Ensuring the Quality of Aggregated General Practice Data: Lessons from the Primary Care Data Quality Programme (PCDQ)

Jeremy van Vlymen, Simon de Lusignan, Nigel Hague, Tom Chan, Billy Dzregah

St. George's - University of London, UK

Abstract

Background: There are large numbers of schemes that collect and aggregate data from primary care computer systems into large databases. These data are then used for market and academic research. How the data is aggregated, cleaned and processed is usually opaque. Making the method transparent allows researchers to compare methods, and users of the output to better understand the strengths and weaknesses of the data.

Objectives: To define the stages of the process of aggregating, processing and cleaning clinical data from multiple data sources.

Methods: Identify errors in design, collection, staging, integration and analysis.


Conclusions: This eight step method provides a taxonomy to enable researchers to compare their methods of data process and aggregation.

Keywords: Medical Informatics, Databases, Primary Medical Care, Computers, Data collection, Computers, Vocabulary controlled

1. Introduction

Computer data routinely collected as part of primary medical care are widely used for research; and the volume of activity in England can be gained from the DocDat website (Directory of Clinical Databases) [1]. There are two methods used to access clinical data for audit, research or other purposes. Either all the data from the general practice computer system is downloaded and analysis conducted on a copy of the whole database (table 1). Two of these extract data from Torex-Isoft [2,3] computer systems; and one is from “In-Practice Systems” IPS) [4]. There are also two new data repositories: one extracts data from IPS [5] the other from EMIS [6]. The collection of data from practices using one manufacturer’s computer system avoids some of the inconsistency problems associated with aggregating data from different systems; and, studies have shown little difference is found when systems are compared [7,8]. These large databases aim to provide a national sample. Other investigators collect data on a more ad hoc basis; extracting the data needed for a particular audit or study. They collect data from across...
the range of different general practice computer systems; usually collecting a specific data set across one or more localities. The commonest tool used for data extraction is a UK Department of Health sponsored tool called MIQUEST (Morbidity Information Query and Export Syntax) [9].

Table 1: Large databases aggregating primary care data in England and Wales

<table>
<thead>
<tr>
<th>Database</th>
<th>Reference providing information about the database:</th>
<th>Computer system</th>
<th>URL</th>
</tr>
</thead>
</table>

As the UK uses a single terminology to record structured data, the Read “codes”; in theory the extraction and aggregation of data should be straightforward. In reality things are far more complex. The range of clinical computer systems offer different user interfaces and will inevitably bias what data is collected and the nature of input errors. The terminology used to code the structured data within the computer system is undergoing constant modification. The early systems contributing to the General Practice Research Database (GPRD) used Oxford Medical Information Systems Codes (OXMIS), the EMIS [10] computer system has many of its own codes [11]. Practitioners using the Read Code system have migrated from version 1, to a 4-character version of version 2, to a 5-character version [12]; and some have now moved on to Version 3: Clinical Terms (CT v3) [13]. To these challenges have to be added the inherent problems associated with data warehousing and quality assurance [14]. Finally, there can be problems associated with the analysis such as sample attrition. Currently, little is written about the process of aggregating, processing and cleaning computer data within the medical literature; and the lessons learnt from this data may have relevance internationally to those involved in processing primary care data drawn from different sources. This paper describes the method used within the primary care data quality project (PCDQ) with the aim of promoting transparency and improving the quality of this process.

2. Method

We reviewed our process and errors that had occurred over the last eight years using eight categories: design, data entry, extraction, migration, integration, cleaning, processing and analysis. This taxonomy is adapted from an error classification published by Berndt et al. [14] who developed this in the context of quality assurance of the healthcare data warehouse. Successful or good quality output is defined using the definition of data quality used in total data quality management (TQDM) as data fit for purpose by its consumers [15]: i.e. in our case, data useable as an educational intervention to improve chronic disease management, to improve the health of populations and for research.
Design: The design process is driven by the purpose of the study or audit we are involved in.

Step 1: Defining the research question or audit criteria for the study.

Step 2: We identify the data set that needs to be extracted to answer the question defined in Step 1. This step of the process includes identifying system specific (e.g. the EMIS computer system has additional codes over and above the national set) or locality specific codes used to record data. We also have to ensure we are collecting sufficient demographic data to define the denominator and characterise individuals. We define labels for all variables, as well as any associated dates and numerical values. These labels will be used in all the successive phases of the process. We use a controlled vocabulary to define these precisely.

Step 3: Identify the information governance issues that might relate to the study. These will include: privacy, confidentiality, and data protection. Ethical committee or other approval may be needed. Generally we extract de-identified data, and take steps to see it is fully anonymised prior to analysis; however all data can be de-identified in its originating practice.

Step 4: Pilot the data extraction and processing. Important lessons may be learnt from this that need to be fed back into the design process. Only once these changes have been incorporated, is the design stage complete.

Data entry issues: Clinicians and clerical staff enter data into their operational computer systems using the codes that their system presents. We set out to identify errors that might occur around data entry. This includes the process of identification of how patients with a clinical condition are represented within the coding system (e.g. what codes are used to identify patients with raised blood pressure). The social context of data entry also has to be considered. For example, a patient the clinician thinks is depressed may wish their problem title to be “headache.”

Extraction: We mainly use MIQUEST queries[9] to extract the data held within the computer system; though we have also experimented with XML (extensible mark-up language) and proprietary extraction tools. We looked to see what form of queries had been most effective in extracting data, and their influence on the down stream processes of migration, integration and analysis. Generally we find that extraction of data into a one line per patient data table is the most flexible arrangement for later analysis. However, we also develop queries that count the size of any dataset or subset. These are run at the same time as an internal validity check – to verify that the correct number of lines of patient data have been returned.

Migration: Migration of data from the general practice system to the data repository requires the transmission of that data. We usually do this using a physical storage device (floppy disk or flash memory card.) The data then has to be migrated into a format whereby it can be integrated into the data repository. We have to have a unique identifier in place throughout the process, so that the data has referential integrity and all the data can, where necessary, be linked to a single patient, practice and locality.

Integration: Different data tables, data about subsets of patients with one disease, serial data collections all have to be linked; so that the project outputs can be delivered. We do this for small projects in a customised Microsoft Access database or for larger projects in My-SQL. De-duplication is performed at this stage.

Cleaning: The cleaning of the data takes place at this final stage. Here the issue of out of range values, inconsistencies such as data entry in more than one type of unit (e.g. heights in centimetres and metres), and other problems with the data are addressed. We usually produce histograms of numerical variables so that we can see if they are normally distributed or if mixed units account for a binomial or other distribution; and the extent to which any...
outliers or strange values (e.g. one UK computer systems allocates a zero when a test request is made) are present.

**Processing:** Processing involves the conversion of extracted code into the plain English text assigned to that code by the coding system; e.g. the code H3z into “chronic obstructive pulmonary disease.” It involves grouping these into categories relevant to the intended analysis; something usually done at the design phase; e.g. sorting anti-psychotic medications into new drugs recommended by national guidelines and older ones. Sometimes more that one variable will contribute information to generate a new derived variable; e.g. patients treated with thyroxine as well as patients with a diagnosis of hypothyroidism might be included in a diagnosis of myxeodema.

**Analysis:** The output from the processing stage is usually a “flat file”. This will have one line per patient and variables in columns. The flat files always include the original un-cleaned data. Ensuring the accuracy of the denominator, and standardising prevalence’s so that comparisons can be made between populations are critical parts of this process.

The first stage is assess the quality of the data (completeness, accuracy, consistency, currency) and where appropriate to calculate the sensitivity and positive predictive value of diagnostic and prescribing data.

Automated reports are generated from this data for feedback to practices or localities; or flat-file data tables for research which are migrated into a standard statistical package.

### 3. Results

Examples are provided of where we have used our method to produce scientific research.

**Design:** We conducted a study of the stroke risk of patients with atrial fibrillation for a primary care organisation. The research question was whether stroke was sub-optimally managed and what the scope was to improve management, especially the use of anti-coagulants. These clear audit criteria enabled us to develop and pilot a dataset to meet their needs [16].

**Data entry:** Data entry issues are often overlooked. Very different picking lists are offered by the different computer systems. The way that the same patients are represented in different computer systems is illustrated in our study of osteoporosis; where we found 100 fold differences in recorded prevalence and that the major UK computing system could not accept results of the gold-standard test for osteoporosis, because the scan result is often a negative number [17]. Coding of bronchitis (H3z if chronic; H06 if acute) is also fraught with difficulty because not only is the difference between chronic and acute easy to confuse when coding but asthma (H33) is a child code of obstructive airways disease (H3) [18].

**Extraction:** There are many practical issues associated with data extraction. The largest query set we have extracted is a study to identify patients with undiagnosed kidney disease. As there are so many possible causes of kidney disease the dataset is very large and has to be broken down into sections for extraction [19].

**Migration:** Migration of data involves coping with all the quirks of computer systems prior to integration. Although there is a national specification, data may be in different formats. We try to hold data as triplets of “date-code-numerical value (if present)” (i.e. Date the code was recorded, the code (e.g. 44p = cholesterol), and the value (e.g. 5 mmol/l.) However, the different clinical systems may export data in varying orders, which then adds to the burden of customising the migration process for particular data sources. Similarly systems may have different variants of the same coding system – e.g. the different versions of the Read “codes”, and proprietary system codes [17].

**Integration:** We generate and use our own unique identifier (UPID) to make each patient’s line of data unique. Most computer systems generate a unique ID for that system – but there
is a chance it might be replicated elsewhere. We therefore compound this UPID with a unique reference for that practice, the research or audit they are participating in, and the local computer unique ID. This is converted into an ASCII format so that it can be used in non-case sensitive relational database joins. The final product, which we call “UPID ASCII” provides the database with full referential integrity.

Cleaning: We could cite many examples of customised data cleaning. One of the best was discovering that one computer system missed out the decimal point in its haemoglobin estimations. It appears that the users of this system have mentally inserted the decimal – as there have been no complaints about it. However, it meant that we found a bimodal distribution of haemoglobin [19].

Processing: This step involves the writing of syntax to recode data into names with meaning, usually categorical variables. We have evolved a set protocol for doing this. We define ranges for numerical variables by consensus with the lead clinicians and remove outliers. We adjust for incorrect units. In an audit looking at the prevalence of obesity we wished to maximise the proportion of patients with a body mass index (BMI) recorded. To do this we extensively cleaned a lot of height and weight data (e.g. if we find “5f10” in a height field we will reinterpret as 5 feet 10 inches and convert to its metric equivalent [20].)

Analysis: This involves the use of a statistical package. Increasingly we are stratifying patient risk outside the computerised medical record e.g. Calculating kidney function and stage of chronic kidney disease [19], stratifying the risk of stroke [16], and so on.

4. Discussion

This paper describes an eight step process for processing primary care data. The process planning is tightly aligned with its purpose: answering a research question or conducting an audit. We would stress the importance of using a controlled vocabulary for variable naming and the importance of developing an understanding of any data entry issues. We believe that are potentially serious weaknesses in analysing data when the researcher does not understand the picking list or other data entry choices that are presented to the person making the data entry; or the social context in which that data entry is made. It is widely recognised that medical diagnostic labels carry stigma and that coding in the consultation is not a neutral process [21]. The volume of publications generated [16-20] from routinely collected data suggests that this represents a reliable process.

Other groups working with routinely collected data should look to put their processes in the public domain so that the generalisable principles can become part of the core theory of primary care informatics [22]. The principals of error classification applied to data warehousing and total data quality management [14,15] need to be more widely adopted by the health informatics community; but adapted to meet our needs. Health informaticians need to carefully evaluate whether other techniques within computer science would enable more effective data warehouse design. There have been very few critical evaluation or comparison of current clinical databases [8] and further research is needed to compare methods of processing data.

5. Conclusions

This eight-step method provides a taxonomy, as well as a guide to the steps involved in processing routinely collected data. We would recommend that all those involved in processing primary care data define and publish their process of analysis and that all study teams working with primary care data include at least one person familiar with using the clinical systems.
6. Acknowledgments

MSD is the principal sponsors of PCDQ through a series of unconditional educational grants.

7. References

[1] London School of Hygiene and Tropical Medicine. Directory of Clinical Databases (DoCDat). URL: http://www.lshtm.ac.uk/docdat
[9] Clinical Information Consultancy. MIQUEST and Health Query Language. URL: http://www.clininf.co.uk/main/miquest.htm
[18] Falconer E, de Lusignan S. An eight step method for assessing diagnostic data quality in practice: Chronic obstructive pulmonary disease (COPD) as an exemplar. Accepted for publication *Informatics in Primary Care*.

**Corresponding author**

Jeremy van Vlymen, Primary Care Informatics, Division of Community Health Sciences, St. George’s, University of London, SW17 0RE., jvanvlym@sgul.ac.uk

---

**Section 13: Public Health Informatics, Clinical Trials**