Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-151

Automated Differentiation of Incident and Prevalent Cases in Primary Care Computerised Medical Records (CMR)

Nadia SMITH[a,1], Valerie LIVINA[a], Rachel BYFORD[b], Filipa FERREIRA[b], Ivelina YONOVA[b,c], and Simon de LUSIGNAN[b,c]

[a]*Data Science Department, National Physical Laboratory, UK*
[b]*Department of Clinical and Experimental Medicine, University of Surrey, UK*
[c]*Royal College of General Practitioners, London, UK*

Abstract. Identifying incident (first or new) episodes of illness is critical in sentinel networks to inform about the seasonal onset of diseases and to give early warning of epidemics, as well as differentiating change in health service utilization from change in pattern of disease. The most reliable way of differentiating incident from prevalent cases is through the clinician assigning episode type to the patient's computerized medical record (CMR). However, episode type assignment is often made inconsistently. The objective of this collaborative study between the Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC), University of Surrey and the National Physical Laboratory (NPL) is to develop a methodology to reconstruct missing or miscoded episode types. The data, gathered from the RCGP RSC network of over 230 practices, are analyzed and poor episode typing reconstructed by disease type. The methodology is tested in practices with good episode type data quality. This method could be used to improve prediction of epidemics, and to improve the quality of historical rates retrospectively.

Keywords. Episode type, general practice, medical record systems, computerized.

1. Introduction

The Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) [1] is an internationally renowned source of information, analysis and interpretation of primary care data. In 2015, it established a new data and analysis hub at the University of Surrey.

Established in 1957, the RSC is an active research and surveillance unit which has been collecting and monitoring respiratory infection and other conditions largely managed in primary care since 1967 [2,3]. A key focus is influenza along with other diseases, collected from over 230 practices across England, and published online as the RCGP Weekly Returns Service (WRS). The WRS monitors the number of patients consulting with incident (first or new) episodes of illness classified by diagnosis in England, and is the key primary care element of the national disease monitoring systems run by Public Health England (PHE); its analyses are also critical for assessing flu vaccine effectiveness [4].

[1] Corresponding Author, Nadia Smith, National Physical Laboratory, Hampton Road, TW11 0LW Middlesex, UK. E-mail: nadia.smith@npl.co.uk.

Identifying incident cases is important for predicting seasonal onset of diseases and early warning of any epidemic at regional and national levels, and for monitoring vaccine uptake and effectiveness. The way to differentiate incident cases from prevalent cases is through the clinician assigning an episode type to a problem title in the patient's computerized medical record (CMR). Episode type (first, new or ongoing) allows new and first diagnoses to be identified for separate analysis. The ability to record episode type was mandated for general practice CMR systems in England, of all brands. Problem titles are recorded using the Read classification [5, 6].

The RCGP RSC is currently working with practices by providing them with feedback and guidance on the importance of good quality entry of medical problems, in which episode typing is key to ensure patients records are of as high quality as possible. This year the RSC has launched a data quality dashboard. Poor quality data in primary care can have many negative consequences at different levels: practice patient care may be compromised, failure to call/recall patients may occur, and practice income may suffer (within a practice); the beginning of epidemics may be missed or the wrong conclusions may be drawn about vaccine benefits/risks (within the RCGP RSC's surveillance programme); decisions about funding and resource allocation for epidemiological research may be mistakenly made based on incorrect evidence and trends.

In the respiratory and other disease surveillance studies that the RCGP RSC currently carries out weekly, a percentage of practices is excluded as the data from those practices do not comply with the quality checks for episode typing. In this study a methodology has been developed by data scientists from the National Physical Laboratory (NPL), to assign missing episode types and reassign episode types judged to be miscoded. The methology has been tested in practices with good episode type data quality.

2. Methods and Analysis

A CMR data analysis has been performed on the weekly data received from practices by the RCGP RSC (all are English General Practices), and a methodology for data mining and episode type reconstruction has been developed and tested on this data set. The methodology has been refined taking into account clinical knowledge.

2.1. Data set

The data have been gathered by the RCGP RSC network (N > 2.3 million > 230 General Practices) that is representative of the national population on a variety of domains, both demographic and clinical [1]. These data are uploaded twice weekly (each extract comprises a 6-week increment of data – the current week plus the previous 5 weeks) to the RSC database and is retained. Data are stored from all consultations which are coded according to encounter, diagnostic and therapeutic procedures, drugs prescribed, and other comorbidities among others (within the RCGP RSC network only coded data is analysed).

2.2. The RCGP RSC Weekly Returns Service

The WRS monitors the number of patients consulting with new episodes of illness classified by diagnosis in England and provides weekly incidence rates per 100,000 population for these new episodes of illness.

The RCGP RSC performs checks to ensure the data quality, which is extremely important as outputs may be biased if the sample is not representative or if there is uncertainty around the denominator [7]. The two main checks performed are of the numerator (number of cases of a particular disease) and denominator (total number of patients) used to calculate the rates:

- Monitored events below threshold (numerator): calculated by practice, the minimum threshold rate for the 37 conditions monitored in the WRS is 0.001. This rate is calculated as the total number of patients coming in for first/new consultations for a particular condition divided by all the patients in that practice. If the threshold rate is lower than 0.001, the practice is excluded from the WRS. A consultation is defined as first or new by checking the episode type.
- Denominator discrepancy: calculated by practice, the number of valid patients received in the data compared to the published practice list size (this is published annually by the Health and Social Care Information Centre, HSCIC). If the list of patients is 15 % over or under this size, the practice is excluded from the WRS.

At present around 15 to 20 practices (an average of 50,000 to 150,000 patients) are excluded every week from the data analysis that is published in the WRS as a result of not complying with these data quality checks, most of the exclusions being due to the numerator check, i.e. miscoded or missing episode typing.

2.3. Reconstructing miscoded or missing episode type

An algorithm has been developed to reconstruct missing or miscoded episode types in the patients' CMR. In collaboration with primary care clinicians, relevant factors have been chosen (the time interval between two independent first/new events, and whether it is a chronic condition or not) to identify the episode type (first, new or ongoing) of a patient consulting with one of the monitored diseases. The algorithm analyses data from one or several overlapping 6-week increments, searching for all events of the monitored diseases and calculates an episode based on a difference in dates between two consecutive events of the same patient. The algorithm adds missing and updates miscoded episode types for 32 out of 37 of the diseases monitored by the RCGP RSC.

The remaining 5 of 37 diseases were excluded as they were considered to be a category too broad to give a time interval between two new episodes and to decide on whether they were a chronic condition (e.g. symptoms involving musculoskeletal problems). Further information such as individual code level and patient history would be required to be able to decide the relevant factors for these diseases.

The time interval between two independent first/new events varies for each disease monitored, with the minimum being 10 days for diseases such as 'common cold' and the maximum being 90 days for 'infectious mononucleosis'. Therefore it was decided that the minimum amount of data needed on which to apply and test the algorithm was 90 days' worth of consecutive data.

2.4. Reporting and validating

The results of the algorithm, i.e. the reconstructed episode types, are reported across the whole ranges of practices. As a validation exercise, we also reported the top quartile of best quality of episode typing practices against the remaining three quartiles of practices.

3. Findings to date

The results presented in Table 1 correspond to the outputs of the algorithm when monitoring 32 diseases in week 52 of the year 2016 (chosen as a relevant week due to being in the middle of the flu season) using 13 weeks of aggregated data (from week 40 to week 52, 2016). The first column provides results for all the practices included in the analysis; the second for the top quartile of best quality episode typing practices; and the third for the remaining three quartiles of practices. The first row shows the percentage of the missing episode types (with respect to the total number of events monitored in that increment of data), defined as those that had a blank episode type field in the patient's CMR for the corresponding consultation. The second row shows the percentage of miscoded episode types, defined as those that had a different episode type calculated by the algorithm to that of the patient's CMR for the corresponding consultation. The final row shows the total percentage of updated episode types, defined as the total of episode types that the algorithm has modified with respect to the patient's CMR and can be calculated as the sum of the missing and miscoded episode types.

Table 1. Outputs of the algorithm when monitoring 32 diseases in week 52 of 2016, using 13 weeks of aggregated data (from week 40 to week 52, 2016).

	All the practices	**Top quartile of "best" practices**	**Remaining practices**
% missing episode types	8.13 %	1.52 %	10.99 %
% miscoded episode types	13.83 %	7.42 %	16.62 %
% updated episode types	21.96 %	8.94 %	27.61 %

Looking at the results for all of the practices included in the analysis for this case study, the results are promising, showing nearly 22 % of updated episode types – 8.13 % of which were missing, and 13.83 % of which were miscoded. Furthermore, nearly all of the miscoded events were found to be miscoded in the CMR as 'ongoing', when they were really 'first' or 'new'. This is particularly important for the data quality checks the RSC performs weekly on the data to be analyzed, as with the reconstructed episode types this data could now be included into the weekly analysis.

From a validation point of view, for this particular case study, it can be seen in Table 1 that the episode types calculated by the algorithm match the original ones much better for the top quartile of best quality episode typing practices than for the remaining ones (7.42 % vs 16.62 %). It can be observed also that the top quartile practices have very few incomplete episode types in the data whereas the remaining have many more (1.52 % vs 11 %).

Further validation tests are currently underway to ensure consistent results with other datasets from different dates, and also with data from a longer period of time. Once

fully validated, the algorithm will be used to reconstruct data weekly and historical data will be analysed retrospectively.

4. Conclusions and future collaboration

A methodology for reconstructing episode types by disease type in primary care computerized medical records has been developed in a collaborative study between the Royal College of General Practitioners Research and Surveillance Centre, University of Surrey and the National Physical Laboratory. It has been run and validated on a test dataset gathered from the RCGP RSC database.

Whilst educational programmes, feedback and dashboards can prospectively improve data quality, these activities have no retrospective impact. To harness historic data, methods such as these are badly needed.

The algorithm has been developed in a heuristic fashion, taking into account primary care consultants' knowledge, limiting it therefore to the diseases under study and those for which an episode type can easily be defined based on a difference in dates between two consecutive events.

Once fully validated the methodology has the potential to improve the data quality of the patients' CMR prospectively, therefore allowing the previously excluded records to be included in the analysis carried out for the WRS done by the RCGP RSC. The methodology can also be applied retrospectively, allowing the reconstructed historical datasets to be analyzed and compared to the previously published results in the RCGP RSC weekly returns service, to see if previously excluded data can now be included and the quality of historical rates can be improved retrospectively.

Acknowledgements:

Patients and practices of RCGP RSC who consent to share data. Apollo medical systems and CMR suppliers – EMIS, TPP, In Practice Systems. The University of Surrey and NPL for supporting the work. Ana Correa.

References

[1] A. Correa *et al.*, Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) sentinel network: a cohort profile. *BMJ Open* **20** 6(4) (2016), e011092.
[2] S. de Lusignan *et al.*, RCGP Research and Surveillance Centre: 50 years' surveillance of influenza, infections, and respiratory conditions. *Br J Gen Pract.* **67**(663) (2017), 440-441.
[3] D.M. Fleming *et al.,* Health monitoring in sentinel practice networks: the contribution of primary care. *Eur J Public Health* **13**(3 Suppl) (2003), 80–4.
[4] R. Pebody *et al.,* End-of-season influenza vaccine effectiveness in adults and children, United Kingdom, 2016/17. *Euro Surveill*. **22**(44) (2017), Nov.
[5] S. de Lusignan, Codes, classifications, terminologies and nomenclatures: definition, development and application in practice. *Inform Prim Care*. **13**(1) (2005), 65-70.
[6] S. de Lusignan and C. van Weel, The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice* **23** (2006), 253–263.
[7] C. Souty *et al.*, Improving disease incidence estimates in primary care surveillance systems. *Population Health Metrics* **12** (2014), 19.