

Automatic Knowledge Extraction from EHRs

Ieva Vasiljeva and Ognjen Arandjelović

University of St Andrews, St Andrews KY16 9SX, United Kingdom

Corresponding author email: ognjen.arandjelovic@gmail.com

Abstract

Increasing efforts in the collection, standardization, and maintenance of large scale longitudinal electronic health care records (EHRs) across the world provide a promising source of real world medical data with the potential of providing major novel insights of benefit both to specific individuals in the context of personalized medicine, as well as on the level of population-wide health care and policy. The present paper builds upon the existing and intensifying efforts at using machine learning to provide predictions on future diagnoses likely to be experienced by a particular individual based on the person's existing diagnostic history. The specific model adopted as the baseline predictive framework is based on the concept of a binary diagnostic history vector representation of a patient's diagnostic medical record. The technical novelty introduced herein concerns the manner in which transitions between diagnostic history vectors are learnt. We demonstrate that the proposed change *prima facie* enables greater learning specificity. We present a series of experiments which demonstrate the effectiveness of the proposed techniques, and which reveal novel insights regarding the most promising future research directions.

1 Introduction

Recent years have witnessed a remarkable convergence of two broad trends. The first of these concerns information i.e. data. Rapid technological advances coupled with an increased presence of computing in nearly every aspect of daily life, have for the first time made it possible to acquire and store massive amounts of highly diverse types of information. Concurrently and in no small part propelled by the described environment, research in artificial intelligence – in machine learning, data mining, and pattern recognition, in particular – has reached a sufficient level of methodological sophistication and maturity to process and analyse the collected data, with the aim of extracting novel and useful knowledge.

Considering the growing appreciation of the financial burden of health care and the associated human cost, it is unsurprising that application domains pertaining to health care

have attracted a significant amount of attention. For example, the diverse information content shared across social media platforms [Abel *et al.*, 2011; Agarwal *et al.*, 2011; Baucorn *et al.*, 2013; Beykikhoshk *et al.*, 2014, 2015c,b; Bollen *et al.*, 2011] is increasingly recognized as a source of valuable insight into the behaviour of individuals, spread of epidemics, and the adoption of health related recommendations and advice. The analysis of medical literature itself is another potential target for data mining algorithms and knowledge extraction [Andrei and Arandjelović, 2016a,b; Beykikhoshk *et al.*, 2015a]. In this paper we are specifically interested in the rich source of heterogeneous health data, collected and stored in the form of large scale longitudinal electronic health records (EHRs) [Arandjelović, 2015a; Christensen and Ellingsen, 2016; Xu *et al.*, 2016].

2 Previous work

Most existing methods constrain their prediction to a narrow specific context, e.g. to admissions to the emergency department [Li and Guo, 2009], to heart failure related admissions [Hammill *et al.*, 2011], to the veteran population [Holloway *et al.*, 1990] etc. In addition to their inherently limited scope the applicability of these methods is further impaired by their frequent reliance on a substantial amount of expert knowledge in the choice of variables used for prediction [Lee-gon *et al.*, 2005]. Notwithstanding these efforts, the performance of the methods described in the literature has largely been disappointing [Kansagara *et al.*, 2011]. Better results have been reported in prediction attempts which simplify the task even further by looking at short-term (usually of approximately 30 days) predictions only [Holman *et al.*, 2005].

One of the possible reasons for the poor performance of the existing methods in the literature lies in their virtually universally overly simplistic inference models. In particular, unlike in the present work, they fail to capture sequential information on historical admissions and diagnoses; rather, they base their predictions on a single cumulative snapshot of a patient's record [Bottle *et al.*, 2006].

The present work builds upon the model recently proposed by Arandjelović [2015b]. Evaluated on a large corpus of real world EHRs the method demonstrated highly promising results on the task of diagnosis prediction. Additional analyses by Vasiljeva and Arandjelović [2016a] and Arandjelović [2016] corroborated the original findings and provided

additional insights regarding the structure of the underlying model, which have led to its further refinements [Vasiljeva and Arandjelović, 2016b].

For completeness and for the sake of clarifying the specific limitations of the original method which the present work addresses, we next summarize the main ideas behind the baseline algorithm we build upon; for full technical detail and additional discussion the reader is referred to the original publications [Arandjelović, 2015a, 2016].

2.1 Original diagnostic history based method

Consider a patient’s hospital diagnosis history H which comprises a sequence of diagnoses d_i :

$$H = d_1 \rightarrow d_2 \rightarrow \dots \rightarrow d_n, \quad (1)$$

where each d_i is a discrete variable whose value is a specific diagnostic code. Examples of diagnosis coding schemes include that provided by the International Statistical Classification of Diseases and Related Health Problems (ICD-10) World Health Organization [2004] and the related Australian Refined Diagnosis-Related Groups (AR-DRGs). The algorithm proposed by Arandjelović [2015a] predicts the most likely next diagnosis d_{n+1}^* by learning the probabilities of transitions from H to all other possible histories which can result from a single follow-up diagnosis d :

$$d_{n+1}^* = \arg \max_{d \in D_{\text{ICD}}} p(H \rightarrow d|H), \quad (2)$$

where D_{ICD} is the set of diagnostic codes. To make the estimation of the probability $p(H \rightarrow d|H)$ tractable and learnable from limited data, a patient’s diagnostic history H is represented using a fixed length binary vector $v(H)$. This representation bears resemblance to the bag of words representation frequently used in text analysis and which has since been successfully adapted to various other application domains too [Sivic and Zisserman, 2003; Arandjelović, 2012]. Each element in $v(H)$ encodes the presence (value 1) or lack thereof (value 0) of a specific salient diagnosis (i.e. the corresponding code) in H , save for the last element which captures jointly all non-salient diagnoses. Saliency is determined by the frequency of the corresponding diagnosis in the entire data corpus (n.b. different saliency criteria can be readily used instead). The probability $p(H \rightarrow d|H)$ in (2) is then estimated by superimposing a Markovian model [Sukkar *et al.*, 2012; Jackson *et al.*, 2003] on the space of history vectors which leads to $H \rightarrow d$ being interpreted as a transition from the state represented by $v(H)$ to the state represented by $v(H \rightarrow d)$. As usual the probabilities parameterizing the Markov model are learnt from a training data corpus. A conceptual illustration of the method is shown in Figure 1.

The key idea behind the described model, supported by a breadth of empirical evidence, is that in a vast number of cases it is the *presence* of past complications which most strongly predicts future ailments [Mudge *et al.*, 2011; Friedman *et al.*, 2008 2009; Dharmarajan *et al.*, 2013; Butler and Kalogeropoulos, 2012], which allows for the space of states over which learning is performed to be reduced dramatically; in particular, this is achieved by employing a fixed length state representation and through binarization of its elements.

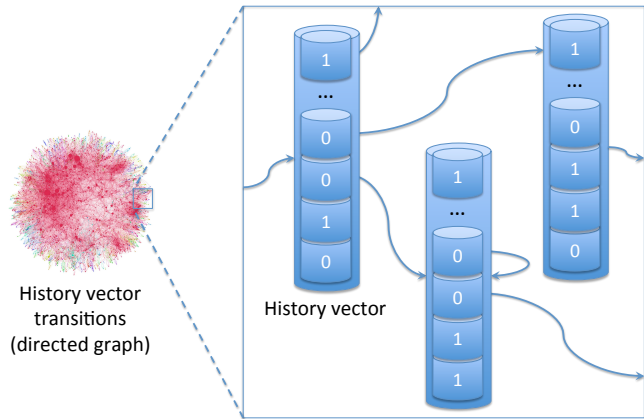


Figure 1: Conceptual illustration of the method proposed by Arandjelović [2015a] which superimposes a Markovian model over a space of history vectors used to represent the medical state of a patient. The inherent ‘forgetfulness’ of the Markovian assumption, which is a major limitation in its use in the application domain of interest in this paper, is overcome by incorporating memory in the state space itself.

3 Technical contributions

In this section we introduce our main technical contribution. A further contribution in the form of novel analyses and empirical results which highlight important and promising future research directions is presented in Section 4.

3.1 Improving the learning model

The first major contribution of the present work goes to the very heart of the learning framework underlying the diagnostic progression model, and concerns the issue of the space over which learning is performed. In other words we propose a paradigm change in terms of what is explicitly learnt.

Recall from the previous section that the method described by Arandjelović [2015a] learns the probabilities of transitions from the space of history vectors to the same space of history vectors i.e. it learns $p(H'|H)$ where H is a patient history vector and H' a possible extension to that history, $H' = H \rightarrow d$. This approach follows naturally from the structure of the problem: both H and H' are states in a Markov chain and indeed the baseline formulation of this class of problems learns amongst other things precisely these transition probabilities. However, the very aspect of the history vector representation which makes it a powerful feature for longitudinal pattern extraction, in this instance introduces a significant practical limitation. Because history vectors are binarized, in general a specific transition does not uniquely determine the diagnosis which caused the transition to occur. In particular this occurs when a diagnosis already recorded in a patient’s history is repeated – the transition from H to itself does not allow the method to distinguish between different diagnoses in the patient’s history and determine which effected the transition. This is a major limitation given that many of the most serious diseases tend to be chronic in nature. In the context of the baseline model described in the

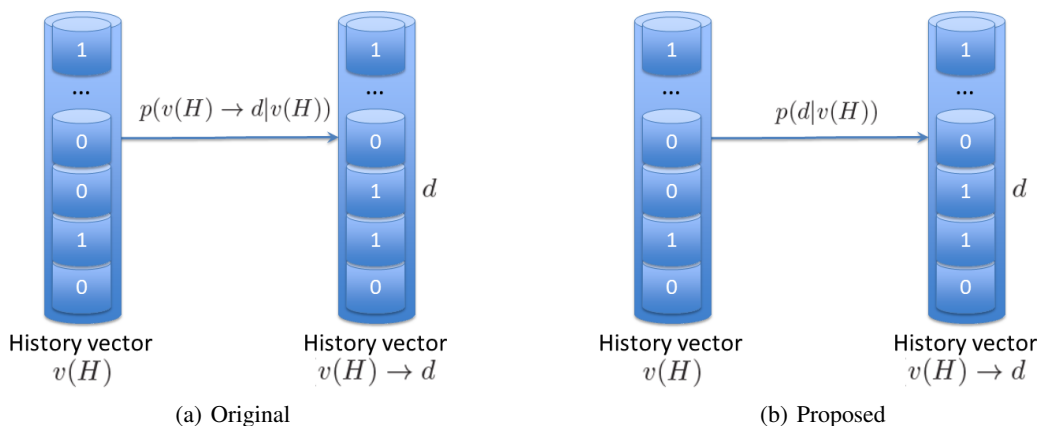


Figure 2: Illustration of the learning paradigm change proposed in the present paper. (a) The original diagnostic history based method of Arandjelović [2015a] learns transition probabilities between different compatible patient histories (histories which correspond to history vectors differing in at most one binary entry). (b) In contrast herein we learn directly the probabilities of all possible diagnostic codes given a specific patient history. This change does not increase the complexity of the learning task at hand while *prima facie* providing greater specificity in prediction.

previous section this means that a repeated diagnostic code effects a transition from a history vector to itself.

The method introduced in the present paper solves the aforementioned problem by changing the space over which learning is performed. In particular, rather than learning the probabilities of transitions between history vectors themselves, we learn the probabilities of follow up diagnoses directly. In other words, rather than learning probabilities of the type:

$$p(v(H) \rightarrow d|v(H)), \quad (3)$$

we propose to learn:

$$p(d|v(H)). \quad (4)$$

It can be readily seen that this is a stronger learning task in the sense that knowing the follow-up diagnosis d allows for the computation of the next Markov chain state $H' = H \rightarrow d$ without ambiguity whereas the opposite is not the case, as described previously.

What makes the proposed learning methodology particularly sensible is that it does not carry the burden of either greater computational complexity nor learning challenge – the dimensionality of the space over which learning is performed stays exactly the same (it is governed by the choice of the number of salient diagnoses), which remains as densely populated as before. Hence this learning paradigm change is unambiguously superior to that described originally.

4 Evaluation and analysis

In this section we summarize some of the experiments we conducted to evaluate the proposed framework and to garner additional useful insight into the structure of its prediction which would help illuminate the most promising avenues for further improvement and future work.

4.1 EHR data

In an effort to reduce the possibility of introducing variability due to confounding variables, we sought to align our evaluation protocol as much as possible with that adopted by previous work. Hence we conducted our experiments using the large collection of EHRs (over 40,000 individuals and over 400,000 diagnostic events) described in Arandjelović [2015b]. For completeness here we summarize the key features of this data set.

The EHRs adopted for evaluation were collected by a large private hospital. The distribution of patient age in the database is 73 ± 15 years, the youngest and oldest patients being 17 months and 102 years old respectively, with the male to female ratio 56 : 44. Approximately 23% of the patients in the database have a date of death associated with their EHR, which means that they are deceased and thus have a record of a terminal diagnosis. The entire EHR collection spans a period of 10 years, with the average number of diagnoses per patient of 10.1 ± 62.2 . Standard ethical procedures in the storage and use of data were followed, and an appropriate institutional ethical approval obtained before the commencement of research.

4.2 A note on performance assessment

Before proceeding with a description the experiments we conducted, we would like to turn our attention briefly to the question of how performance of a method addressing the problem at hand should be measured. It is little short of a truism to note that this issue is crucial in ensuring that the choice between different competing models is made on meaningful basis. [Arandjelović, 2015a] touched upon this discussion but herein we would like to add some further insight.

To place the present discussion in context, let us first summarize how performance assessment was approached in the original work by [Arandjelović, 2015a]. [Arandjelović, 2015a] reported the performance of the baseline diagnos-

tic history based method on two different prediction tasks, namely (i) the first follow-up diagnosis prediction, and (ii) long term prediction of the most likely diagnosis sequence. In both scenarios prediction was made on the basis of the patient’s current diagnostic history.

The accuracy of the first follow-up diagnosis prediction was made initially by examining the proportion of the predictions which were correct (i.e. which correctly matched the actual follow-up diagnosis in the data) and was analysed in further detail using cumulative match characteristic (CMC) curves which capture the prediction accuracy at rank- N for different values of N . In the present paper we would like to put forward an argument that rank-1 accuracy is not a good performance assessment measure for the problem at hand. Our argument is based on a simple observation which we will support further with empirical evidence in Section 4.3. Specifically, we observed that in practice, serious chronic conditions with multiple possible comorbidities which exhibit predictable longitudinal patterns, are interlaced with sporadic diagnoses of more common ailments which are often at most weakly related to the former conditions. This means that by their very nature these sporadic episodes are virtually unpredictable and the failure of a model to predict them should not weigh as heavily as an error in the prediction of chronic conditions. Therefore we would argue that prediction accuracy not at rank-1 level but rather at rank-2 to rank-4, is more insightful and reflects a method’s performance better.

To assess long term prediction accuracy in his original work Arandjelović [2015a] used model likelihood computed using the actual diagnostic progression observed to compare different competing models. This manner of performance assessment can be seen to be less sensitive to the confounding aspects described in the context of follow-up diagnosis prediction. The primary reason is that by its very nature this measure necessitates a comparison between two alternatives (i.e. two model) – the likelihood value itself is not particularly meaningful in answering the question of how good a method is. Rather, it is the comparison of the likelihoods (ratio thereof) of two different models that informs the decision on which of the two should be preferred. In this context, the effects of spurious and ‘unpredictable’ diagnoses cancel out.

4.3 Results and discussion

Using the collection of EHRs described in Section 4.1, we conducted a series of experiments to evaluate the performance of the modified diagnostic history based method proposed in this paper, and validate the different hypotheses that underlie the model. To make our method readily comparable with those already described in the existing literature and the results interpretable in the context of previous findings, we followed the standard cross-validation framework used in previous work.

Experiment 1: Follow-up diagnosis prediction

Following previous work, in our first experiment we sought to examine the performance of the proposed method on the task of follow-up diagnosis prediction. For each possible partial patient history in our data set we compared the prediction made by our algorithm and compared it with the ground truth

i.e. the actual follow-up diagnosis. Our findings are summarized using the standard cumulative match characteristic curve shown in Figure 3.

It can be readily seen that the proposed method achieves impressive accuracy of approximately 78% already for rank-1 prediction. Moreover, from this already high baseline rank- N accuracy increases rather rapidly with N , achieving approximately 89% for rank-3 and 92% for rank-5 prediction. As argued in Section 4.2 these values are more meaningful measures of performance in the context of the problem at hand.

It is interesting to compare the plot in Figure 3 with the analogous plot in [Arandjelović, 2015a]. While these are not directly comparable, given that we directly predict the follow-up diagnosis whereas Arandjelović [2015a] predicts the follow-up history, it is insightful to analyse the magnitude and the nature of the difference between the two CMCs. As expected, rank-1 accuracy in Figure 3 is lower than that in [Arandjelović, 2015a]. However, the difference diminishes for prediction at higher ranks, effectively vanishing by rank-10.

Experiment 2: Coarse diagnosis type prediction

As already noted in the past by Arandjelović [2016] amongst others (also see [RGI-CGHR Collaborators, 2009]), the hierarchical structure of most commonly used diagnosis coding schemes encodes important information which has received little attention to date [Vasiljeva and Arandjelović, 2016b]. Hence, in order to illuminate possible directions for future work in our second set of experiments we sought to garner additional insight into the nature of errors of our algorithm on a coarser coding level. We did this by examining each erroneous prediction in turn and checking whether the prediction concerns the same category of diagnoses as the correct one, as readily inferred from the aforementioned hierarchy [Arandjelović, 2016]. For example, an erroneous prediction of a specific circulatory system diagnosis which predicted a different circulatory system diagnosis can be reasonably considered as being less serious than an erroneous prediction which predicted, say, a disease of the skin and subcutaneous tissue.

Our results are highly insightful. Already at rank-1 the obtained accuracy was 94% which by rank-4 reached 100% thereby corroborating the hypothesis put forward in Section 4.2.

A more detailed examination of errors at rank-1 level also revealed interesting findings. In particular, even when the coarse follow-up diagnosis category was not correctly identified, we found that nearly universally the prediction was highly meaningful. For example, to take some of the more frequent errors we observed, circulatory disorders were confused with blood disorders. In other instances we observed precisely what we argued from theory: that longitudinally patterned, chronic conditions were interlaced with episodic and sporadic ailments which are inherently not strongly predictable on the basis of a patient’s past diagnostic history. This corroboration of our hypothesis motivates the development of methods capable of separating such interlaced patterns and of incorporating temporal information in the predictive process, thereby providing more nuanced, temporal

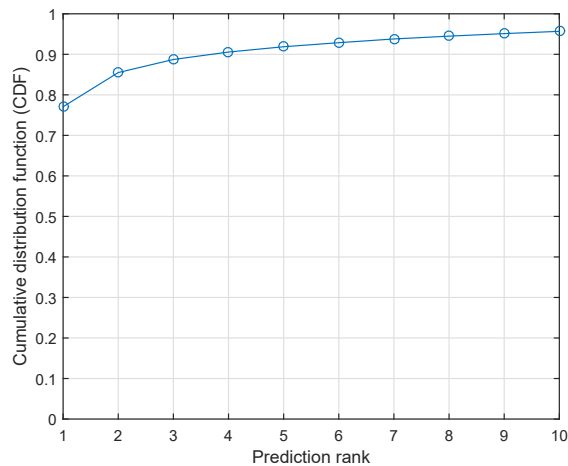


Figure 3: The cumulative match characteristic of the first follow-up diagnosis prediction obtained by our method. As argued in Section 4.2 we consider the performance across the rank range of approximately 3–5 to be most relevant in practice. From the plot it can be seen that the corresponding prediction accuracy is 89–92%.

rather than sequential prediction.

5 Summary and future work

In this paper we introduced a novel algorithm that uses machine learning on EHR collections for the discovery of longitudinal patterns in the diagnoses of diseases. The key technical novelty concerned a learning paradigm change which enables greater learning specificity. A series of experiments were presented to demonstrate the effectiveness of the proposed technique. Novel insights resulting from our experimental findings were also discussed and highlighted.

As regards possible future work directions, a number of possibilities were proposed by the author of the original history vector based approach that the present method was partly inspired by. While we agree with most of these in broad terms, our contributions, experiments, and results suggest what we believe to be more promising immediate alternatives. In particular while we agree with the authors of the original method that the presence of a particular episode of care is a predictive factor not much weaker than the exact number of episodes (which would require prohibitively large amount of training data to learn), we believe that history vector binarization is an overly harsh step for the reduction of the learning space. Following the spirit of the method introduced in the present paper we intend to explore the possibility of automatically detecting chronic types of episodes of care (such as dialysis, for example) and then using a binary representation for non-chronic, and a more graded representation for chronic conditions.

References

F. Abel, Q. Gao, G. J. Houben, and K. Tao. Analyzing user modeling on Twitter for personalized news recommenda-

tions. *In Proc. International Conference User Modeling, Adaptation and Personalization*, pages 1–12, 2011.

- A. Agarwal, B. Xie, I Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of Twitter data. *In Proc. Workshop on Language in Social Media*, pages 30–38, 2011.
- V. Andrei and O. Arandjelović. Identification of promising research directions using machine learning aided medical literature analysis. *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016.
- V. Andrei and O. Arandjelović. Temporal quasi-semantic visualization and exploration of large scientific publication corpora. *In Proc. International Joint Conference on Artificial Intelligence Workshop on Big Scholarly Data*, 2016.
- O. Arandjelović. Object matching using boundary descriptors. *In Proc. British Machine Vision Conference*, 2012. DOI: 10.5244/C.26.85.
- O. Arandjelović. Discovering hospital admission patterns using models learnt from electronic hospital records. *Bioinformatics*, 31(24):3970–3976, 2015.
- O. Arandjelović. Prediction of health outcomes using big (health) data. *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2543–2546, August 2015.
- O. Arandjelović. On the discovery of hospital admission patterns – a clarification. *Bioinformatics*, 2016. DOI: 10.1093/bioinformatics/btw049.
- E. Baucom, A. Sanjari, X. Liu, and M. Chen. Mirroring the real world in social media: Twitter, geolocation, and sentiment analysis. *In Proc. International Workshop on Mining Unstructured Big Data Using Natural Language Processing*, pages 61–68, 2013.
- A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli. Data-mining Twitter and the autism spectrum disorder: a pilot study. *In Proc. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pages 349–356, 2014.
- A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh. Hierarchical Dirichlet process for tracking complex topical structure evolution and its application to autism research literature. *In Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1:550–562, 2015.
- A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh. Overcoming data scarcity of Twitter: using tweets as bootstrap with application to autism-related topic content analysis. *In Proc. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pages 1354–1361, August 2015.
- A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli. Using Twitter to learn about the autism community. *Social Network Analysis and Mining*, 5(1):5–22, 2015.

- J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *In Proc. International Conference on Weblogs and Social Media*, pages 450–453, 2011.
- A. Bottle, P. Aylin, and A. Majeed. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *J R Soc Med*, 99(8):406–414, 2006.
- J. Butler and A. Kalogeropoulos. Hospital strategies to reduce heart failure readmissions. *J Am Coll Cardiol*, 60(7):615–617, 2012.
- B. Christensen and G. Ellingsen. Evaluating model-driven development for large-scale EHRs through the openEHR approach. *Int J Med Inform*, 89:43–54, 2016.
- K. Dharmarajan, A. F. Hsieh, Z. Lin, H. Bueno, J. S. Ross, I. Horwitz, J. A. Barreto-Filho, N. Kim, S. M. Bernheim, L. G. Suter, E. E. Drye, and H. M. Krumholz. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA*, 309(4):355–363, 2013.
- B. Friedman, H. J. Jiang, and A. Elixhauser. Costly hospital readmissions and complex chronic illness. *Inquiry*, 45(4):408–421, 2008–2009.
- B. G. Hammill, L. H. Curtis, and G. C. et al. Fonarow. Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circ Cardiovasc Qual Outcome*, 4(1):60–67, 2011.
- J. J. Holloway, S. V. Medendorp, and J. Bromberg. Risk factors for early readmission among veterans. *Health Serv Res*, 25(1):213–237, 1990.
- C. D. A. J. Holman, D. B. Preen, N. J. Baynham, J. C. Finn, and J. B. Semmens. A multipurpose comorbidity scoring system performed better than the charlson index. *J Clin Epidemiol*, 58(10):1006–1014, 2005.
- C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society, Series D*, 52(2):193–209, 2003.
- D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani. Risk prediction models for hospital readmission: a systematic review. *JAMA*, 306(15):1688–1698, 2011.
- J. Leegon, I. Jones, K. Lanaghan, and D. Aronsky. Predicting hospital admission for emergency department patients using a bayesian network. *AMIA Annu Symp Proc*, page 1022, 2005.
- J. Li and L. Guo. Hospital admission prediction using pre-hospital variables. *BIBM Conf Proc*, pages 283–286, 2009.
- A. M. Mudge, K. Kasper, A. Clair, H. Redfern, J. J. Bell, M. A. Barras, G. Dip, and N. A. Pachana. Recurrent readmissions in medical patients: a prospective study. *J Hosp Med*, 6(2):61–67, 2011.
- RGI-CGHR Collaborators. Report on the causes of death in India: 2001–2003. *Office of the Registrar General of India*, 2009.
- J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. *In Proc. IEEE International Conference on Computer Vision*, 2:1470–1477, 2003.
- R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman. Disease progression modeling using hidden Markov models. *In Proc. IEEE International Conference on Engineering in Medicine and Biology Society*, pages 2845–2848, 2012.
- I. Vasiljeva and O. Arandjelović. Prediction of future hospital admissions – what is the tradeoff between specificity and accuracy? *In Proc. International Conference on Bioinformatics and Computational Biology*, 2016.
- I. Vasiljeva and O. Arandjelović. Towards sophisticated learning from EHRs: increasing prediction specificity and accuracy using clinically meaningful risk criteria. *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016.
- World Health Organization. *International statistical classification of diseases and related health problems.*, volume 1. World Health Organization, 2004.
- L. Xu, D. Wen, X. Zhang, and J. Lei. Assessing and comparing the usability of Chinese EHRs used in two Peking University hospitals to EHRs used in the US: A method of RUA. *Int J Med Inform*, 89:32–42, 2016.