Bioinformatics, 32(13), 2016, 2079–2080
doi: 10.1093/bioinformatics/btw078
Advance Access Publication Date: 15 February 2016
Letter to the Editor

Data and text mining

Comment on 'Discovering hospital admission patterns using models learnt from electronic hospital records'. The importance of using the right codes

Guillermo H. Lopez-Campos[1,]*, Fernando Martin-Sanchez[2] and Kathleen Gray[1]

[1]Health and Biomedical Informatics Centre, The University of Melbourne, Parkville, VIC, Australia and [2]Department of Healthcare Policy and Research. Division of Health Informatics. Weill Cornell Medicine, New York, NY, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 2, 2015; revised on December 18, 2015; accepted on December 22, 2015

Contact: guillermo.lopez@unimelb.edu.au

In 'Discovering hospital admission patterns using models learnt from electronic hospital records' Arandjelovic proposes a new model for prediction of hospital admissions based on representing a patient's medical record as a binary history record. There is no doubt that the use of electronic medical records to predict aspects of healthcare such as readmission risk, length of stay or even likelihood of death is an important research area in medical informatics. Such studies have a potentially broad impact on the evaluation and improvement of the quality of care, ranging from assessing clinical risks to measuring the costs associated with them (Cai *et al.*, 2015; McCoy *et al.*, 2015). We agree with Arandjelovic on the relevance of this topic and the volume of publications focused on specific diseases and interventions. Nonetheless, for these very same reasons, we have serious concerns regarding the data handling methods and thus the conclusions reached by Arandjelovic in this article. The dataset used in the analysis is described sketchily, lacks some relevant information that would facilitate the interpretation of this work and contains several other elements of concern.

Initially it was the apparent misconception in the (mis-) use of the codes included in the analyses that alarmed us the most. According to the author this article is based on the analysis of International Classification of Diseases (ICD) codes extracted from clinical records, but the author fails to state clearly what version or versions of the ICD codes was used for the development of the methodology. There are several versions of ICD codes that have been used for the codification of diagnosis/diseases/procedures in health records, and these versions differ in several important ways—such as the use of different codification strategies (e.g. alphanumeric in ICD-10 versus numeric in ICD-9) and the number of chapters. The bibliographic reference provided in this article and its description of the use of alphanumeric characters suggest the use of ICD-10 (http://www.who.int/classifications/icd/en, Accessed October 26 2015), but the number of chapters, 12, does not match the 22 chapters included in ICD-10. Of further concern is the article's selection of the code E62, which it describes as 'respiratory infections/inflammations'. The author places this code, according to its starting letter 'E', under chapter four (IV) of ICD-10 ('Endocrine, nutritional and metabolic diseases'). The conceptual mismatch alone should raise doubts even among readers unfamiliar with ICD-10, and indeed further inspection would reveal that E62 does not exist as an ICD-10 code (http://apps.who.int/classifications/icd10/browse/2016/en, Accessed October 26 2015).

Because of the lack of details about the dataset used and given the aforementioned example, we proceeded to analyze in further detail the rest of the codes provided in this article, comparing them with the ICD-10 codes, to check whether this was an unfortunate error limited to the introduction or a misconception propagated in further sections of the article. Our analysis of the codes in Table 1 of the article, which are said to account for 75% of the analyzed admissions, showed that they are either incorrect—meaning that they do not match the description provided (24 out of 30, 80%)—or non-existent (the remaining 6 out of 30, 20%).

After this initial analysis, we attempted to identify a plausible origin for the codes used in this article, comparing them with other medical informatics codes. The results of our analysis point toward the use of the Australian Refined Diagnosis Related Groups (AR-DRGs). Determining the exact version used is impossible due the removal of the additional letter that builds the code) (https://www.

accd.net.au/Downloads.aspx Accessed October 26 2015). This coding system refers to different episodes of care rather than diagnosis, with a focus on funding and management of the services provided by hospitals, relating patient records to the costs incurred by the hospital. AR-DRG codes are derived from existing hospital data, dynamic, and frequently reviewed (approximately every two years). Although clinically coherent about patient demographics, diagnosis and interventions, these codes are not intended to provide diagnostics insights. They reduce the complexity of ICD codes tremendously, through merging several ICD codes under the same DRG, and they take into account other parameters such as the level of resource utilization and statistical soundness. DRG codes are split according to different variables such as age and separation mode, by adding an additional letter to the code. With that additional letter (e.g. B70A, B70B, B70C, B70D or Z40Z) DRGs include information regarding the complexity of the procedures carried out.

Assuming that the codes in Arandjelovic's Table 1 are DRG codes, the dataset is oversimplified by eliminating the DRG splitting criteria, thus going against the spirit of the DRG codes. Conflating DRGs regardless of this second letter coding would introduce bias in the results; for example, handling both B70A and B70D as B70 would consider 'Stroke and Other Cerebrovascular Disorders, Major Complexity' and 'Stroke and Other Cerebrovascular Disorders, Transferred <5 Days' as the same entity. This is another potential misleading factor when interpreting findings and conclusions from this article's data analysis, because the final results could be affected by the criteria applied in the selection of the DRG code.

Moreover, the wording used in different sections of this article makes it difficult to interpret its contribution, or even to determine whether the described method aims to 'develop a framework which allows the health practitioner to understand the available patient information', 'predict future diagnosis' or 'predict next admission'. In any case the results reported in this article could well be interpreted by future readers operating under the assumption that the correct codification was applied in the analyses. If indeed the analysis in this article has used a coding system that was not designed for diagnostics representation and related purposes, the claims associated with the ability of the method to predict the first future diagnosis or prognosis—as per the results section of the abstract and the conclusion section of the article—are not supported by the data. Instead the method at best would suggest the first future episode of care (which is captured already in DRG codes).

However, even allowing this interpretation and accepting that the aim is hospital admissions pattern discovery in electronic health records using DRG codes, the article still requires a more detailed explanation of the dataset used. For instance, if it spanned a period of years (as could be inferred from the existence of patients with hundreds and thousands of admissions in Figure 4a) it might have used different versions of the DRG codes. Hence the methods should have explained how the consolidation of these different versions was addressed.

The article claims that this system outperforms previous approaches in the literature but does not present any support for this claim. Given our comments above on coding system/s, it might well be that this comparison rests on tools and methods based in analysis of different coding systems for different purposes, making the comparisons of performance meaningless. The only comparisons presented in this article are between the proposed method and a Markov model developed by the same author (similar to those presented in his reference 3, Arandjelovic, 2015a, b).

For us to do a more detailed analysis in the interests of replicating this study would have required access to the original dataset. Despite the journal policy encouraging dataset publication, we understand if the dataset cannot be made available due to ethical and confidentiality issues associated with the secondary use of data extracted from clinical records. For this very reason, it is concerning to note that this article lacks any mention of ethical approval or procedures used for data anonymization, nor even acknowledgement of the original data source.

This article's disregard for the meaning of clinical codes, its loose use of key healthcare terminology (the term 'admission' is indistinctly used referring to codes, admission causes, diagnosis and apparently even for actual hospital admission events) and its critical omission of key information about the health data source all emphasize the need for greater collaboration between the bioinformatics and medical informatics communities to advance research of this kind. This is essential to ensure that there is fundamental understanding of the context where data are derived and where new methods and systems will be applied.

Conflict of interest: none declared.

References

Arandjelovic,O. (2015a) Discovering hospital admission patterns using models learnt from electronic hospital records. *Bioinformatics*, **31**, 3970–3976.

Arandjelovic, O. (2015b) Prediction of health outcomes using big (health) data. *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc*, **2015**, 2543–2546.

Cai,X. *et al.* (2015) Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *J. Am. Med. Inf. Assoc.*, doi:10.1093/jamia/ocv110.

McCoy,T.H. *et al.* (2015) Sentiment Measurement in hospital Discharge Notes is Associated with readmission and mortality risk: an electronic health record study. *PLos One*, **10**. e0136341.