Conference papers

A system of metadata to control the process of query, aggregating, cleaning and analysing large datasets of primary care data

Jeremy van Vlymen BEng MSc
Research Assistant

Simon de Lusignan BSc MSc MD MRCGP
Senior Lecturer
Primary Care Informatics, St George’s – University of London, London, UK

ABSTRACT

**Background**  Metadata is data that describes other data or resources. It has a defined number of named elements that convey meaning. Medical data are complex to process. For example, in the Primary Care Data Quality (PCDQ) renal programme, we need to collect over 300 variables because there are so many possible causes of renal disease. These variables are not just single columns of data – all are extracted as code plus date, while others are code–date–value. Metadata has the potential to improve the reliability of processing large datasets.

**Objective** To define unique and unambiguous metadata headings for clinical data and derived variables.

**Method** We defined the look-up tables we would use as a controlled vocabulary to name the core clinical concepts within the metadata. We added six other elements to describe data: (1) the study or audit name; (2) the query used to extract the data; (3) the data collection number; (4) the type of data, including specifying the units; (5) the repeat number (if the variable was extracted more than once); and (6) a processing suffix that defines how the data have been processed.

**Results** The metadata system has enabled the development of a query library and an analysis syntax library that make data processing and analysis more efficient. Its stability means greater effort can be put into more complex data processing, and some semi-automation of processes. However, the system has had implementation problems. It has been particularly hard to stop clinicians using multiple synonyms for the same variable.

**Conclusions** The PCDQ metadata system provides an auditable method of data processing. It is a method that should improve the reliability, validity and efficiency of processing routinely collected clinical data. This paper sets out to demystify our data processing method and makes the PCDQ metadata system available to clinicians and data processors who might wish to adopt it.

**Keywords:** data processing methods, metadata, primary care data quality

Introduction

Metadata is often referred to as ‘data about data’; it is structured data that describes a resource. Its purpose is to prevent misinterpretation of data and to facilitate finding resources. It is defined using agreed rules, and is capable of being stored in a central repository. A metadata schema will usually have the following characteristics:

- a defined and limited number of elements
- each element is named
- each element conveys meaning.
The concept of metadata predates the World Wide Web; it is not dependent on computer technology.\textsuperscript{3} The library index card is a good example. However, the speed at which computers can process large volumes of data, including large numbers of index cards, makes computer-readable metadata an important enabler of digital libraries.\textsuperscript{4} The use of metadata is widespread in tools that search and find online information.\textsuperscript{5} The World Wide Web Consortium now defines metadata as ‘machine understandable information for the web’.\textsuperscript{6} A large number of metadata schemata have been developed; perhaps the best known is the Dublin Core Metadata Initiative (DCMI).\textsuperscript{5} Whilst not strictly part of the schema, if the metadata can be encoded it will allow automated processing. Commonly, metadata schemes are enabled by HTML (Hypertext Mark-up Language), XML (eXtensible Mark-up Language), RDF (Resource Description Framework) and MARC (MAchine Readable Cataloging).\textsuperscript{7–10}

The causation of disease is complex, and the study of conditions using routinely collected computer data often requires the collection of multiple variables. For example, in the Primary Care Data Quality (PCDQ) renal programme, we need to collect over 300 variables because there are so many possible causes of renal disease.\textsuperscript{11} These variables are not just single columns of data – many are extracted as code plus date, while the remainder are a triple of code–date–value (for example, cholesterol will have an associated date of the result, and value in mmol/L). Data processing involves the addition of unique patient identifiers and the creation of derived variables (for instance, patients with cholesterol over 5 mmol/L).\textsuperscript{12} This can result in the final data table having over 500 variables. Although most clinical datasets contain common data elements requiring the same analysis (such as blood pressure data), very often each study would have its own idiosyncratic data labelling system. To control this process a system is required that links the concepts in the research question to the query generated to answer it, which link through the whole process to the variable in the final data table used for analysis. This paper reports the systems of metadata we have developed within the PCDQ programme to define unique and unambiguous metadata headings for clinical data and derived variables.

Method

We defined our project aim as creating metadata that would link the data requirements identified at the inception of a research project through to the final output (see Figure 1). Consequently the metadata had to be derived from the research question or audit criteria for the study. The research question or audit criteria are divided into diagnostic concepts, comorbidities and risk factors.\textsuperscript{13} Experienced researchers and clinicians would identify the coded data needed to answer the research question. We next involve experienced clinicians using all the different brands of GP computer systems in the study population – as we believe clinicians are more likely to select codes near the top of any picking list rather than necessarily the ontologically most appropriate.\textsuperscript{14,15} The codes required to collect the necessary data are listed; queries are created to extract the data from the various GP computer systems; and the final variable names and labels are derived from them. Consequently, the metadata enables traceability back to the source.

![Figure 1 Elements of the PCDQ metadata schema](image)
query and research question. Once the queries have been written they are stored in a query library. The query library is updated in total whenever an amendment is made to any query within it. The queries have names that inform:

- the programme the query relates to (for example, \( C_2 = \) cardiovascular programme no. 2)
- whether this is a ‘generic’ (labelled ‘G’) or a special query set related to a particular brand of computer system (for instance, we label queries for the EMIS system ‘E’ because it has a different drug code dictionary).

The metadata we have developed has a unique name for each data column; its format is compatible with that allowed in databases and analysis tools (for example, variable names must not start with a number in SPSS [Statistical Package for Social Sciences]). The core of the metadata is the rubric from the coding system that best describes the variable – we call this element of the metadata the ‘core clinical concept’ (CCC).

Our system of metadata requires that a coding or classification system is defined along with the browser or look-up table that will be used to identify the code. This is important because the exact syntax that appears in the look-up table is what will be used in the metadata; the description used to describe it will be used in exactly that format as the variable label. This exact copy must use the same choice of upper and lower-case letters as well as spacing and punctuation. Using an exact copy also means that the coding and classification system can act as a controlled vocabulary. At any stage in the process, from translation of concept within the research question to final output, the rubric in the variable name can be translated to the core clinical concept and vice versa. This avoids confusion between multiple medical near-synonyms (such as CHD [coronary heart disease], IHD [ischaemic heart disease], etc.) and ‘creep’ in meaning as variable labels are progressively reinvented. We have seen migration of meaning in variable labels from more- to less-, as well as less- to more-specific meaning. For example, the variable label ‘paroxysmal atrial fibrillation’ evolved to ‘atrial fibrillation’ then to ‘atrial flutter and fibrillation’ – theoretically larger populations at each step.

We then created prefixes and suffixes to qualify this term and make it unique as well as to convey information about the type of information contained within the variable. Each element of the metadata is separated by an ‘underscore’ character. The prefixes identify the query set that generated the query; if this is time-series data the sequence of the data collections and the type of data (for example, diagnosis, history, examination finding, lab result, or date or numeric associated with a variable). The suffixes inform whether this is the latest, if the same variable has been extracted more than once, and if this is the original data or if it has been cleaned or processed.

The first element is the query set used (QSU) to extract these data. These relate to the originating query set (for example, C4 for our fourth set of cardiac queries). As query sets inevitably evolve, for example to overcome technical problems or to incorporate new drugs or codes, the combination of information about which brand of computer system and the date of the data collection allow the query used to extract these data to be readily identified within the query library. The second element of the metadata is the query name (QNA).

The third element of the prefix indicates the data collection number (DCN). This field is blank if it is a single data collection. If successive data collections are performed (for instance, after an intervention) then these subsequent data collections are labelled ‘aa’, ‘bb’, ‘cc’, and so on. If there were ever more than 26 data collections then the series would restart: ‘aaa’, ‘abb’, ‘acc’, and so on.

The fourth element of the metadata prefix is the type of data (DTY). We need to indicate data type because we are usually extracting data as ‘couples’ or ‘triples’ of related data. We nearly always collect the associated date with any codes we extract to make a pair (that is, CODE + DATE). Where there is an associated numerical value we also collect this (that is, CODE + DATE + VALUE). The type of data element is the only element of a variable name which will differentiate whether the code, date or value is extracted. For example: we use ‘MY’ to indicate a date; ‘MI’ to indicate that the units are millimol per litre, and ‘44P’ for serum cholesterol. MY_44P will be the elements that define this as the date of the cholesterol test. This system also helps enable efficient data processing – as heights (in our reference code look-up 229 is the code for O/E height) exported from clinical systems in centimetres would have the element labels cm_229; whereas heights exported in metres would be me_229.

Numerical data type codes are shown in Table 1, and those we use to differentiate clinical data types in Table 2.

There are two metadata elements in the suffix to the CCC element of the metadata; these are the repeat number (RNO) and the processing indicator (PRO). The former, RNO, indicates the number of repeats of a given variable that appear in the dataset. For example, in one study we extracted the last three creatinine readings in order to look at rate of change in renal function. ¹¹ If a variable is only collected once this metadata element can be void. Our convention is that the latest recording is labelled 1, the penultimate recording 2, and so on. We use an ‘f’ when we collect the first occurrence of an event which we sometimes need as part of an analysis. If a sequence from the earliest recording is collected we use ‘f1’, ‘f2’, etc. The
The metadata also contains names for different identifiers (see Table 3) and rules for the creation of derived variables, which may come from a range of sources. These will vary dependent upon local circumstances.

**Table 1** Data-type elements used in the PCDQ programme for numerical values

<table>
<thead>
<tr>
<th>Bigram</th>
<th>Meaning</th>
<th>In full</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td>mmol/l</td>
<td>Millimol per litre</td>
<td>NB. We can’t use ml as we may need this for volume at some stage</td>
</tr>
<tr>
<td>ML</td>
<td>ml</td>
<td>Millilitre</td>
<td>See above</td>
</tr>
<tr>
<td>PI</td>
<td>pmol/l</td>
<td>Picomol per litre</td>
<td></td>
</tr>
<tr>
<td>GD</td>
<td>g/dl</td>
<td>Grams per decilitre</td>
<td>Used in haemoglobin, and US measures of creatinine and cholesterol</td>
</tr>
<tr>
<td>UI</td>
<td>micromol/l</td>
<td>Micromol per litre</td>
<td>Measure of serum creatinine concentration</td>
</tr>
<tr>
<td>KG</td>
<td>kg</td>
<td>Kilogram</td>
<td>Used to measure weight</td>
</tr>
<tr>
<td>BM</td>
<td>kg/m²</td>
<td>Kilograms per metre squared</td>
<td>The units of body mass index</td>
</tr>
<tr>
<td>ME</td>
<td>m</td>
<td>Metre</td>
<td>Sometimes used as a measure of height</td>
</tr>
<tr>
<td>CM</td>
<td>cm</td>
<td>Centimetre</td>
<td>Sometimes used as a measure of height</td>
</tr>
<tr>
<td>IU</td>
<td>IU</td>
<td>International Unit</td>
<td>Used to measure some drugs, e.g. heparin</td>
</tr>
<tr>
<td>MG</td>
<td>mg</td>
<td>Milligram</td>
<td>Used to measure dose of some drugs</td>
</tr>
<tr>
<td>TB</td>
<td>Tablets</td>
<td>Number of tablets, capsules or other unit dose of a drug</td>
<td>The number associated with every prescription – unfortunately this does not follow a standard format (e.g. can be ‘1 op’ - 1 original pack; or 28 tabs; or other variant)</td>
</tr>
<tr>
<td>HG</td>
<td>mmHg</td>
<td>Millimetres of mercury</td>
<td>How blood pressure is measured</td>
</tr>
<tr>
<td>MY</td>
<td>‘Month–Year’</td>
<td>Used to describe a date</td>
<td>NB. Please always use: dd mm yyyy format for all dates in SPSS</td>
</tr>
<tr>
<td>UN</td>
<td>units</td>
<td>Units</td>
<td>Non-specific numeric to be used for numerical lists not defined above</td>
</tr>
</tbody>
</table>
### Table 2  Metadata elements used to describe different types of clinical data

<table>
<thead>
<tr>
<th>Bigram</th>
<th>Meaning</th>
<th>Read code characteristic</th>
<th>Example code</th>
<th>Notes/meaning of example code</th>
</tr>
</thead>
<tbody>
<tr>
<td>DI</td>
<td>Diagnosis</td>
<td>Read code starts with a capital letter</td>
<td>G3</td>
<td>Use DI for diagnosis when the column header primarily describes patient with a diagnosis</td>
</tr>
<tr>
<td>RX</td>
<td>Drugs prescription</td>
<td>Read code starts with a lower case letter</td>
<td>bxd1</td>
<td>Use RX where the column contains a drug, e.g. simvastatin</td>
</tr>
<tr>
<td>OC</td>
<td>Occupation</td>
<td>Read code starts with the zero numeric</td>
<td>031</td>
<td>University academic staff</td>
</tr>
<tr>
<td>HO</td>
<td>History symptoms</td>
<td>Read code starts with a number 1</td>
<td>14A5</td>
<td>History of angina</td>
</tr>
<tr>
<td>OE</td>
<td>Examination signs</td>
<td>Read code starts with a number 2</td>
<td>246</td>
<td>OE BP reading</td>
</tr>
<tr>
<td>DP</td>
<td>Diagnostic procedures</td>
<td>Read code starts with a number 3</td>
<td>324</td>
<td>ECG left ventricle hypertrophy</td>
</tr>
<tr>
<td>LP</td>
<td>Laboratory procedures</td>
<td>Read code starts with a number 4</td>
<td>44P</td>
<td>Serum cholesterol</td>
</tr>
<tr>
<td>XR</td>
<td>Radiology physics in medicine</td>
<td>Read code starts with a number 5</td>
<td>5543</td>
<td>Coronary arteriography abnormal</td>
</tr>
<tr>
<td>PR</td>
<td>Preventive procedures</td>
<td>Read code starts with a number 6</td>
<td>66AD</td>
<td>Fundoscopy – diabetic check</td>
</tr>
<tr>
<td>OP</td>
<td>Operations</td>
<td>Read code starts with a number 7</td>
<td>79232</td>
<td>Prosthetic replacement coronary artery</td>
</tr>
<tr>
<td>TH</td>
<td>Other therapeutic procedures</td>
<td>Read code starts with a number 8</td>
<td>8CAL</td>
<td>Smoking cessation advice</td>
</tr>
<tr>
<td>AD</td>
<td>Administration</td>
<td>Read code starts with a number 9</td>
<td>9S2</td>
<td>Black Caribbean</td>
</tr>
</tbody>
</table>

### Table 3  Names for identifiers and administrative data

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>MIQUEST-generated unique identifier which can be generated in a non-case-sensitive manner</td>
</tr>
<tr>
<td>UPID</td>
<td>Patient unique identifier derived from REF (non-case-sensitive ASCII format)</td>
</tr>
<tr>
<td>PCOID</td>
<td>Primary care organisation (PCO) identifier</td>
</tr>
<tr>
<td>PRACNUM</td>
<td>Practice number – arbitrary unique ID generated locally. Decoding this requires access to a secure table of practice ID</td>
</tr>
<tr>
<td>PRACID</td>
<td>PCOID–PRACNUM combination. This makes each practice truly unique</td>
</tr>
<tr>
<td>UPIDASCII</td>
<td>Patient ASCII-encoded unique identifier</td>
</tr>
<tr>
<td>COLDATE</td>
<td>Collection date</td>
</tr>
<tr>
<td>GPSYS</td>
<td>GP clinical system and version</td>
</tr>
<tr>
<td>CLASS</td>
<td>Coding system and set</td>
</tr>
</tbody>
</table>
Results

Definition of look-up tables that generate the CCC metadata element is critical if the metadata is to be effective. We currently use the Read version 2, 5-byte code set, as it is the most commonly used coding system in the data we are processing; however, we could have used any other coding or classification system. We also defined the coding system browser associated with each release of the metadata standard. We initially used the Clue Read 2, 5-byte browser; more recently we have used the NHS Information Authority Clinical Terminology Browser (Triset).16,17 Occasionally we have to define an additional dictionary; for example, the EMIS clinical computer system uses a different drug dictionary.18 This is not a permanent decision but one that needs to be defined and recorded. We do this in our query library so that the appropriate look-up tables are associated with every release of queries. Historically we used Read version 2, 4-byte, but we recognise that we may need to migrate to Read Clinical Terms version 3 or SNOMED-CT (Systematised Nomenclature of Medicine – Clinical Terms).19

A mechanism of linking the research question to the queries written to extract data and then to the variable labels used in analysis is vital to maintain quality control. The research terms used to express the research question are captured in a structured list (see Figure 2).

The terms are then coded into metadata (see Figure 3). The metadata elements are defined from the research programme, the name of the query and the data type. They are unique and can be linked back to the individual query that extracted them. This means that the concept can be tracked from the research question to the query, through extraction and processing, to the final flat file used to produce research output.

The linkage between the query library, the relevant syntax used to process these data and the Read code look-up engine are shown in Figure 4. Consistent use of variable labels has resulted in greater efficiency in processing, as syntax to sort data can be reused in different projects.

The repeat index within the metadata allows for automation of otherwise onerous processing such as ‘in case’ sorting of ‘date–code–value’ triples in SPSS.20 Whilst it is easy to sort single types of data, it is very hard to sort doubles or triples. In the example shown in Figure 5, serum creatinine values are sorted by latest date order to ensure the most up-to-date valid reading is used in the calculation of estimated glomerular filtration rate. What is described is a bubble-sort algorithm, making use of the VECTOR and LOOP commands in SPSS.21

The sorting of large numbers of pairs (code–date) or triples (code–date–value) in a large dataset (for example, 400-plus variables consisting of a pair or triple) is performed programmatically, again taking advantage of our metadata structure. In Figure 6 we demonstrate the use of Practical Extraction and Reporting Language (PERL) to sort variable names.22 Without a stable metadata structure it would require new codes to be written each time a sort was performed.

Finally, we present the example of an algorithm to sort diabetic patients (see Figure 7). This takes account of a number of factors: the diagnostic code used, the age and gender of the patient, their body mass index.
Figure 3 Coded research concepts list

Figure 4 Medical concept in extraction syntax
Discussion

Principal findings

A system of metadata has enabled us to ensure that we collect data to answer the research question; our process is auditable and is more reliable as a result. It also enables us to link between the variable name and data extraction query—which is important should results show unexpected findings. The metadata standard has also improved processing efficiency and been an enabler of partial automation. Finally, the metadata also enables the original data to be readily identified within the final flat file.

Implications

Developing metadata standards has the potential to make the cleaning and processing of data more transparent. It will also allow better auditing and quality control of research or audit output based on routinely collected clinical data. We would like to see standard metadata element names for data cleaning functions. We believe that a shared system of metadata could make a significant contribution to research governance.

Limitations of the method

The metadata cannot as yet be machine processed. To gain maximum leverage from the schemata developed we need to develop methods of automated processing. We have outlined the first steps we feel should be taken to achieve this in our call for further research.

The system has also had implementation problems. It has been particularly hard for clinicians used to using multiple synonyms for the same variable (such as heart disease, IHD, CHD) to adjust to its use. Also, insisting on the use of the coding system look-up engine can result in some ‘unnatural’ phrases. For example, the precise label for Read code 2469 is ‘O/E – systolic BP reading’, which maybe provides too much information. However, overall we have found that precise use of the look-up engine as a controlled vocabulary has reduced processing errors.

Comparison with the literature

There is a dearth of literature describing the methods of processing routinely collected clinical data. At best the data sources are described, or look to validate findings through comparisons with other data collection schemes. XML has been used within medical records, and also MML (Medical Mark-up Language). These languages offer scope for making our metadata capable of computer processing.

Call for further research

The elements of this metadata could be machine processed if an appropriate encoding of the tags were...
A system of metadata to control the processing of large datasets of primary care data

Figure 6  Sorting of metadata in PERL for SPSS syntax vector processing
carried out. We believe it should be possible to auto-
generate the metadata labels. The mapping from con-
cepts in a research question to the appropriate codes
could be achieved by assigning MeSH (Medical Sub-
ject Headings), then using the UMLS (Unified Medi-
cal Language System) Metathesaurus to link to the
coding system.\textsuperscript{27,28} If this were shown to be reliable it
would enable the computer generation of metadata.

\textbf{Conclusions}

The PCDQ metadata system provides a mechanism
for improving the reliability, validity and efficiency of
processing routinely collected clinical data. We accept
that an object orientated approach would represent
the gold standard in data definition and processing.
However, where the evidence base, treatment thresh-
olds, coding system, clinical system and needs of the
research project are all dynamic, this approach pro-
vides a pragmatic way of working. This paper sets out
to demystify our data processing method and makes
the PCDQ metadata system available to clinicians and
data processors who might wish to adopt it.

\textbf{ACKNOWLEDGEMENTS}

We would like to acknowledge the contributions of
the primary care informatics team at St George’s;
colleagues in British Computer Society (BCS) working
groups and the Health Informatics Forum; and
members of the European Federation for Medical
Informatics (EFMI) Primary Care Working Group.
Funding from PCDQ project sponsors has facilitated
the development of this system.

\textbf{REFERENCES}

1 Howe D. Free online dictionary of computing
(FOLDOC), 1996. wombat.doc.ic.ac.uk/
2 Taylor C. An introduction to Metadata. University
ctmeta4.html
3 Miller P. Metadata for the Masses. www.ariadne.ac.uk/
issues/metadata-masses/howe_96
4 Lifesign project. Metadata literature review. www.
lifesign.ac.uk/documents/Metadata\%20literature\%20
review.doc
5 Dublin Core Metadata Initiative (DCMI). Using Dublin
Core. dublincore.org/documents/usageguide/
6 The World Wide Web Consortium (W3C). Metadata
and Resource Description. www.w3.org/Metadata/
7 The World Wide Web Consortium (W3C). Hypertext
Mark-up Language (HTML). www.w3.org/MarkUp/
8 The World Wide Web Consortium (W3C). Extensible
Mark-up Language (XHTML). www.w3.org/XML/
9 The World Wide Web Consortium (W3C). Resource
Description Framework (RDF). www.w3.org/RDF/
10 Library of Congress. Network Development and MARC
(MAchine Readable Cataloging) Standards. www.loc.
gov/marc/
11 de Lusignan S, Chan T, Stevens P, O’Donoghue D,
Hague N, Dzregah B \textit{et al}. Identifying patients with
chronic kidney disease from general practice computer


Clinical Information Consultancy. Clue Look-up Engine. www.clininfo.co.uk/main/clue.htm

NHS Information Authority Clinical Terminology Browser (Triset): www.nhsia.nhs.uk/terms/pages/ct_browser.asp?om=m1


SPSS (Statistical Package for Social Sciences): www.spss.com/


The PERL Directory. www.perl.org


This paper was presented at the PHCSG Annual Conference 2005, Downing College, Cambridge, September 2005.

CONFLICTS OF INTEREST

None.

ADDRESS FOR CORRESPONDENCE

Simon de Lusignan
Senior Lecturer, Primary Care Informatics
Hunter Wing
St George’s – University of London
London SW17 0RE
UK
Tel: +44 (0)208 725 5661
Fax: +44 (0)208 767 7697
Email: slusigna@sgul.ac.uk

Accepted September 2005