Building Capacity for Health Informatics in the Future
F. Lau et al. (Eds.)

doi:10.3233/978-1-61499-742-9-29

Mapping Local Codes to Read Codes

Wilfred BONNEY [a,1], James GALLOWAY [a], Christopher HALL [a], Mikhail GHATTAS [a], Leandro TRAMMA [a], Thomas NIND [a], Louise DONNELLY [a], Emily JEFFERSON [a] and Alexander DONEY [a]
[a] *Health Informatics Centre, University of Dundee, Dundee, Scotland, United Kingdom*

Abstract. *Background & Objectives*: Legacy laboratory test codes make it difficult to use clinical datasets for meaningful translational research, where populations are followed for disease risk and outcomes over many years. The Health Informatics Centre (HIC) at the University of Dundee hosts continuous biochemistry data from the clinical laboratories in Tayside and Fife dating back as far as 1987. However, the HIC-managed biochemistry dataset is coupled with incoherent sample types and unstandardised legacy local test codes, which increases the complexity of using the dataset for reasonable population health outcomes. The objective of this study was to map the legacy local test codes to the Scottish 5-byte Version 2 Read Codes using biochemistry data extracted from the repository of the Scottish Care Information (SCI) Store. *Methods*: Data mapping methodology was used to map legacy local test codes from clinical biochemistry laboratories within Tayside and Fife to the Scottish 5-byte Version 2 Read Codes. *Results*: The methodology resulted in the mapping of 485 legacy laboratory test codes, spanning 25 years, to 124 Read Codes. *Conclusion*: The data mapping methodology not only facilitated the restructuring of the HIC-managed biochemistry dataset to support easier cohort identification and selection, but it also made it easier for the standardised local laboratory test codes, in the Scottish 5-byte Version 2 Read Codes, to be mapped to other health data standards such as Clinical Terms Version 3 (CTV3); LOINC; and SNOMED CT.

Keywords. Clinical Datasets, Read Codes, Data Mapping, Health Data Standards

1. Introduction

Advancements in Health Informatics have led to an increasing recognition that routinely collected clinical datasets are valuable resources for research relating to population health outcomes. However, clinical datasets often involve continually changing codes and data standards not only over time, but also by place with many centres adopting variable protocols. In order to optimise the research use of such datasets, semantic interoperability is required to achieve meaningful exchange across time, care settings, datasets, and standards [1]. Data mapping is one approach adopted by many healthcare providers and organisations to support semantic interoperability and meaningful exchange of healthcare data across the continuum of different care settings and providers [1-3]. The International Organization for Standardization Technical Committee on Health Informatics (ISO/TC 215) [4] defined mapping as the "process of associating concepts or terms from one coding system to concepts or terms in another coding system and defining their equivalence in accordance with a

[1] Corresponding Author: Wilfred Bonney, Email: w.bonney@dundee.ac.uk

documented rationale and a given purpose" (p. 1). More importantly, the purpose or use case of data mapping is to develop links between concepts within one source dataset to the same or substantially similar concepts in another target dataset [1, 3].

Clinical data mapping is very important because there is no single medical terminology for primary care data, making it difficult to understand and translate meanings across the different heterogeneous data sources and terminologies developed for different uses of healthcare data [3]. Data mapping seeks to link the variety of clinical code sets, classification and terminology systems that are often used to document a wide range of clinical and administrative content in healthcare delivery [3]. This mapping linkage provides a common medical language necessary for recording structured data in healthcare information systems as well as supporting the generation of high quality data reports [2]. Mapping is important when routinely collected primary care data are coded in a specific format and the same data are needed for a different purpose [1, 3, 4].

The Health Informatics Centre (HIC) at the University of Dundee functions as a data research portal providing clinical data, from the National Health Service (NHS) Scotland, for use by researchers through psuedonymised extracts of cohorts with appropriate governance approvals. HIC hosts continuous biochemistry data from the clinical laboratories in Tayside and Fife dating back as far as 1987. In common with probably most Health Boards, the raw data are encoded with local specific test codes which have often varied over this time period. For example, in Tayside alone, the test codes for *Serum alkaline phosphatase* have changed over the years with varied codes such as AALKP, ALK_PHOS, ALP, ALPH, AP, MAALKP, and MALKP. The objective of this study was to map the legacy laboratory test codes to the Scottish 5-byte Version 2 Read Codes using a sample of the most recent data in SCI-Store and standardise them across two different Health Boards in Scotland so as to achieve semantic interoperability. The mapped Read Codes were then employed as a toolkit to restructure the HIC-managed biochemistry dataset, thereby, facilitating easier cohort identification and selection.

2. Read Clinical Codes

The Read Clinical classification, also known as Read Codes, is a comprehensive computer-based classification, nomenclature and coding system for medicine [5]. Read Codes were initially developed, in the early 1980s, by Dr. James Read, a general practitioner from Loughborough working with Abies Informatics Ltd. [5-7]. Read Codes comprise of a super-set of existing health data standards such as International Classification of Diseases, Injuries and Causes of Death (ICD-9), the International Classification of Diseases Clinical Modification (ICD 9-CM), the International Classification of Procedures In Medicine (ICPM), British National Formulary (BNF), the Office of Population Censuses and Surveys classification of surgical operations and procedures (OPCS 4), and OPCS Classification of Occupations [5, 8, 9].

The introduction of Read Codes in general practice revolutionised General Practitioners' (GPs) computing in the UK. According to Benson [8], the Read Codes have been used by all GPs in the UK. This increasing use of Read Codes by GPs might also help explain why GPs use computers and hospital doctors do not in the UK [10, 11]. The success of Read Codes is often attributed to the fact that the codes were written by a single responsible GP for GPs, and thus considered fit-for-purpose [8].

Also, Systematized Nomenclature for Medicine – Clinical Terms (SNOMED CT) is a merger of the Read Codes Version 3 with SNOMED RT, which is the original SNOMED reference terminology developed by the American College of Pathologists [8, 12, 13].

Read Codes were designed for use with computers and are easy to implement in software applications. No paper version of Read Codes was ever published, as they are intended to be used in recording clinical information on computers, thereby, facilitating multiple updates and extensions [7-9]. The first version of Read Codes, which was also known as the 4-byte GP set, was adopted by the British National Health Service in April 1990. The Read Codes Version 2, which became known as the 5-byte or unified set, extended the utility of Read Codes across the boundary between primary care and hospitals to provide a mechanism for hospitals in cross-mapping their data to ICD-9 [7, 14]. Although there exists Read Codes Version 3 (i.e. Clinical Terms Version 3 (CTV3)) [7, 8, 12], the version used in this study is the Scottish 5-byte Version 2 Read Codes. The Scottish 5-byte Version 2 Read Codes is the de facto standard for Scotland [15]. The CTV3 are used predominantly in England [15].

3. Materials & Methods

3.1. SCI-Store Data

The SCI-Store data was derived from the Scottish Care Information (SCI) repository, also known as SCI-Store. SCI-Store [16] is a national "data repository which retains patient information at a Health Board level" (p. 1). The SCI-Store web services interface provides access to several laboratory report types, including Biochemistry, Haematology, Pathology, Microbiology, and Radiology. SCI-Store is used in each of the 15 NHS Health Board areas within Scotland as the area Master Patient Index (MPI), and generally contains one Electronic Health Record (EHR) per patient for the given Health Board area [16]. For the purposes of this study, the biochemistry laboratory report of individuals from Fife and Tayside Health Board areas were considered. In most recent years (i.e. October 2004 for Fife and January 2007 for Tayside), the SCI-Store data are already encoded with Read Codes.

Biochemistry sample data were pooled randomly from the SCI-Store Version 8.1 XML feed. Extracted data for Fife were between the years 2004 and 2013, inclusive. Similarly, for Tayside, the extracted data were from the year period 2007-2013. Biochemistry data could not be retrieved from SCI-Store for Fife and Tayside for the years earlier than 2004 and 2007 respectively. The pooled sample data were then aggregated with their corresponding Read Codes and exported to QlikView [17] for statistical analysis and interpretation. The SCI-Store biochemistry data was used as a trusted source of correct patient laboratory data for both Fife and Tayside Health Boards.

3.2. Tayside and Fife Legacy Data

The Tayside and Fife legacy data consisted of biochemistry data held concurrently in different legacy laboratory systems including: Win Carter (1987-1993); Pinnacle (1993-1998); and Master Labs (1998-2007). The legacy data were also aggregated and exported to QlikView. The complete legacy incorporated data captured from SCI-Store,

as well as data previously loaded/received from legacy laboratory systems' data dumps. The aggregated legacy data on QlikView provided a reliable platform to map redundant HIC local test codes to unique Read Codes.

3.3. Data Mapping

The data mapping process utilised the best practice guidelines published by the American Health Information Management Association (AHIMA). AHIMA [1] specified six basic steps for mapping data contained in the repository of EHRs. These six steps involved: (a) developing a business case; (b) defining a specific use case; (c) developing heuristics/rules for implementation; (d) planning a pilot phase to test the rules; (e) developing full content with periodic testing; and (f) communicating with source and target data owners. As part of this study, all the six steps in the best practice guidelines were implemented. Abhyankar et al. [18] and Bonney et al. [19] utilised similar steps in their mapping of clinical laboratory data to LOINC.

The data mapping exercise concentrated on the top biochemistry analytes used most frequently by researchers. These top analytes were mapped individually to the Scottish 5-byte version Read Codes using the aggregated data on QlikView. Analytes, sample types, and laboratory test codes were grouped under existing Read Codes to create a master *source-of-truth* table as shown in Table 1. Table 1 was subsequently referred to facilitate mapping to legacy test codes and sample types, which provides a representative set of the data. The NHS Clinical Terminology Browser v1.04 (with the Scottish 5-byte Version 2 Read Codes) was used as a toolkit in selecting the appropriate descriptions for each of the identified Read Codes from SCI-Store. This was necessary because some of the Read Codes descriptions were not consistent with the actual Read Codes values.

3.4. Consolidation of Local Test Codes

A consolidation process was conducted, in consultation with domain experts and biochemists in NHS Tayside and Fife, to map most of the existing legacy test codes to the Scottish 5-byte Version 2 Read Codes. The consolidation process ensured that all existing legacy laboratory test codes of interest that were not retrievable from SCI-Store were also included and considered in the review process. The inclusion and exclusion criteria, for the consolidation process, involved manual review of the frequency distributions and mean averages of each individual analyte or test code with clinical expertise and/or biochemists. The review utilised: (a) the business intelligence of QlikView to quickly highlight issues and gaps in the data; and (b) the experiential knowledge of the domain experts in interpreting the output to derive the underlying semantic meaning of the biochemistry data. Validation was through QlikView models using mean values and frequency distributions to ensure test codes were correct. Researchers used the data to assess biological accuracy in conjunction with domain experts and legacy metadata in relation to recalibration events.

4. Results

4.1. SCI-Store

The extracted biochemistry data from SCI-Store contained, in total, 2,093,087 biochemistry results for 134,627 unique individuals/patients. 838 distinct test codes were identified from the pooled sample data. *Serum creatinine* (44J3.), *Serum sodium* (44I5.) and *Serum potassium* (44I4.) were found to be the top three Read Codes from the pooled sample data from SCI-Store. Note that the tenth test code (i.e. *Glomerular filtration rate calculated by abbreviated Modification of Diet in Renal Disease Study Group calculation adjusted for African American origin* (451G.)) was only used in the Tayside region of Scotland, but not Fife.

Critical review and analysis of the Fife and Tayside biochemistry data from SCI-Store revealed significant differences in the use of sample types in Fife and Tayside regions of Scotland. The Fife biochemistry data was more consistent with the assigned Read Codes values than that of the Tayside biochemistry data. Whereas Fife used mostly Blood, Serum and Urine as sample types; Tayside mostly used Blood, Fluoride Oxalate, and Urine as sample types. The Fife biochemistry data did not contain Fluoride Oxalate as a sample type. In contrast, the Tayside biochemistry data never used Serum as a sample type, even though the Read Codes equivalents of the results were reported as originating from either serum or plasma. For example, in glucose measurement, Tayside assigned Fluoride Oxalate as the sample type, but the result was assigned with a Read Code value of "44g..", indicating that it is a Plasma glucose level.

4.2. Data Mapping

The data mapping methodology resulted in the mapping of 485 legacy biochemistry test codes, spanning 25 years, to 124 Read Codes. Since the mapping concentrated on the top analytes frequently used by researchers, none of the test codes was ever discarded. Also, all legacy test codes related to the top analytes were mapped correctly with less ambiguity. In the case of Serum creatinine, the inference that was derived from the aggregates was that the HIC local test codes: CR, CRE, Creatinine, and DCRE could be represented by the Read Code value "*44J3.*". In other words, if the HIC local test codes contain CR, CRE, Creatinine, and DCRE; and the *Sample Type* is Serum for Fife and Blood for Tayside; and the *Unit* is umol/L; then they can be assigned a Read Code value of "*44J3.*". Similarly, local test codes: GL, GLP, and Glucose were assigned the Read Code value of "*44g..*".

4.3. Consolidation of Local Test Codes

Whereas the review of the frequency distributions of the test codes enabled identification of gaps and subsequent identification of additional missing test codes for the Read Codes mapping, the review of the mean averages enabled the confirmation of the alignment of each single test results.

4.4. Restructuring of HIC-managed Biochemistry Dataset

The results from the Read Codes mapping exercise (shown in Table 1) were used to restructure and standardise the release of the HIC-managed biochemistry dataset. Table 1 shows an instant comparison of the unstandardised dataset versus the standardised format. In this example, three key unstandardised fields: *Sample type, Test_code* and *Description*, were being replaced with standardised unique sample types with their associated Scottish 5-byte Version 2 Read Codes values.

Table 1. Sample of HIC-managed biochemistry dataset restructured using the Scottish 5-byte Version 2 Read Codes

Unstructured and Unstandardised Data Format	Standardised and Restructured Data Format
Prochi|Sample_date|Sample_type|Test_code|Description|Result|Source_code|Lab_system|CTC_source	Prochi|Sample_date|**Sample_type**|**Read_code**|Result|Source_code|Lab_system|CTC_source
1. ABC5348143|12/12/2008|B - FO|CRE|CREATININE|62|STANLEY|6|6	1. ABC5348143|12/12/2008|**SERUM**|**44J3**.|62|STANLEY|6|6
2. ABC1932792|30/07/2008|FO – B|HB1C|HbA1c|6.2|WHPER1|6|6	2. ABC1932792|30/07/2008|**BLOOD**|**42W4**.|6.2|WHPER1|6|6
3. ABC8075943|13/04/2008|B - FO|AP|ALK. PHOSPHATASE|109|N22OP|6|6	3. ABC8075943|13/04/2008|**SERUM**|**44F..**|109|N22OP|6|6
4. ABC6172182|28/09/2008|B|K|POTASSIUM|4.4|N9|6|6	4. ABC6172182|28/09/2008|**SERUM**|**44I4.**|4.4|N9|6|6
5. ABC7294046|23/05/2000|NULL|UCRE|URINE CREATININE|14.7|N22OP|6|6	5. ABC7294046|23/05/2000|**URINE**|**46M7**.|14.7|N22OP|6|6
6. ABC9990519|11/01/2001|NULL|GLPF| FASTING GLUCOSE|4.6|PARK|6|6	6. ABC9990519|11/01/2001|**PLASMA**|**44g1.**|4.6|PARK|6|6

5. Discussion & Conclusion

Data mapping methodology presents a gold standard and an easier approach for standardising clinical datasets for secondary data uses [1]. Mapping data elements in EMRs to a reference classification and/or terminology system not only facilitate reuse of primary care data for multiple purposes, but they also support data analysis, health information exchange and interoperability, and data comparison across the continuum of different healthcare providers [1, 3]. More importantly, data mapping improves the quality of the research output derived from EMRs. It is in this regard that Hammond et al. [14] asserted that health data standards are required "when excessive diversity creates inefficiencies or impedes effectiveness" (p. 212).

This study has demonstrated how data mapping methodology could be used to standardise legacy test codes as well as restructure clinical datasets and to provide consistency when comparing data from different Health Boards. The study established a baseline approach for standardising and restructuring of the HIC-managed biochemistry dataset with the Scottish 5-byte Version 2 Read Codes, which is the de facto standard for Scotland. The applied methodology has improved the efficient grouping of related or common biochemistry test codes, thereby, facilitating easier cohort identification and selection. The restructuring approach has also resulted in the release of the HIC-managed biochemistry dataset in a format that is coherent and standardised for use in translational research, as Read Codes are optimised for secondary data uses [2]. For example, the methodology has allowed continuous analytes to be monitored over a period of 25 years. The methodology is also important not only for generating researchable data from routine NHS healthcare data, but also for linking researchable data (in some cases going back 25 years) to very large genomic resources for translational research.

Although there is no-one-size-fits-all mapping for healthcare data [1], the key aspect of the applied mapping methodology is that the standardised legacy test codes, in the Scottish 5-byte Version 2 Read Codes, can now be easily mapped to other health data standards such as Read Codes Version 3 (i.e. CTV3); LOINC; and SNOMED CT. The full mapping of the test codes from the Scottish 5-byte Version 2 Read Codes to SNOMED CT will be performed as part of future work, as SNOMED CT is considered to become the terminology standard of choice covering both primary and secondary care [15]. Moreover, the applied data mapping and restructuring methodology will be replicated for other HIC-managed laboratory datasets such as Haematology, Immunology, Microbiology, and Virology. In the near future, HIC could release the biochemistry dataset in any of the four standardised formats (i.e. Scottish 5-byte Version 2 Read Codes, CTV3, LOINC, and/or SNOMED-CT), if so desired by researchers.

There are continuing national efforts to create a single standardised laboratory dataset across different Health Boards in Scotland. The work presented here to standardise biochemistry data across two Health Boards could be utilised by this national project, as the study makes it easier to participate in geographically distributed research projects, where coding standards are diverse. The study also adds to the growing body of knowledge in Health Informatics literature that existing health data standards capture most of the clinical concepts used routinely in healthcare delivery.

6. Acknowledgements

This work was supported by the Medical Research Council (MRC) grant number MR/M501633/1 and the Wellcome Trust grant number WT086113 through the Scottish Health Informatics Programme (SHIP). SHIP is a collaboration between the Universities of Aberdeen, Dundee, Edinburgh, Glasgow and St Andrews and the Information Services Division of NHS Scotland. The authors acknowledge the support from the UK Health Informatics Research Network and the Farr Institute of Health Informatics Research. The authors also acknowledge the support of Dundee University Medical School.

References

[1] AHIMA, Data mapping best practices. *J AHIMA* **82**(4) (2011), 46-52.
[2] K. E. Campbell, K. Giannangelo. Language barrier: Getting past the classifications and terminologies roadblock. *J AHIMA* **78**(2) (2007), 44-6, 48.
[3] M. Foley, et al., Translation, please: Mapping translates clinical data between the many languages that document it. *J AHIMA* **78**(2) (2007), 34-38.
[4] International Organization for Standardization (ISO/TC 215), *Health informatics -- Principles of mapping between terminological systems*. 2014, ISO: Geneva, Switzerland. p. 38.
[5] J. D. Read, T. J. Benson. Comprehensive coding. *Br J Health Care Comput* **May** (1986), 22-25.
[6] J. D. Read. Computerizing medical language. *Br J Health Care Comput* (1990), 203-208.
[7] N. Booth. What are the Read Codes? *Health Libr Rev* **11**(3)(1994), 177-182.
[8] T. Benson. The history of the Read Codes: the inaugural James Read Memorial Lecture 2011. *Inform Prim Care* **19**(3) (2011), 173-82.
[9] J. Chisholm. The Read clinical classification. *BMJ* **300**(6732)(1990), 1092.
[10] T. Benson. Why general practitioners use computers and hospital doctors do not - Part 1: incentives. *Br Med J* **325**(7372)(2002), 1086-1089.

[11] T. Benson. Why general practitioners use computers and hospital doctors do not - Part 2: scalability. *Br Med J* **325**(7372) (2002), 1090-1093.
[12] M. O'Neil, C. Payne, J. Read, Read Codes Version 3: a user led terminology. *Methods Inf Med* **34**(1-2) (1995), 187-92.
[13] K. A. Spackman, K.E. Campbell, R.A. Côté, SNOMED RT: a reference terminology for health care. *Proc AMIA Symp* (1997), 640-644.
[14] W. E. Hammond, et al., Standards in Biomedical Informatics, in *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, E.H. Shortliffe and J.J. Cimino, Editors. 2014, Springer-Verlag: New York. p. 211-254.
[15] Information Service Division (ISD)., *Coding & Terminology Systems*. 2010 [cited 2015 Jan 27]; Available from: http://www.isdscotland.org/Products-and-Services/Terminology-Services/Coding-and-Terminology-Systems/.
[16] Scottish Care Information (SCI)., *SCI Store Product Overview*. [cited 2015 Jan 27]; Available from: http://www.sci.scot.nhs.uk/products/store/General/SCI%20Store%20-%20Product%20Description.pdf.
[17] QlikView, *QlikView Overview*. [cited 2015 Feb 12]; Available from: http://www.qlik.com/us/explore/products/overview.
[18] S. Abhyankar, D. Demner-Fushman, C.J. McDonald, Standardizing clinical laboratory data for secondary use. *J Biomed Inform* **45**(4) (2012), 642-50.
[19] W. Bonney, A. Doney, E. Jefferson, Standardizing biochemistry dataset for medical research, in *Proceedings of HEALTHINF 2014: International Conference on Health Informatics*, M. Bienkiewicz, et al., Editors. 2014, SciTePress: Angers, France. p. 205-210.