardiovascular disease management in UK primary care. Methods Inf Med. 2007;46(1):57-62.

[4] de Lusignan S, Gallagher H, Chan T, Thomas N, van Vlymen J, Nation M, Jain N, Tahir A, du Bois E, Crinson I, Hague N, Reid F, Harris K. The QICKD study protocol: a cluster randomised trial to compare quality improvement interventions to lower systolic BP in chronic kidney disease (CKD) in primary care. Impl Sci. 2009 Jul 14;4:39.

[5] Elmagarmid A, Rusinkiewicz M, Sheth A (Eds). Management of Heterogeneous and Autonomous Database Systems. San Francisco; Morgan Kaufmann, 1998.

[6] Hassey A, Gerrett D, Wilson A. A survey of validity and utility of electronic patient records in a general practice. Fisher Medical Centre, Millfields, Skipton. BMJ. 2001 Jun 9;322(7299):1401-5.

[7] de Lusignan S. Codes, classifications, terminologies, and nomenclatures: definition, development and application in practice: a theme of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCI WG). Inform Prim Care 2005;13(1):65-69.

[8] Tatum J T, Wilkinson IV C W, Jannarone R J. Automatic data extraction, error correction and forecasting system. Netuitive Inc, Alpharetta, GA. US Patent 6591255.

[9] Moulene MV, de Lusignan S, Freeman G, van Vlymen J, Sheeler I, Singleton A, Kumarapeli P. Assessing the Impact of Recording Quality Target Data on the GP Consultation Using Multi-Channel Video. MedInfo 2007.

[10] Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. Fisher Medical Centre, Millfields, Skipton. BMJ 2003;326:1070.

[11] Madhavaram M, Ali DL, Zhou M. Integrating Heterogeneous Distributed Database System. Dept of Computer Science, University of Southern Mississippi, Computers ind. Engng Vol. 31, No. 1/2, pp. 315-318, 1996.

[12] Oxfordshire MAAG. Case study: a review of the Oxfordshire scheme. Collection of health data from general practice. Oxford: PC Inform Services (PRIMIS), 2000.

[13] University of Nottingham. PRIMIS+ Discussion Board. URL: http://forum.primis.nottingham.ac.uk/

Address for correspondence

Simon de Lusignan - Reader in General Practice&Informatics
Division of Population Health Sciences and Education
St. George's - University of London, SW17 0RE, UK
Tel: +44(0)20 8725 56661 Email: slusigna@sgul.ac.uk