Informatics in Primary Care 2011;19:127–34

Leading article

Defining datasets and creating data dictionaries for quality improvement and research in chronic disease using routinely collected data: an ontology-driven approach

Simon de Lusignan MSc MD(Res) FHEA FBCS CITP FRCGP
Professor of Primary Care and Clinical Informatics, Department of Health Care Management and Policy, University of Surrey, Guildford, UK

Siaw-Teng Liaw PhD FRACGP FACHI
Professor of General Practice, University of New South Wales, Director, General Practice Unit, SW Sydney Local Health District, Sydney, Australia

Georgios Michalakidis MSc
Doctoral Student, Computing Department

Simon Jones MSc PhD
Research Professor, Department of Health Care Management and Policy

University of Surrey, Guildford, UK

ABSTRACT

Background The burden of chronic disease is increasing, and research and quality improvement will be less effective if case finding strategies are suboptimal.

Objective To describe an ontology-driven approach to case finding in chronic disease and how this approach can be used to create a data dictionary and make the codes used in case finding transparent.

Method A five-step process: (1) identifying a reference coding system or terminology; (2) using an ontology-driven approach to identify cases; (3) developing metadata that can be used to identify the extracted data; (4) mapping the extracted data to the reference terminology; and (5) creating the data dictionary.

Results Hypertension is presented as an exemplar. A patient with hypertension can be represented by a range of codes including diagnostic, history and administrative. Metadata can link the coding system and data extraction queries to the correct data mapping and translation tool, which then maps it to the equivalent code in the reference terminology. The code extracted, the term, its domain and subdomain, and the name of the data extraction query can then be automatically grouped and published online as a readily searchable data dictionary. An exemplar online is: www.clininf.eu/qickd-data-dictionary.html

Conclusion Adopting an ontology-driven approach to case finding could improve the quality of disease registers and of research based on routine data. It would offer considerable advantages over using limited datasets to define cases. This approach should be considered by those involved in research and quality improvement projects which utilise routine data.

Keywords: classification, medical informatics, medical records systems, computerised

Introduction

Growing burden of chronic disease

Internationally, there is a growing burden of chronic disease, and a need to reorientate health services towards the provision of chronic care.[1] Computerised medical records systems may have a role in improving management by enabling the ready identification of cases and in monitoring quality.[2,3] These computerised disease registers are likely to be important in areas where there are quantitative measures that define whether you have a particular disease and for measuring the quality of care. Diabetes[4] and the secondary prevention of cardiovascular disease including the management of hypertension[5,6] provide examples of where computerised medical records enable quality improvement even though there remains scope for refinement.[7,8]

Practical approaches to case finding in chronic disease: ontologies and data dictionaries

Two practical approaches are given to ensure that we identify cases and systematically list the codes required to conduct research or quality improvement. Ontologies provide insight into what might be extracted from a clinical system to provide the data we require and data dictionaries provide an accessible list of the extracted variables.

Ontologies to define cases with a chronic disease

Ontologies provide a method for describing concepts and relationships within a domain. The principal use of ontologies in informatics is to enable human and machine communication, by defining the terms used to describe an area of knowledge. Ontologies usually have the following components:

- classes – general types of entities in the domain
- relationships that can exist among and between the things within the domain
- the properties (or attributes) those things may have.[9]

Another recognised use of ontologies is for the retrieval of data.[10] However, this approach has not been widely used in quality improvement or research into the management of chronic disease. There is the potential to use ontologies to define datasets that might be used to identify people with a chronic condition for quality improvement or research.

One of the best known definitions of ontologies in informatics emphasises both their machine-processable and human interpretability:

> Ontologies are: collections of formal, machine-processable and human-interpretable representations of the entities, and the relations among those entities, within a defined application domain—are helping researchers manage the information explosion by providing explicit descriptions of biomedical entities and an approach to annotating, analyzing the results of clinical and scientific research.
>
> Ontologies are useful because they provide regimentations of terminology that can support the reusability and integration of data and thereby support the development of useful systems for purposes such as decision support, data annotation, information retrieval, and natural-language processing.[11]

Data dictionary a centralised repository of the dataset that defines a case

A data dictionary is a centralised repository of information about data such as the meaning, relationships to other data, origin, usage and format.[12] A data dictionary could capture the classes of information, some relationships and properties of the data. Data dictionaries are a potential mechanism for ensuring the transfer of meaning into clinical information systems and ultimately improve care efficiency.[13] Data dictionaries can also play an important role in modelling and in the specification and requirements analysis with the use of metadata.[14]

Objective

This leading article proposes that ontologically rich approaches should be used to define datasets to identify cases in quality improvement and research projects. If this were done, it would substantially improve the identification of cases within routinely collected data. We propose how a dataset might be constructed and how variable lists for all studies using substantial datasets might be displayed in an accompanying data dictionary.

Method

Overview

We propose a five-step process: (1) identifying a reference coding system or terminology; (2) using an ontology-driven approach to identify relevant concepts and relationships that might define a case; (3) developing metadata that can be used to identify the source and nature of extracted data; (4) mapping

the extracted data to the reference terminology; and (5) creating the data dictionary.

Identifying a reference coding system or terminology

We recommend selecting a comprehensive coding system for use as the reference coding system for a project. The coding system selected should be the most commonly used for that particular study and be capable of having the core relevant concepts mapped to it. Increasingly, data for a study are recorded using more than one coding system. For example, in the UK, the coding systems used are:[15]

- Read version 2 (hierarchical)
- Read Clinical Terms version 3 (CTv3)
- Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)
- International Classification of Disease version 10 (ICD-10)
- Office of Population Census and Surveys version 4 (OPCS4).

By way of contrast in Australia they use:

- International Classification of Primary Care (ICPC)
- Doctors Command Language (DOCLE)[16]
- ICD-10-AM (Australian Modification).

Historically, much research was based on data collected from single brands of computer system and such single EPR supplier research networks have been extremely successful.[17] However, more and more research involved linking between databases, so that the effect of an intervention in one part of the health system can be seen in another. We, for example, have demonstrated how improving access to psychological therapies (IAPT) has a positive impact on accident and emergency uptake.[18]

Defining the relevant concepts and relationships within the reference terminology

A person having a chronic disease may be identified from a range of codes. A disease code will usually signify that a person has a chronic condition; e.g. the disease code for 'Essential Hypertension' in Read codes version 2 is 'G20z'. However, codes from other parts of the classification may also signify that someone has hypertension and codes are sometimes inserted in error. We suggest using a tabular approach in which each chapter of a coding system is explored to see if codes that might represent a person with hypertension indicate that the person has the condition.

Metadata to control extraction and uploading

Data extraction is not consistent between different brands of electronic patient record (EPR) systems or coding systems. It is necessary to create a metadata system that links and labels the coding system used in the site from which data is extracted; the brand of EPR and version, for example, can affect the drug dictionary used and the data extraction query. Metadata are data that describe other data and therefore can be used to control and manage processes.[19]

Mapping data from different sources

The uploaded data from different sources, labelled by system metadata, then needs to be mapped using validated processes wherever possible. If not available, this needs to be done involving clinicians in the field who understand its ontological significance.

Creating the data dictionary

The data dictionary should be readily searchable and display the code and term, the domain and a link to the relevant data extraction query. The metadata drives the creation of the data dictionary for all the terms returned by the data extraction queries. It links extracted data from different coding systems to a common list of subdomains and domains, as well as to the data extraction query.

For example:

- G20z is an example code
- the term is 'Essential Hypertension NOS'
- it belongs to the subdomain called 'G2 Hypertensive Disease'
- the related domain is 'G: Circulatory System Disease'
- it was extracted by a query called 'Cardiovascular co-morbidities...'.

Results

Identifying a reference coding system or terminology

We generally use the most commonly used in a particular study. We currently use Read version 2, 5-Byte for UK primary care studies; and ICD-10 for hospital studies; using OPCS4 where operations or procedures are the primary focus. However, this choice can vary according to the usual practice in the areas under investigation. Where a single brand of computer system is used we may have to include local

codes. These remain much used in the EMIS system, and for other brands we may use CTv3.

Ontologically rich approach to identifying cases

We look for codes that might enable us to identify cases by systematically searching across coding hierarchies, or identifying alternative codes in other non-hierarchical systems which may indicate that the patient has the condition. We look for cases by searching for 'History of' codes (e.g. 14A2, History of Hypertension), 'Diagnosis' codes (e.g. G20z, Essential Hypertension), 'Procedure' codes (e.g. 6628, Poor hypertensive control), 'Administration' codes (e.g. 901% Hypertension monitoring administration) and 'Therapy' codes which imply the condition (e.g. b2% Thiazide diuretics). Our method for identifying cases of hypertension is shown in Table 1. Further refinements include the use of '%' as the end of a code when all child codes are included and a '.' (full stop) when just the code listed is required. We also list codes within a hierarchy we wish to exclude.

We also indicate the likelihood of a code to truly map to a condition. We develop rules on a study by study basis. The most complex we have developed were to enable the machine processing of a diagnosis of diabetes into definite, probable, possible or not having the condition.[20]

Metadata to control the process

We initially developed metadata to make our data processing more efficient and consistent.[21] However, this was developed when we were principally working with just primary care data and complex methods are needed to cope with linking heterogeneous datasets. We have subsequently developed and added a solution-orientated taxonomy to report data extraction errors, so we can understand any gaps in our data.[22] We use Java and Another Tool for Language Recognition (ANTLR) for the parsing of data[23] to ensure their consistency with the reference terminology.

Mapping data from different sources

Our system collects all individual de-duplicated (via parallel processing) clinical codes from different extraction samples and stores them in a memory-efficient data store for further processing.[24] The results carry heavy metadata (e.g. coding system used, original extracted set for traceability).[25] We carry out our mapping and translate codes to definitions within our reference coding systems. Wherever possible we use validated translation schemas provided by Technology Reference Data Update Distribution (TRUD), NHS – for mapping Read Clinical Terms version 3 (CTv3) to Read 2. We also use translations provided by EPR vendors, for example, Egton Medical Information Systems (EMIS) mapping to convert EMIS drug codes to standard Read 2 codes. Only exceptionally will we devise manual schemas for mapping. Where we do, we classify our mapping into 'Direct', 'Partial' and 'No clear' mapping.[26] Where mapped codes appear in the data dictionary they appear with the code to which they are mapped, e.g. the CTv3 code 'XE0Uc' appears with the comment 'Essential hypertension (Read 2 equivalent: G20)'. The mapped codes do not display a domain or subdomain, but do display the name of the data extraction query as this will be different to the query used to extract data ('Collection request') from practices using Read 2 codes.

Creating the data dictionary

We have created a method whereby code hierarchy is dynamically generated, with the identification and translation of the code domain (e.g. 1: History/Symptoms) and subdomain (e.g. 12: Family History) for each individual code as well as the extraction metadata. An online system for web and mobile representation of the dictionary data for all the Clinical Informatics research group's current projects are now placed online and are freely available (e.g. Osteoporosis data dictionary is available at: www.clininf.eu/osteoporosis-data-dictionary.html) This allows for dynamic searches in sets with thousands of codes and a view of the complete dataset each study holds (Figure 1). It also means that investigators and collaborators can readily identify the data available for a particular project.

Discussion

This paper describes a way of using ontologies to ensure the high chance of identifying people who have chronic diseases from routinely collected clinical data; and data dictionaries can provide browsable lists of variables extracted. Data in primary care may not be complete or accurate, or current,[27] and there may be measurable gaps in data quality.[28,29] Therefore, we need to extract and process data in a way that takes account of its limitations,[30] and this should include taking account of the presence or absence of ontological relationships.

Many researchers and others involved in extracting routinely collected data who understand issues about data quality may already be addressing the principles

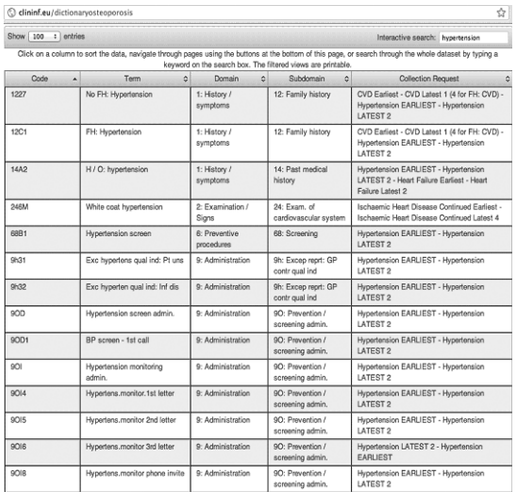
Table 1 An ontologically rich approach to identifying cases of essential hypertension

Code category	Included				Excluded			
	Domain	Subdomain	Term	Code	Domain	Subdomain	Term	Code
Disease	Circulatory system disease	Hypertensive disease	Hypertensive disease	G2.				
			Essential hypertension	G20%				
			Secondary hypertension	G24%	Circulatory system disease	Hypertensive disease	Hypertension secondary to drug	G24z1
			Other specified hypertensive disease	G2y%				
			Hypertensive disease NOS	G2z%				
	Nervous system and sense organ diseases	Disorders of eye and adnexa	Hypertensive retinopathy	F4213				
	Pregnancy/childbirth/ puerperium	Pregnancy complications	Pre-exist hypertens compl pregnancy	L128%				
History	Past medical history	H/O Cardiovascular disease	H/O Hypertension	14A2				
Examination	Examination/signs	Exam. Cardiovascular system	O/E BP reading	2469				
			O/E Systolic BP reading	246%				
			O/E Diastolic BP reading	246A				
			White coat hypertension	246M				

Table 1 Continued

Code category	Included				Excluded			
	Domain	Subdomain	Term	Code	Domain	Subdomain	Term	Code
Investigations	Diagnostic procedures	Electrocardiography	EGC left ventricular hypertrophy	342%	Diagnostic procedures	Electrocardiography	ECG no LVH	324..
Procedures	Preventive procedures	Cardiac disease monitoring	Good hypertension control	6627				
			Poor hypertension control	6628				
	Other therapeutic procedure	Referral for further care	Referral to hypertension clinic	8HTS				
Administrative	Administration	Prevention/ screening admin	Hypertension monitoring admin	901%				
		Patient encounter admin data	Seen in hypertension clinic	9NO3%				
Therapy	Cardiovascular drugs		Thiazide diruetics	b2%				
			Beta-adrenoreceptor blockers	bd%				
			Angiotensin converting enzyme inhibitor	bi%				
			Other antihypertensives	bk%				

Key: %= collect all child codes; "." Just the specific term; H/O = history of; NOS = not otherwise specified


cininf.eu/dictionaryosteoporosis

Show 100 entries Interactive search: hypertension

Click on a column to sort the data, navigate through pages using the buttons at the bottom of this page, or search through the whole dataset by typing a keyword on the search box. The filtered views are printable.

Code	Term	Domain	Subdomain	Collection Request
1227	No FH: Hypertension	1: History / symptoms	12: Family history	CVD Earliest - CVD Latest 1 (4 for FH: CVD) - Hypertension EARLIEST - Hypertension LATEST 2
12C1	FH: Hypertension	1: History / symptoms	12: Family history	CVD Earliest - CVD Latest 1 (4 for FH: CVD) - Hypertension EARLIEST - Hypertension LATEST 2
14A2	H / O: hypertension	1: History / symptoms	14: Past medical history	Hypertension EARLIEST - Hypertension LATEST 2 - Heart Failure Earliest - Heart Failure Latest 2
246M	White coat hypertension	2: Examination / Signs	24: Exam. of cardiovascular system	Ischaemic Heart Disease Continued Earliest - Ischaemic Heart Disease Continued Latest 4
68B1	Hypertension screen	6: Preventive procedures	68: Screening	Hypertension EARLIEST - Hypertension LATEST 2
9h31	Exc hypertens qual ind: Pt uns	9: Administration	9h: Excep reprt: GP contr qual ind	Hypertension EARLIEST - Hypertension LATEST 2
9h32	Exc hyperten qual ind: Inf dis	9: Administration	9h: Excep reprt: GP contr qual ind	Hypertension EARLIEST - Hypertension LATEST 2
9OD	Hypertension screen admin.	9: Administration	9O: Prevention / screening admin.	Hypertension EARLIEST - Hypertension LATEST 2
9OD1	BP screen - 1st call	9: Administration	9O: Prevention / screening admin.	Hypertension EARLIEST - Hypertension LATEST 2
9OI	Hypertension monitoring admin.	9: Administration	9O: Prevention / screening admin.	Hypertension EARLIEST - Hypertension LATEST 2
9OI4	Hypertens.monitor.1st letter	9: Administration	9O: Prevention / screening admin.	Hypertension EARLIEST - Hypertension LATEST 2
9OI5	Hypertens.monitor 2nd letter	9: Administration	9O: Prevention / screening admin.	Hypertension EARLIEST - Hypertension LATEST 2
9OI6	Hypertens.monitor 3rd letter	9: Administration	9O: Prevention / screening admin.	Hypertension LATEST 2 - Hypertension EARLIEST
9OI8	Hypertens.monitor phone invite	9: Administration	9O: Prevention / screening admin.	Hypertension EARLIEST - Hypertension LATEST 2

Figure 1 Example online data dictionary, searched using the term 'Hyperten'

set out in this paper. However, others come to work on routinely collected data without contextual insight as to the range of codes that might be used to represent a case.[31] Data dictionaries make explicit the link between code and term, to subdomain and domain, and data extraction query. They can be generated dynamically from systems that have developed metadata to link these items and to flag mapping between coding systems.

However good the ontologically rich process of defining cases or of setting out of the terms used in our data dictionary there will be limitations. Concepts often evolve and relationships change. Not all relationships will be perfect. For example, it appears impossible to avoid extracting family history of hypertension codes when looking for the codes for hypertension. This can be adjusted for in the final analysis of data, but illustrates that it is not always possible to make perfect ontological links. The 'Chocolate teapot not otherwise specified' discussion paper illustrates this point well and provides good counsel against ontological obsessionalism.[32,33] Not all concepts have direct mapping to a single diagnosis, and sometimes an operation, procedure or other process of care code may be the only indication that a person might have the condition. Others have recognised that there may be mandatory, multiple or numeric criteria for formalising description logic ontologies.[34] A similar approach to identify patients with diabetes, using a combination of diagnostic terms as well as medications and laboratory tests has been used in Australian primary care.[35]

A final advantage of the ontology-driven approach to defining cases is that it will be inclusive rather than limited. Hayes, in his principles, decries the 'dataset mentality'.[36] This is effectively an arbitrary list of codes which signifies that an individual has a condition. The downside of the limited dataset approach is that it will inevitably miss cases represented elsewhere within the clinical record. Whether for research or as part of a disease register for quality improvement, adopting an ontology-driven approach is likely to create a list of variables that are inclusive of patients with a particular condition; albeit that some of the mappings will be partial.

Conclusion

The process from case finding to data extraction to creating a data dictionary should be seen as a continuum. Data dictionaries can link extracted codes and terms to clinical domains and data extraction queries. An automated method which has proved more efficient than manual approaches (people extracting routine data to identify cases with chronic disease) may be more likely to identify cases if they take an ontologically rich approach.

REFERENCES

1 Boult C, Karm L and Groves C. Improving chronic care: the 'guided care' model. *The Permanenete Journal* 2008; 12(1):50–4.
2 Muttitt SC and Alvarez RC. Chronic disease management: it's time for transformational change! *Healthcare Papers* 2007;7(4):43–7; discussion 68–70.
3 Samoutis GA, Soteriades ES, Stoffers HE, Zachariadou T, Philalithis A and Lionis C. Designing a multifaceted quality improvement intervention in primary care in a country where general practice is seeking recognition: the case of Cyprus. *BMC Health Service Research* 2008; 8:181.
4 O'Mullane M, McHugh S and Bradley CP. Informing the development of a national diabetes register in Ireland: a literature review of the impact of patient registration on diabetes care. *Informatics in Primary Care* 2010;18(3): 157–68.
5 Belsey J, de Lusignan S, Chan T, van Vlymen J and Hague N. Abnormal lipids in high-risk patients achieving cholesterol targets: a cross-sectional study of routinely collected UK general practice data. *Current Medical Research and Opinion* 2008;24(9):2551–60.
6 Saxena S, Car J, Eldred D, Soljak M and Majeed A. Practice size, caseload, deprivation and quality of care of patients with coronary heart disease, hypertension and stroke in primary care: national cross-sectional study. *BMC Health Service Research* 2007 Jun 27;7:96.
7 Debar S, Kumarapeli P, Kaski JC and de Lusignan S. Addressing modifiable risk factors for coronary heart disease in primary care: an evidence-base lost in translation. *Family Practice* 2010 Aug;27(4):370–8.
8 de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D and Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. *Diabetes Medicine*

2012 Feb;29(2):181–9. doi: 10.1111/j.1464–5491.2011.03419.x
9 Gruber T. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies* 1995;43(4–5):907–28.
10 Smith B and Ceusters W. Ontological realism: a methodology for coordinated evolution of scientific ontologies. *Applied Ontology* 2010;5:139–88.
11 Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M *et al*. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS* 2006 Summer;10(2):185–98.
12 McDaniel G (Ed). *IBM Dictionary of Computing*, 10th edition. New York: McGraw-Hill, 1994.
13 Anderson J. Data dictionaries – a way forward to write meaning and terminology into medical information systems. *Methods of Information in Medicine* 1986 Jul; 25(3):137–8.
14 Navathe S and Kerschber L. Role of data dictionaries in information resource management. *Information and Management* 1986:10(1);21–46.
15 de Lusignan S. Codes, classifications, terminologies and nomenclatures: definition, development and application in practice. *Informatics in Primary Care* 2005; 13(1):65–70.
16 de Lusignan S and Teasdale S. Achieving benefit for patients in primary care informatics: the report of an international consensus workshop at Medinfo 2007. *Informatics in Primary Care* 2007;15(4):255–61.
17 de Lusignan S and van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice* 2006; 23(2):253–63.
18 de Lusignan S, Chan T, Parry G, Dent-Brown K and Kendrick T. Referral to a new psychological therapy service is associated with reduced utilisation of healthcare and sickness absence by people with common mental health problems: a before and after comparison. *Journal of Epidemiology and Community Health* 2011 Oct 3. [Epub ahead of print].
19 Park J-R. Metadata quality in digital repositories: a survey of the current state of the art. *Cataloging & Classification Quarterly* 2009;47(3–4):213–28.
20 Sadek AR, van Vlymen J, Khunti K and de Lusignan S. Automated identification of miscoded and misclassified cases of diabetes from computer records. *Diabetes Medicine* 2011 Sep 14. doi: 10.1111/j.1464–5491.2011.03457.x.
21 van Vlymen J and 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(4):281–91.
22 Michalakidis G, Kumarapeli P, Ring A, van Vlymen J, Krause P and 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. *Studies in Health Technology and Informatics* 2010;160(Pt 1):724–8.
23 Parr TJ and Quong RW. ANTLR: a predicated-LL(k) parser generator. *Software—Practice & Experience* 1995; 25(7):789–810. doi: 10.1002/spe.4380250705.
24 Moreira J, Midkiff S, Gupta M and Lawrence R, IBM T. J. Watson Research Center. High Performance Computing with the Array Package for Java: A Case Study using Data Mining. Supercomputing, ACM/IEEE 1999 Conference 1999;10. doi: 10.1109/SC.1999.10025.
25 de Lusignan S, Liaw ST, Krause P, Curcin V, Vicente MT, Michalakidis G *et al*. Key concepts to assess the readiness of data for international research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation. Contribution of the IMIA Primary Health Care Informatics Working Group. *Yearbook of Medical Informatics* 2011;6(1):112–20.
26 Rollason W, Khunti K and de Lusignan S. Variation in the recording of diabetes diagnostic data in primary care computer systems: implications for the quality of care. *Informatics in Primary Care* 2009;17(2):113–19.
27 Williams JG. Measuring the completeness and currency of codified clinical information. *Methods of Information in Medicine* 2003;42(4):482–8.
28 Thiru K, Hassey A and Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ* 2003;326(7398):1070.
29 Brown PJ and Warmington V. Info-tsunami: surviving the storm with data quality probes. *Informatics in Primary Care* 2003;11(4):229–33; discussion 234–7.
30 de Lusignan S, Valentin T, Chan T, Hague N, Wood O, van Vlymen J and Dhoul N. Problems with primary care data quality: osteoporosis as an exemplar. *Informatics in Primary Care* 2004;12(3):147–56.
31 Stevens PE, Farmer CK and de Lusignan S. Effect of pay for performance on hypertension in the United Kingdom. *American Journal of Kidney Disease* 2011 Oct;58(4):508–11.
32 Duncan M. Medical terminology version control discussion paper: The chocolate teapot (Version 2.3). URL: www.mrtablet.demon.co.uk/chocolate_teapot_lite.htm
33 HL-7 Connection. Chocolate Teapot Not Otherwise Classified. URL: www.hl7connection.com/2011/05/chocolate-teapot-not-otherwise-classified/
34 Bertaud-Gounot V, Duvauferrier R and Burgun A. Ontology and medical diagnosis. *Informatics for Health and Social Care* 2012;37(1):22–32.
35 Liaw ST, Chen HY, Maneze D, Taggart J, Dennis S, Vagholkar S *et al*. Health reform: is current electronic information fit for purpose? *Emergency Medicine Australasia* 2011 (Sep): doi: 10.1111/j.1742–6723. 2011.01486.x
36 de Lusignan S and Krause P. The Hayes principles: learning from the national pilot of information technology and core generalisable theory in informatics. *Informatics in Primary Care* 2010;18(2):73–7.

ADDRESS FOR CORRESPONDENCE

Simon de Lusignan
Professor of Primary Care and Clinical Informatics
Department of Health Care Management and Policy
University of Surrey
Guildford
GU2 7XH
UK
Email: s.lusignan@surrey.ac.uk