Name:	Matthew Levine
Affiliation:	Department of Biomedical Informatics, Columbia University
Email:	mel2193@cumc.columbia.edu
Presentation type:	Poster

Comparing lagged linear methods for uncovering associations in EHR data

Matthew E. Levine, BA[1], David J. Albers, Ph.D.[1], George Hripcsak, M.D., M.S.[1]
[1]Department of Biomedical Informatics, Columbia University, New York, New York, USA

Abstract

Time series analysis methods have been shown to reveal clinical and biological associations in data collected in the electronic health record. We wish to develop reliable high-throughput methods for identifying adverse drug effects that are easy to implement and produce readily interpretable results. To move toward this goal, we used univariate and multivariate lagged regression models to investigate associations between twenty pairs of drug orders and laboratory measurements. Multivariate lagged regression models exhibited higher sensitivity and specificity than univariate lagged regression in the 20 examples, and incorporating autoregressive terms for labs and drugs produced more robust signals in cases of known associations. Moreover, including inpatient admission terms in the model attenuated the signals for some cases of unlikely associations, suggesting that multivariate lagged regression models' explicit handling of context-based variables provides a simple way to probe for health-care processes that confound analyses of EHR data.

Introduction

With the increasing collection and storage of patient electronic health data around the world comes a proportionally growing impetus to use that information to improve clinical care. We hope to move towards reliable high-throughput methods for determining adverse drug effects that can be applied to large clinical data repositories, like that collected by Observational Health Data Sciences and Informatics (OHDSI), which contains over 600 million patient records [1]. Many research inquiries can be satisfied with simple determinations of whether a patient ever had a particular condition, and it is often sufficient to consider events that occur over relevant time windows with respect to a condition of interest [2]. However, it can be useful to consider methods with the potential to reveal fine temporal structure in EHR data, and recent advances in such methods have been applied to machine-learning approaches during phenotyping [3,4], pattern discovery [5–7], temporal abstraction over intervals [8], and dynamic Bayesian networks [9]. Many of these approaches to time-series analysis rely on assumptions of stationarity (roughly, having consistent mean and variance through a time window of interest) that are frequently broken by clinical data. This issue is compounded by the simple fact that patients are sampled with greater frequency when they are ill [10].

Our past work has revealed informative results about temporal processes in the EHR by applying lagged linear correlation to time series constructed using linear temporal interpolation and intra-patient normalization of clinical signout note and laboratory test data [11]. Similarly, time-delayed mutual information reveal lagged linear structure as well as nonlinear dynamical processes related to physiology [12,13] despite EHR-data complexities and homo- or heterogeneity among patient populations [14–17]. Our most recent efforts to characterize temporal processes in the EHR are motivated by our previous findings that 1) temporal clinical and physiologic processes can be described through lagged linear correlation of concepts extracted from signout notes and laboratory values [11], 2) time series data, under some clinical circumstances, are better parameterized by their raw sequence than their clock measurements [17], and 3) health-care process events such as inpatient admission are systematically correlated with concepts and laboratory values [18].

Methods

In this study, we used multivariate distributed lag models to incorporate additional context-related variables in lagged linear analysis of temporal processes to better characterize both intended and unintended physiologic effects of drugs. In order to broaden the applicability of the method, we designed a time series preparation methodology that can use drug-order records as inputs, which, unlike physician notes, are readily available in data collected by OHDSI. As part of optimizing time series construction methods, we investigated the effects of two pre-processing steps: intra-patient normalization of laboratory tests and different data preparation strategies (e.g. regressing on differences between interpolated values of the time series). In order to evaluate these methods, we applied them to twenty pairings of drugs and laboratory measurements.

We used the 27-year-old clinical data warehouse at NewYork-Presbyterian Hospital, which contains electronic health records for over 3 million patients, to examine pairwise relationships between drug order records and laboratory measurements. The subset of data used is also available in OMOP CDM v4. We considered five drugs—simvastatin, amphotericin B, spironolactone, and warfarin—and four laboratory tests (total creatine kinase (CK), creatinine, potassium, and hemoglobin), and a patient cohort was identified for each of the 20 drug-lab pairs in the experiment. We identified eight drug-lab pairs for which existing clinical evidence suggests significant physiologic associations; we did not find conclusive evidence for associations between the remaining 12 drug-lab pairs.

Because our goal is to minimize bias and confounding, we employed two techniques to minimize bias. We used a particular form of lagged regression, known as Granger causality [19], to assess the effect of one variable (drug) over another (laboratory measurement) beyond that accounted for by the target variable's autocorrelation. We used an extension of Granger causality, vector autoregression [20], to also account for a third variable (inpatient admission) as an example of a health care process confounder. We performed 100 iterations of a bootstrap by sampling patients with replacement in order to obtain confidence intervals for estimated coefficients. In our evaluation, we focus on the estimates of lagged drug coefficients, and evaluate the effect of additional variables not by examining their coefficients directly, but rather by evaluating how their presence affected the drug coefficients.

After performing a univariate lagged linear regression, we considered a simple multivariate lagged regression that only incorporates lagged drug values, and jointly estimates all lagged drug coefficients, which we refer to as the "multivariate lagged drug model". We then evaluated how adding lagged terms to represent previous laboratory values affects drug coefficients by fitting a multivariate autoregressive drug and lab model in the form of Granger causality [19]. We also introduce an additional context variable to represent the inpatient admission timeline, and fit a further augmented multivariate autoregressive drug, lab, and context model in vector autoregression form [20]. Each lagged variable had 30 lagged terms, for which each lag represented one unit in sequence time. So, the "autoregressive drug, lab, and context" model uses the last 30 interpolated laboratory values, the last 30 interpolated drug values, and the last 30 interpolated admission values from the constructed time series to predict a present measurement. This alignment of previous data is performed for each laboratory measurement, and is aggregated within each patient, then across patients, creating a matrix with 91 columns (90 explanatory values and 1 predicted value) and a length equivalent to the number of qualifying laboratory measurements in the cohort.

Results

By developing a method for constructing time series of continuous and categorical variables, we were able to compare univariate and multivariate lagged regression models that incorporate lab measurements, drug orders, and inpatient admissions. Both univariate and multivariate lagged methods performed best, overall, with intra-patient normalized laboratory values, and multivariate methods performed best when "differences" were used during pre-processing stages. All multivariate methods identified the same six out of eight expected physiologic effects documented in clinical literature, whereas the univariate approach identified five. Adding variables that describe patient context (lagged lab measurements and lagged admission events) increased the number and magnitude of significant drug coefficients in the expected cases and improved discrimination against unlikely associations. We found that adding context-based variables to autoregressive models allowed for explicit handling of confounding variables and provided a simple way to evaluate the temporal effects of ordered drugs on physiology.

Conclusion

By comparing univariate and multivariate lagged regression models, we established methods for timeline construction that yielded consistent results across model implementations. We found that drug effects were best characterized, as compared to clinical literature, by multivariate lagged models that incorporate drug orders, laboratory measurements, and inpatient admission events. These results suggest that simple autoregressive models of commonly available EHR data can be used to detect real physiologic drug effects in the presence of confounding health-care processes.

Acknowledgment

This work was funded by National Library of Medicine grant R01 LM006910.

References

[1] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, Stud Health Technol Inform. 216 (2015) 574–578.

[2] C.A. McCarty, R.L. Chisholm, C.G. Chute, I.J. Kullo, G.P. Jarvik, E.B. Larson, R. Li, D.R. Masys, M.D. Ritchie, D.M. Roden, others, The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Medical Genomics. 4 (2011) 13.

[3] T.A. Lasko, J.C. Denny, M.A. Levy, Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data, PLoS ONE. 8 (2013) e66341. doi:10.1371/journal.pone.0066341.

[4] Z. Liu, M. Hauskrecht, Sparse linear dynamical system with its application in multivariate clinical time series, arXiv Preprint arXiv:1311.7071. (2013). http://arxiv.org/abs/1311.7071 (accessed March 6, 2016).

[5] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012: pp. 453–461. http://dl.acm.org/citation.cfm?id=2339605 (accessed March 6, 2016).

[6] I. Batal, H. Valizadegan, G.F. Cooper, M. Hauskrecht, A Pattern Mining Approach for Classifying Multivariate Temporal Data, in: IEEE, 2011: pp. 358–365. doi:10.1109/BIBM.2011.39.

[7] G.N. Norén, J. Hopstadius, A. Bate, K. Star, I.R. Edwards, Temporal pattern discovery in longitudinal electronic patient records, Data Mining and Knowledge Discovery. 20 (2010) 361–387. doi:10.1007/s10618-009-0152-3.

[8] R. Moskovitch, Y. Shahar, Medical temporal-knowledge discovery via temporal abstraction., in: AMIA, 2009. http://medinfo.ise.bgu.ac.il/medLab/MembersHomePages/RobPapers/Moskovitch.MedicalKarmaLego.AMIA09.pdf (accessed March 6, 2016).

[9] M. Ramati, Y. Shahar, Irregular-time Bayesian networks, arXiv Preprint arXiv:1203.3510. (2012). http://arxiv.org/abs/1203.3510 (accessed March 6, 2016).

[10] A. Rusanov, N.G. Weiskopf, S. Wang, C. Weng, Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research, BMC Medical Informatics and Decision Making. 14 (2014) 1.

[11] G. Hripcsak, D.J. Albers, A. Perotte, Exploiting time in electronic health record correlations, Journal of the American Medical Informatics Association. 18 (2011) i109–i115. doi:10.1136/amiajnl-2011-000463.

[12] D.J. Albers, G. Hripcsak, M. Schmidt, Population Physiology: Leveraging Electronic Health Record Data to Understand Human Endocrine Dynamics, PLoS ONE. 7 (2012) e48058. doi:10.1371/journal.pone.0048058.

[13] D.J. Albers, N. Elhadad, E. Tabak, A. Perotte, G. Hripcsak, Dynamical Phenotyping: Using Temporal Analysis of Clinically Collected Physiologic Data to Stratify Populations, PLOS ONE. 9 (2014) e96443. doi:10.1371/journal.pone.0096443.

[14] D.J. Albers, G. Hripcsak, Using time-delayed mutual information to discover and interpret temporal correlation structure in complex populations, Chaos: An Interdisciplinary Journal of Nonlinear Science. 22 (2012) 013111.

[15] D.J. Albers, G. Hripcsak, Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series, Chaos, Solitons & Fractals. 45 (2012) 853–860. doi:10.1016/j.chaos.2012.03.003.

[16] D.J. Albers, G. Hripcsak, A statistical dynamics approach to the study of human health data: Resolving population scale diurnal variation in laboratory data, Physics Letters A. 374 (2010) 1159–1164. doi:10.1016/j.physleta.2009.12.067.

[17] G. Hripcsak, D.J. Albers, A. Perotte, Parameterizing time in electronic health record studies, Journal of the American Medical Informatics Association. 22 (2015) 794–804. doi:10.1093/jamia/ocu051.

[18] G. Hripcsak, D.J. Albers, Correlating electronic health record concepts with healthcare process events, Journal of the American Medical Informatics Association. 20 (2013) e311–e318. doi:10.1136/amiajnl-2013-001922.

[19] C.W.J. Granger, Investigating Causal Relations by Econometric Models and Cross-spectral Methods, Econometrica. 37 (1969) 424–438. doi:10.2307/1912791.

[20] J. Durbin, S.J. Koopman, Time series analysis by state space methods, 2nd ed, Oxford University Press, Oxford, 2012.