Physics of the Medical Record: Handling Time in Health Record Studies

George Hripcsak(✉)

Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, USA
hripcsak@columbia.edu

The rapid increase in adoption of electronic health records (EHRs) creates the possibility of tracking billions of patient visits per year and exploiting them for clinical research. The international observational research collaboration, Observational Health Data Sciences and Informatics (OHDSI), has counted 682 million patient records that have been converted to a common format known at the OMOP Common Data Model [1]. While this number includes duplicates and records that have not been made broadly available to researchers, its scale demonstrates that converting the world population to a common format is feasible.

Yet even with massive amounts of data available and even in a common format, a number of challenges exist. Data can be inaccurate, complex, and missing, and the health care process affects the measurement and recording of information to cause bias [2]. For example, tests from the middle of the night are more likely to be abnormal because patients will most likely be tested then only because they are very ill. Previous work demonstrated some of the human factors that can affect recording, such as not entering symptoms for deceased patients [3] with a consequent large effect on outcomes studies. Not surprisingly, clinical variables are correlated with variables related to health care processes like admission, and they each do so in ways that are distinctive but follow patterns such that related concepts have similar patterns [4]. The result of these challenges is that existing statistical and machine learning data analysis methods, if used naïvely on EHR data, will produced biased results.

Therefore, as both a source of information—e.g., to tease out causality—and as a source of bias to be avoided, time is pervasive in EHR studies. Time has been studied since the creation of the field of biomedical informatics and the use of artificial intelligence in medicine. The remainder of this paper covers topics in the collection and analysis of temporal data, with an emphasis on correcting bias associated with the temporal data. The term, "physics of the medical record," refers to the study of the record as an object of interest in itself (as opposed to the study of the patient) to better understand the health care processes that create the record and the resulting biases associated with the recording of data. It also refers to the general method of building and testing models, aggregating across units, and in some cases employing methods drawn from non-linear time series analysis.

The first challenge is collecting the temporal information. Structured clinical data are usually associated with one or more timestamps, and one of them is often identifiable as a primary time of interest [5]. Narrative clinical data, which today provide a deeper view of the patient in terms of symptoms and clinicians' motivation, require natural language

J.H. Holmes et al. (Eds.): AIME 2015, LNAI 9105, pp. 3–6, 2015.
DOI: 10.1007/978-3-319-19551-3_1

processing, and assessing the time of events abstracted from narrative reports is complex. For example, the time that an event really occurred differs from the time that a patient reports it to a clinician or the time that the clinician writes a note about it. Progress has been made in temporal processing [6] with surprisingly good performance both on predefined tasks [7] and even on general tasks for which the system was not optimized [8,9]. More research is required, however, including understanding how temporal concepts are used. It was previously found, for example, that the uncertainty of temporal declarations is predictable and can be modeled with a regression equation [10] to supply information to downstream analysis.

Temporal data may be analyzed in many ways. The literature is filled with a diverse set of techniques to include time in analyses. This includes machine-learning approaches during phenotyping [11,12], pattern discovery [13,14,15], temporal abstraction over intervals [16,17], and dynamic Bayesian networks [18,19]. Several of these directly handle the irregularity of time [12,13,18,19].

In keeping with the "physics of the medical record" theme, it may be useful to step back and better understand the temporal properties of the EHR. As noted above, data are collected irregularly and in a biased fashion such that patients are sampled more often when they are more ill. Thus they are not at all sampled at random. Furthermore, physiology is by its nature non-stationary—we hope so because our goal is ultimately to change a person's state from ill to healthy. This non-stationarity can affect algorithms. For example, predictive algorithms generally assume that the distributions of data and model parameters remain constant over the prediction period. It was previously demonstrated that the predictability of clinical data is better correlated with sampling frequency than with actual time [20]. Further analysis showed that sequential measurements have roughly constant variability over a broad range of time scales [21]. That is, it appears that clinicians sample patients at a rate commensurate with the change in variability, sampling more often when patients are more ill and generally more variable. It was recently shown that clinicians sometimes over- and sometimes under-correct for changes in variability [22]. By parameterizing a problem not by actual time, but by units of sampling—i.e., arbitrarily define the time between sequential measurements as one—one can achieve greater stationarity and at least in one example improve predictive power [21].

While an EHR may have many health records, the data available for any given patient is often limited, especially once one selects variables relevant to some specific task. The result is a large number of short time series. Combining the information from these disparate patients is challenging because there may be too few measurements in each patient to draw reliable conclusions about each one's time series. Instead, the information implicit in the short time series must be aggregated. In a study of predictability [23], as quantified by the mutual information between different time points within a patient, which was referred to as time-delayed mutual information (TDMI), it was shown that such aggregation could produce interpretable results. A method was developed to decide when such aggregation is warranted [24], how to assess baseline mutual information and therefore excess information, and bias. [25].

Similar work outside of biomedicine [26] demonstrates the generalizability of such research and the benefit of looking broadly outside of biomedicine for new methods.

Given the ability to aggregate time series, one can exploit health record data to uncover correlations among variables, carrying out tasks like pharmacovigilance. A relatively simple algorithm using only lagged linear correlation, linear temporal interpolation, and within-patient normalization, produced informative results about temporal processes based on definitional, physiologic, and intentional associations [27] despite being applied blindly to all health record data regardless of source or clinical context.

With advances in understanding and applying time series methods to health record data, we will be better able to exploit health record data. For example, one study revealed whether seizures in the setting of intracerebral bleed are merely symptoms or whether they cause further morbidity [28]. Further work is needed studying the EHR, studying the health care processes that underlie it, and developing new methods to analyze it.

References

1. Hripcsak, G., Duke, J.D., Shah, N.H., Reich, C.G., Huser, V., Schuemie, M.J., Suchard, M.A., Park, R.W., Wong, I.C.K., Rijnbeek, P.R., van der Lei, J., Pratt, N., Norén, G.N., Lim, Y.C., Stang, P.E., Madigan, D., Ryan, P.B.: Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. In: MEDINFO 2015, São Paulo, Brazil, August 19-23 (2015)
2. Hripcsak, G., Albers, D.J.: Next-generation phenotyping of electronic health records. J. Am. Med. Inform. Assoc. 20, 117–121 (2013), doi:10.1136/amiajnl-2012-001145.
3. Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., Melton, G.B.: Bias associated with mining electronic health records. J. Biomed. Discov. Collab. 6, 48–52 (2011), PMC3149555
4. Hripcsak, G., Albers, D.J.: Correlating electronic health record concepts with health care process events. J. Am. Med. Inform. Assoc. 20(e2), e311-e318 (2013), doi:10.1136/amiajnl-2013-001922.
5. Hripcsak, G., Ludemann, P., Pryor, T.A., Wigertz, O.B., Clayton, P.D.: Rationale for the Arden Syntax. Comput. Biomed. Res. 27, 291–324 (1994)
6. Zhou, L., Hripcsak, G.: Temporal reasoning with medical data - A review with emphasis on medical natural language processing. J. Biomed. Inform. 40, 183–202 (2007)
7. Uzuner, Ö., Stubbs, A., Sun, W.: Chronology of your health events: Approaches to extracting temporal relations from medical narratives. J. Biomed. Inform. 46, S1–S4 (2013)
8. Zhou, L., Parsons, S., Hripcsak, G.: The evaluation of a temporal reasoning system in processing clinical discharge summaries. J. Am. Med. Inform. Assoc. 15, 99–106 (2008), PMC2274869
9. Sun, W., Rumshisky, A., Uzuner, O.: Temporal reasoning over clinical text: the state of the art. J. Am. Med. Inform. Assoc. 20, 814–819 (2013)
10. Hripcsak, G., Elhadad, N., Chen, C., Zhou, L., Morrison, F.P.: Using empirical semantic correlation to interpret temporal assertions in clinical texts. J. Am. Med. Inform. Assoc. 16, 220–227 (2009), PMC2649319
11. Lasko, T.A., Denny, J.C., Levy, M.: Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One 8, e66341 (2013)

12. Liu, Z., Hauskrecht, M.: Sparse linear dynamical system with its application in multivariate clinical time series. In: NIPS 2013 Workshop on Machine Learning for Clinical Data Analysis and Healthcare (December 2013)
13. Wang, F., Lee, N., Hu, J., Sun, J., Ebadollahi, S.: Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In: KDD 2012, Beijing, China, August 12-16, pp. 453–461 (2012)
14. Batal, I., Valizadegan, H., Cooper, G.F., Hauskrecht, M.: A pattern mining approach for classifying multivariate temporal data. In: Proceedings IEEE Int. Conf. Bioinformatics Biomed., pp. 358–365 (2011)
15. Noren, G.N., Hopstadius, J., Bate, A., Star, K., Edwards, I.R.: Temporal pattern discovery in longitudinal electronic patient records. Data Min. Knowl. Discov. 20, 361–387 (2010)
16. Shahar, Y.: A framework for knowledge-based temporal abstraction. Artificial Intelligence 90(1-2), 79–133 (1997)
17. Moskovitch, R., Shahar, Y.: Medical temporal-knowledge discovery via temporal abstraction. In: AMIA Annu. Symp. Proc., pp. 452–456 (2009)
18. Sebastiani, P., Mandl, K.D., Szolovits, P., Kohane, I.S., Ramoni, M.F.: A Bayesian dynamic model for influenza surveillance. Stat. Med. 25(11), 1803–1816 (2006)
19. Ramati, M., Shahar, Y.: Irregular-time Bayesian networks. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010), Catalina Island, CA, USA (2010)
20. Albers, D.J., Hripcsak, G.: An information-theoretic approach to the phenome (abstract). In: AMIA Summit on Translational Bioinformatics, March 15-17, San Francisco, CA (2009)
21. Hripcsak, G., Albers, D.J., Perotte, A.: Parameterizing time in electronic health record studies. J. Am. Med. Inform. Assoc. (February 26, 2015), pii: ocu051, doi: 10.1093/jamia/ocu051.
22. Lasko, T.A.: Nonstationary Gaussian process regression for evaluating repeated clinical laboratory tests. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, January 25-30 (2015)
23. Albers, D.J., Hripcsak, G.: A statistical dynamics approach to the study of human health data: resolving population scale diurnal variation in laboratory data. Physics Letters A 374, 1159–1164 (2010), PMC2882798
24. Albers, D.J., Hripcsak, G.: Using time-delayed mutual information to discover and interpret temporal correlation structure in complex populations. Chaos 22, 013111 (2012), doi:10.1063/1.3675621
25. Albers, D.J., Hripcsak, G.: Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series. Chaos, Solitons & Fractals 45, 853–860 (2012), PMC3332129
26. Komalapriya, C., Thiel, M., Ramano, M.C., Marwan, N., Schwarz, U., Kurths, J.: Reconstruction of a system's dynamics from short trajectories. Phys. Rev. E 78, 066217 (2008)
27. Hripcsak, G., Albers, D.J., Perotte, A.: Exploiting time in electronic health record correlations. J. Am. Med. Inform. Assoc. 18(suppl. 1), i109–i115 (2011)
28. Claassen, J., Albers, D., Schmidt, J.M., De Marchis, G.M., Pugin, D., Falo, C.M., Mayer, S.A., Cremers, S., Agarwal, S., Elkind, M.S.V., Connolly, E.S., Dukic, V., Hripcsak, G., Badjatia, N.: Nonconvulsive seizures in subarachnoid hemorrhage link inflammation and outcome. Annals of Neurology (in press)