

Comparing lagged linear correlation, lagged regression, Granger causality, and vector autoregression for uncovering associations in EHR data

Matthew E. Levine, BA¹, David J. Albers, Ph.D.¹, George Hripcsak, M.D., M.S.¹

¹Department 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 among the 20 example pairings. Moreover, including inpatient admission terms in the model attenuated the signals for some cases of unlikely associations, demonstrating how multivariate lagged regression models' explicit handling of context-based variables can provide 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. These improvements can range from workflow and operations optimization to pharmacovigilance studies, but the central feature for effectively exploiting the electronic health record (EHR) is our ability to learn from the data collected. 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 is to be expected, even desired, since the primary goal of medicine is to drive patients from problematic to healthy states. This issue is compounded by the simple fact that patients are sampled with greater frequency when they are ill [10]. In fact, it appears that clinicians sample patients at rates proportional to their health variability, effectively inducing stationarity by indexing the time series not by clock-time, but rather by mere measurement sequence with single units of time imposed between each measurement [11].

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 [12]. These results indicated temporal processes that were definitional (e.g. low potassium levels associated with hypokalemia), physiologic (e.g. a potassium-sparing diuretic preceded increases in potassium levels), or intentional (e.g. a potassium-sparing diuretic was ordered in patients with low potassium levels), and used clock-time as the lagged time variable. Similarly, time-delayed mutual information reveal lagged linear structure as well as nonlinear dynamical processes related to physiology [13,14] despite EHR-data complexities and homo- or heterogeneity among patient populations [11,15–17].

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 [12], 2) time series data, under some clinical circumstances, are better parameterized by their raw sequence than their clock measurements [11], and 3) health-care process events such as inpatient admission are systematically correlated with concepts and laboratory values [18].

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 are readily available in more contexts than physician notes. In order to evaluate these methods, we applied them to twenty pairings of drugs and laboratory measurements. As part of optimizing time series construction methods, we also investigated the effects of two pre-processing steps: intra-patient normalization of laboratory tests and different data preparation strategies.

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.

Methods

Experimental design

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. 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 clinical evidence suggested significant physiologic associations (shown in Table 1); we did not find conclusive evidence for associations between the remaining 12 drug-lab pairs. Patients were included in a drug-lab cohort if they met the following criteria: 1) at least 2 of the laboratory measurements of interest on record, 2) at least 1 order for the drug of interest, and 3) more than 30 combined data points between laboratory measurements of interest and total drug orders (any drug). No attempts were made to remove or correct outliers.

Building a time series from clinical data

Laboratory measurements, drug orders, and inpatient admission events for each patient in each cohort were extracted from the clinical data warehouse. A piecewise-defined linear drug-lab timeline was constructed for each patient as described by Hripcsak *et al.* using linear temporal interpolation (see Figure 1 in Hripcsak *et al.* for an example) [12]. Laboratory values were continuous, and orders for the drug of interest were represented as 1 (present), whereas orders for other drugs were represented as 0 (absent). Although orders for other drugs do not necessarily indicate cessation of a drug of interest, such orders were treated as evidence of absence to avoid incorporating external domain knowledge about drug administration that might produce artifact associations. Inpatient admission timelines were defined with a 1 at the time of admission, and zeros at 24hrs before and after admission, effectively creating spikes at times of admission. Smoothness and differentiability of the admission spike are unimportant when using discretized convolutional approaches, and there are many ways of constructing the spike such that it has mass to contribute during the convolution. For every time point where there was a concept (lab, drug, or inpatient admission), the values of each other variable at that time point were interpolated as the clock-time weighted mean of the preceding and succeeding value of each respective variable (or as the closest measurement if there was no value on one side). Clock-time weighting is computed by weighting the 2 bordering values by their temporal distances from the time-point at which we interpolate. Thus, all concepts, whether from categorical or real-valued sources, took on rational values that were paired at each time point. Because each time-point typically has only 1 reported event, each time-tuple is comprised of 1 true data point and 2 interpolated values.

Pre-processing of time series data

Two types of pre-processing of interpolated time series were designed and evaluated. First, each patient's time series of laboratory values were normalized to have 0 mean and 1 variance by subtracting the mean and dividing by standard deviation [12]. This operation removed inter-patient effects. Second, we replaced each interpolated value in the time series with its difference from the immediately preceding interpolated value, such that time series values represented changes in values. This was an important means of reducing dependence between lagged variables in our novel application of multivariate lagged regression models to interpolated clinical data. As such, we effectively considered 4 time series construction methods—1) no pre-processing, 2) intra-patient normalization, 3) differences, and 4) normalization and differences.

Sequence time

Although clock-time was used for weighted interpolations and pre-processing steps, it was discarded in favor of raw sequence time for subsequent lagged linear analyses due to recent examples that demonstrated greater information and greater stationarity in clinical time series data that are parameterized by their sequence [11]. Real-time may prove to be a more sensible choice in other circumstances, especially when data is stationary. All time intervals between interpolated, pre-processed values were set to unit 1 length, effectively converting from clock-time to sequence time.

Univariate lagged linear regression (ULLR)

We compute lagged linear regression coefficients, β_τ , for the following distributed lag model [20], where y_t represents a laboratory value at sequence time t and x represents the interpolated drug value at time $t-\tau$. This model performs the same computations as our previous lagged linear correlation experiment [12], but supplies a different statistic, namely the lagged drug coefficient β_τ .

$$y_t = c_\tau + \beta_\tau x_{t-\tau} + \epsilon_\tau \quad (1)$$

Multivariate lagged regression (MLLR)

We aim to leverage the relationship between the context of each data point and the variables they predict by including context-dependent factors in multivariate autoregressive models. In general, a multivariate distributed lag model for L lags and N variables (for which the i^{th} variable is denoted u^i), can be used to define the lagged coefficient for each variable u^i at each lag τ (denoted as $\beta_\tau^{u^i}$), and is written as [20]

$$y_t = c + \sum_{i=1}^N \sum_{\tau=1}^L \beta_{u^i, \tau} u_{t-\tau}^i + \epsilon \quad (2)$$

,where $\beta_{u^i, \tau}$ is the coefficient for lag τ of the variable u^i . Because the parameters of these models are estimated jointly, adding explanatory variables that are related to both the predicted variables and variables of interest can change the values of coefficients of interest. Concretely, we considered a simple multivariate lagged regression that only incorporates lagged drug values, and jointly estimates all lagged drug coefficients β_τ , with $L=30$, according to the following model, which we refer to as the “multivariate lagged drug model”

$$y_t = c + \sum_{\tau=1}^L \beta_\tau x_{t-\tau} + \epsilon \quad (3)$$

We then evaluated how adding lagged terms to represent previous laboratory values affects drug coefficients by fitting the following “autoregressive drug and lab” model with $L=30$; this is in the form of Granger causality [19].

$$y_t = c + \sum_{\tau=1}^L \beta_{y, \tau} y_{t-\tau} + \sum_{\tau=1}^L \beta_{x, \tau} x_{t-\tau} + \epsilon \quad (4)$$

We also introduce an additional context variable, z , to represent the inpatient admission timeline, and fit a further augmented “autoregressive drug, lab, and context” model with $L=30$; this is in the form of vector autoregression [20].

$$y_t = c + \sum_{\tau=1}^L \beta_{y, \tau} y_{t-\tau} + \sum_{\tau=1}^L \beta_{x, \tau} x_{t-\tau} + \sum_{\tau=1}^L \beta_{z, \tau} z_{t-\tau} + \epsilon \quad (5)$$

Intuitively, this 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. We did not perform any feature selection procedures, such as Bayesian information criterion, as this was out of the scope of our case study for

method comparison—employing such selection criteria is a key step in determining true Granger causality, which we did not attempt.

Bootstrap

After cohort identification, timeline construction, and pre-processing, we aggregated pairs of lagged interpolated values across patients to construct a sparse model matrix for each drug-lab pair that conformed to the specified dimensionality of each model. Coefficients for each model were estimated by performing sparse linear least squares regression with Cholesky factorization from the MatrixModels package in R [21]. One hundred iterations of a bootstrap were performed for each matrix by sampling patients with replacement [22] in order to obtain empirical estimates of statistical significance. Coefficient estimates were labeled as statistically significant if zero was not included in their 95% Confidence Interval (CI) as computed by the bootstrap, which was defined as

$$\left(\mathbb{E}[\beta_\tau] - 1.96 \sqrt{\text{var}(\beta_\tau)}, \mathbb{E}[\beta_\tau] + 1.96 \sqrt{\text{var}(\beta_\tau)} \right)$$

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.

Results

Intra-patient normalization in univariate lagged linear regression

Univariate lagged linear regression (ULLR) with intra-patient normalization was performed for each of the 20 drug-lab pairs of interest, eight of which we hypothesized, based on clinical literature, to have a significant directional (i.e. increasing or decreasing) effect. The number of significantly positive and negative lagged drug coefficients are compared with results that could be expected from the literature in Table 1. Analysis of normalized data with ULLR detected 4 out of 8 expected signals correctly with appropriate directionality (3 positive relationships and 1 negative), and reported all other cases to have statistically significant relationships. While this demonstrates real statistical correlations between variables, it does not necessarily implicate a physiologic association. The many biases in EHR data can be misleading when drawing statistical conclusions, so it is important to find ways of systematically focusing the analysis to reveal only the effects of interest—in this case, ones rooted in physiology.

Univariate LLR of drugs predicting labs

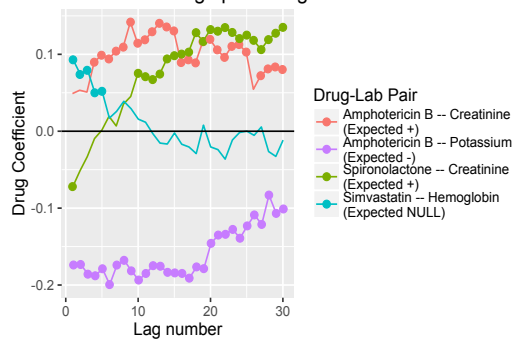


Figure 1. The lagged drug coefficients for four drug-lab pairs are plotted, and statistically significant (95% CI does not include 0) coefficients are denoted with points on the line plots. Lag number 1 indicates the interpolated drug values immediately preceding each laboratory measurement, and lag number 30 indicates the interpolated drug value 30 points in sequence time preceding each laboratory measurement. Significance on the left of the plot indicates rapid effects, while significance on the right indicates slower effects. Amphotericin B is shown to increase creatinine and decrease Potassium, as expected [23,24]. Simvastatin is shown to increase Hemoglobin levels early on, and spironolactone is shown to decrease creatinine initially, then raise it. These latter effects are likely explained, at least in part, by health care processes.

likely to be sick (possibly with high creatinine) and subsequently improve due to treatment. Creatinine elevation due to the drug, then, would be on a longer time scale than creatinine-lowering treatments. Table 1 shows that the multivariate models remove this effect, likely by jointly considering previous orders of the drug.

In Figure 1, we show that univariate LLR analysis of normalized data reveals clinically characterized trends, such as amphotericin B's tendency to decrease potassium levels and increase creatinine. It shows an overall trend linking spironolactone to increases in creatinine levels (a known phenomenon), but also finds a statistically significant negative relationship at lag of 1 in sequence time. Figure 1 also indicates a significant association between simvastatin and increases in hemoglobin levels, for which we do not have a particular biological interpretation. This result is characteristic of the significant signals detected by ULLR in the 12 drug-lab pairs for which we did not expect a physiologic association. These results suggest that univariate LLR analysis, like other analytic approaches to clinical data, is vulnerable to health-care process effects in EHR data. For example, the short-term negative relationship of spironolactone and could be attributed to a treatment pattern in which patients first prescribed the drug are

Adding autoregressive terms: multivariate lagged linear regression

We adopted a multivariate LLR model in order to address the confounding effects of health care process when tasked with detecting true physiologic effects of drugs. We first considered a model that estimates all lagged drug coefficients jointly (equation 3, “multivariate lagged drug model”) with the intent of incorporating drug timeline history into each prediction. We also evaluated an augmented version of this model that adds autoregressive terms of previous laboratory values as well as drug orders (equation 4, “autoregressive drug and lab model”)—this formulation is very similar to the autoregressive model used in Granger causality, although we do not perform model selection to choose the number of lags, nor do we address potential unit root issues as is typical in Granger causal analysis [19]. We used intra-patient normalization and applied the differences pre-processing step described earlier (replacing values of the time series with their difference from the previous value) in all multivariate LLR analyses to reduce dependence between interpolated values (independence of lagged variables is an important requirement in multiple regression).

The multivariate lagged drug model showed statistically significant relationships with correct directionality between 6 of the 8 drug-lab pairs with known biological activity and rejected 3 of the 12 uncharacterized pairings. The autoregressive model of drug and lab histories detected the same 6 of the 8 known pairs, and also rejected 3 uncharacterized pairings. Moreover, in cases of true associations, drug and lab autoregression revealed, on average, more significant coefficients with greater magnitude. This finding suggests that the multivariate drug and lab model will have better sensitivity than the lagged drug model without sacrificing specificity.

Figure 2 shows how univariate LLR, multivariate LLR with drugs, and multivariate LLR with drugs and labs describe two cases: 1) ibuprofen predicting creatinine levels, and 2) simvastatin predicting changes in hemoglobin. It is known that ibuprofen can cause acute renal failure, for which high creatinine is a common symptom, and indeed we see that multivariate LLR (and not ULLR) predicts the correct directionality of the effect, with larger and more significant coefficients produced when using autoregressive terms of lab and drug values. An effect of simvastatin on hemoglobin levels, however, was not supported by evidence from our literature search, suggesting that apparent effects are attributable to non-biological phenomena. Figure 2b shows that the MLLR drug model predicts a significant negative effect, whereas the lab and drug model indicates no significant relationships. This implies that previous hemoglobin measurements helped to explain future drops in hemoglobin (i.e. this model may account for the number of anemic patients on simvastatin). The univariate model, however, showed a positive effect during early

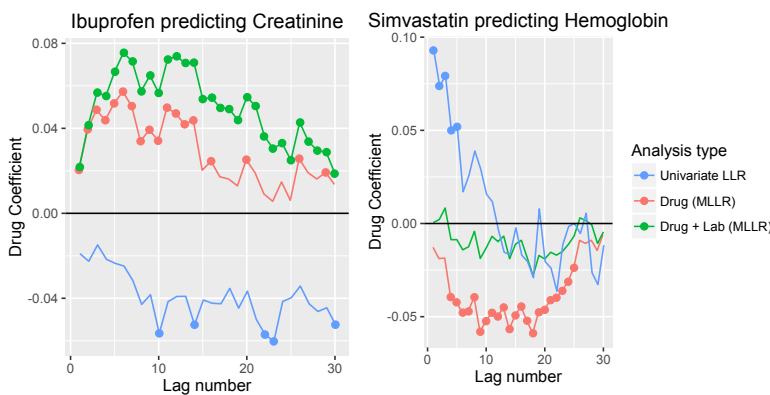


Figure 2. a) Ibuprofen is shown to predict elevated levels of creatinine when using multivariate lagged linear regression (MLLR) models, with a more robust signal coming from the lab and drug model. Univariate LLR (ULLR) predicted a negative relationship, with only 5 significant coefficients. b) Simvastatin is suggested to have a short-term positive effect on hemoglobin by ULLR, and a sustained negative effect by drug MLLR. MLLR with lab and drug terms, however, reports no significant drug coefficients.

time points, indicating something more likely related to health care processes that were accounted for by the multivariate models through joint parameter estimation. The performance of multivariate LLR methods was dependent on both intra-patient normalization and taking differences of the time series data sets. Omitting both pre-processing steps resulted in 0 correctly identified signals for the drug-only model, and 1 correctly

identified signal for the drug and lab model. Using normalization without differences allowed the drug and lab model to detect 1 additional signal (2 in total), while using differences without normalization restored performance closer to levels seen with both pre-processing steps. The drug-only model detected the same 6 of 8 hypothesized drug-lab pairs, and had 10 apparent false positives. The lab and drug model significantly underperformed without both differences and normalization, detecting half the number of expected associations when used without normalization.

Furthermore, the combination of normalization and differences yielded more robust signals. Figure 3 illustrates that the combination of both pre-processing steps produces the most robust signal for predicting creatinine elevation by

amphotericin B, as measured by the number and magnitude of correctly oriented significant coefficients. In this, example both differences and normalization are required to identify a positive signal with confidence.

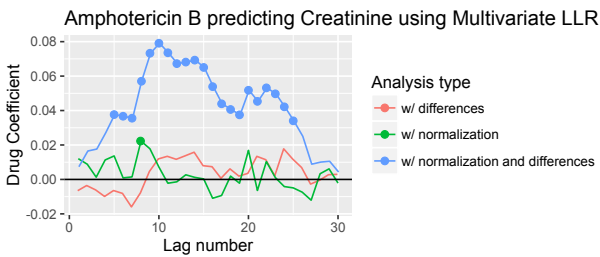


Figure 3. The resulting drug coefficients are plotted for amphotericin B predicting creatinine using the autoregressive drug and lab model with different pre-processing steps. In this case both normalization and differences were necessary to produce the expected positive association [37].

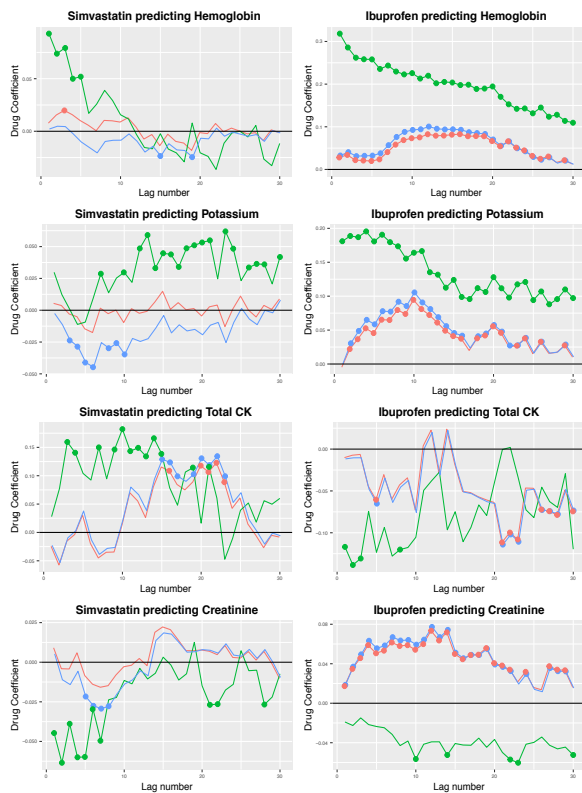


Figure 4. Drug coefficients estimated from the univariate lagged linear regression (ULLR) model, the autoregressive drug and lab model (drug + lab MLLR), and the autoregressive context model (drug + lab + admit MLLR) (see legend of Figure 3) were plotted for all four investigated labs for **a)** simvastatin and **b)** ibuprofen. The context model accentuated the expected effect on CK by corroborating drug coefficient estimates of the other models, but rejecting significant signals for the other three labs. The context model was less impactful in the case of ibuprofen, in which it exclusively corroborates the autoregressive drug and lab model without adjustment. This suggests that inpatient admission is more relevant for evaluating effects of simvastatin than ibuprofen in the EHR.

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

Adding context-related variables to multivariate lagged linear regression

We evaluated lagged drug coefficients for a multivariate autoregressive model that incorporates patients’ admission timelines as well as drug and lab measurement histories for all 20 drug-lab pairs using intra-patient normalized time series of differences. This allowed us to consider the extent to which contextual information, like inpatient admission, can recalibrate estimated effects of drugs on physiologic measurements. Using this method, we detected the same 6 known drug effects that were captured using

the autoregressive drug and lab model. Little difference was seen in number or magnitude of significant drug coefficients between the autoregressive drug and lab model and the context model *in cases of expected physiologic drug effects*. However, the context model did produce significantly different results for some of the uncharacterized drug-lab pairs.

Figure 4a demonstrates that autoregressive context terms “explain away” the effect of simvastatin on creatinine, potassium, and hemoglobin (hemoglobin has only 1 significant coefficient in the context model) by adjusting the lagged drug coefficients to insignificant quantities, while keeping intact simvastatin’s propensity to increase Total CK via muscle damage. Figure 4b shows that admission does not, however, contribute additional information toward studying the effect of ibuprofen on the four considered lab tests. This result is striking, and suggests that inpatient admission is an important confounding variable to consider when analyzing temporal effects of simvastatin, but is largely unimportant for analyzing the effects of ibuprofen. While the significant relationships for ibuprofen may in fact be legitimate, it is equally possible that they are explained away by other process-related context variables.

Discussion

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. All lagged methods showed highest specificity and sensitivity, overall, with intra-patient normalized laboratory values, and multivariate methods performed best in these metrics when differences were used during pre-processing stages. All multivariate methods identified the same six physiologic effects documented in clinical literature. Adding variables that

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.

It is useful to note that the inpatient admission variables did not affect drug coefficients for warfarin or ibuprofen for most labs. This simply indicates that there is little correlation between admission, the drug, and each lab that cannot be explained by the drugs and labs alone. In this study, we are not interested in confounder coefficient magnitude or their relative explanatory contributions—rather, we are looking for the co-linearities that *eliminate* or *enable* rightfully significant drug coefficients (i.e. explaining away). Adding context-variables to autoregressive models appears to be a simple way to probe for confounders in EHR data without biasing the analysis. Selecting other possible confounders, like surgery, is a worthwhile exercise in better explaining the trends we observed with little physiologic interpretation. However, adding multiple new variables effectively adds new time points for interpolation in the time series, making each lag represent less average clock-time. It is unclear how many variables can be included in the timeline construction without detrimental distortion of results.

It is difficult to discuss the detected positive associations that are not reported in clinical literature. On one hand, data from the EHR are liable to bias, and confounders can only be accounted for explicitly in our formulation. On the other hand, subtle drug effects are likely to be unstudied, yet omnipresent in medical practice. For this reason, we do not formally compute specificity or sensitivity metrics. We were also unable to detect drops in hemoglobin to implicate ibuprofen and warfarin in bleeding events. Our results suggested that warfarin and ibuprofen increase hemoglobin levels, which contradicts strong clinical evidence that they are causative agents of intestinal bleeding as defined by low hemoglobin levels [23]. It may be that physicians are very careful to maintain normal or even high hemoglobin levels when concerned about bleeding. This is an open question that merits further investigation, and we hypothesize that introducing relevant context variables (e.g. surgery, other labs) in the autoregression could help untangle this problem. In general, different data sources may impose different limitations on inference. It is unclear why we did not observe a drop in hemoglobin, but it may be a natural limitation of the data we used, which may lack the measurement frequency necessary to capture short-term fluctuations in hemoglobin levels.

More broadly, we wish to better understand how temporal dependence between lagged variables (even after taking differences) manifests in the coefficient estimation. We found that changing the number of estimated lags often does not significantly alter the trend of coefficients. Specifically, similar trends for amphotericin B predicting elevated creatinine persist whether estimating 10, 30, or 60 lags. The plots, not shown, are all close to zero at their endpoints and peak in the middle. This is clearly an effect of co-linearity between lags of a particular variable, and it may be important to employ model selection methods, like Akaike information criterion [24] and Bayesian information criterion [25], to reduce the number of variables and their co-dependencies. These and other methods may assist in automating and optimizing feature selection (we hand-selected drugs and labs that are consistently taken across large patient cohorts at our center). Model selection criteria are also important for formally evaluating Granger causality between correlated variables, and we hope to have a method that rigorously evaluates causality.

We seek a reliable high-throughput method that produces meaningful and interpretable results that will allow us to uncover new associations—we can only build the methodology for detecting known associations *de novo*, but it is hard to know if such a method will be as good at detecting previously hidden associations. Moreover, given that drug coefficients are lagged by sequence, it is difficult to map results back to actual clock-time. In principle, each lag τ is associated with a distribution of time lengths across all patients, and the summary statistics for those distributions may provide sufficient insight into the unique time-scale of each temporal process we detect.

Linear and non-linear distributed lag models have been used widely to study the temporal relationships between environmental factors and rates of health-related incidents, like suicide, mortality, and infectious disease [26,26–29], but their adoption in work with EHR data has been less common. These methods offer the advantage of explicit handling of confounding variables by including their autoregressive terms in the regression. More complex approaches also exist for removing confounders in distributed lag models, and Bahadori and Liu have proposed a methodology for applying Granger causality to medical data that is designed to learn the effects of unobserved confounders [30]. Ghassemi *et al.* used multi-task Gaussian process models for multivariate time series modeling of data collected in intensive care units (ICU) [31], and Joshi and Szolovits have applied novel unsupervised data mining techniques to characterize severity of patient physiology in ICUs [32]. Other noteworthy approaches for temporal pattern discovery have been applied to EHR data [5–7,33]. Jung and Shah evaluated the effect of non-stationarity in EHR data on different machine learning models, and found suboptimal performance of complex methods that ignore non-stationarity [34]. In addition, Lasko *et al.* have demonstrated machine learning methods that evaluate sampling rates and biases in laboratory measurements [35] and model temporal effects for both

continuous and categorical variables from the EHR [3,36]. While machine learning and pattern recognition methods can uncover complex relationships in high dimensional data, the lagged linear regression methods we use are advantageous due to their simplicity of implementation and interpretability.

This study was limited to EHR data from one medical center, and was limited to eight hypotheses based on review of clinical literature. However, there does not exist a ground truth for these hypotheses. Similarly, the twelve drug-lab pairs that were not found to be significantly linked in clinical literature cannot be verified and were thus exploratory, rather than formal negative controls.

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 for 20 example drug and lab pairs. 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, and a more thorough study with a larger feature set is warranted.

Table 1. Model performance*

Drug	Variables in model			lab_drug_context	lab_drug	lab_drug	lab_drug	lab_drug	drug	drug
	Extra variable in TS			inpatient	inpatient	none	none	none	none	none
	Differences			diffs	diffs	diffs	no diffs	diffs	diffs	no diffs
	Normalized			norm	norm	norm	norm	no norm	norm	norm
	Model type			Multivariate	Multivariate	Multivariate	Multivariate	Multivariate	Multivariate	Univariate
Lab	Expected	Citation								
amphotericinB	Hemoglobin			0+, 21-	0+,21-	0+,16-	1+, 3-	0+,14-	0+,4-	0+, 30-
amphotericinB	Total CK			0+, 5-	0+,5-	0+,8-	1+, 2-	0+,0-	0+,6-	0+, 30-
amphotericinB	creatinine	pos	[37]	22+, 0-	22+,0-	21+,0-	1+, 0-	0+,0-	19+,0-	26+, 0-
amphotericinB	potassium	neg	[38]	0+, 22-	0+,22-	0+,22-	1+, 0-	0+,20-	0+,10-	0+, 30-
ibuprofen	Hemoglobin	neg	[23]	28+, 0-	28+,0-	27+,0-	5+, 0-	0+,0-	13+,0-	30+, 0-
ibuprofen	Total CK			0+, 8-	0+,8-	0+,1-	0+, 2-	0+,0-	0+,0-	0+, 4-
ibuprofen	creatinine	pos	[39,40]	26+, 0-	26+,0-	30+,0-	2+, 0-	9+,0-	18+,0-	0+, 5-
ibuprofen	potassium	Possible pos	[39,40]	23+, 0-	24+,0-	23+,0-	7+, 0-	23+,0-	8+,0-	30+, 0-
simvastatin	Hemoglobin		[41]	1+, 0-	0+,2-	0+,0-	0+, 3-	0+,1-	0+,22-	5+, 0-
simvastatin	Total CK	pos	[42]	5+, 0-	8+,0-	8+,0-	2+, 1-	9+,0-	7+,0-	12+, 0-
simvastatin	creatinine			0+, 0-	0+,4-	0+,6-	0+, 0-	2+,0-	0+,7-	0+, 10-
simvastatin	potassium			0+, 0-	0+,7-	0+,11-	1+, 0-	0+,0-	0+,8-	18+, 0-
spironolactone	Hemoglobin			28+, 0-	27+,0-	28+,0-	2+, 2-	28+,0-	21+,0-	0+, 5-
spironolactone	Total CK			0+, 0-	0+,0-	0+,2-	0+, 1-	0+,0-	0+,0-	5+, 0-
spironolactone	creatinine	pos	[43]	23+, 0-	23+,0-	25+,0-	7+, 0-	2+,1-	27+,0-	21+, 1-
spironolactone	potassium	pos	[44]	28+, 0-	28+,0-	29+,0-	8+, 1-	27+,0-	27+,0-	30+, 0-
warfarin	Hemoglobin	neg	[23]	30+, 0-	30+,0-	30+,0-	10+, 3-	30+,0-	29+,0-	30+, 0-
warfarin	Total CK			0+, 4-	0+,5-	0+,16-	0+, 1-	0+,0-	1+,0-	1+, 14-
warfarin	creatinine			28+, 0-	27+,0-	29+,0-	7+, 1-	25+,1-	28+,0-	0+, 18-
warfarin	potassium			30+, 0-	30+,0-	30+,0-	8+, 2-	28+,0-	18+,0-	30+, 0-

*Pairs show the number of statistically significant positive and negative lags (e.g., “2+,1-” implies two positive lags and one negative lag). Green implies predominantly positive association, red implies predominantly negative, and grey implies minimal (less than 3) or mixed.

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, et al., 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, et al., 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, Parameterizing time in electronic health record studies, *Journal of the American Medical Informatics Association*. 22 (2015) 794–804. doi:10.1093/jamia/ocu051.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [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.
- [21] D. Bates, M. Maechler, *MatrixModels: Modelling with Sparse And Dense Matrices*, 2015. <https://CRAN.R-project.org/package=MatrixModels>.
- [22] A.C. Davison, D.V. Hinkley, *Bootstrap methods and their application*, Cambridge University Press, Cambridge ; New York, NY, USA, 1997.
- [23] J.P. Hreinsson, S. Palsdóttir, E.S. Bjornsson, The Association of Drugs With Severity and Specific Causes of Acute Lower Gastrointestinal Bleeding: A Prospective Study, *J. Clin. Gastroenterol.* (2015). doi:10.1097/MCG.0000000000000393.
- [24] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*. 19 (1974) 716–723. doi:10.1109/TAC.1974.1100705.
- [25] G. Schwarz, Estimating the Dimension of a Model, *Ann. Statist.* 6 (1978) 461–464. doi:10.1214/aos/1176344136.

- [26] A. Gasparrini, Y. Guo, M. Hashizume, E. Lavigne, A. Zanobetti, J. Schwartz, et al., Mortality risk attributable to high and low ambient temperature: a multicountry observational study, *The Lancet*. 386 (2015) 369–375. doi:10.1016/S0140-6736(14)62114-0.
- [27] A. Gasparrini, B. Armstrong, M.G. Kenward, Distributed lag non-linear models, *Stat Med*. 29 (2010) 2224–2234. doi:10.1002/sim.3940.
- [28] D.M. Stephen, A.G. Barnett, Effect of temperature and precipitation on salmonellosis cases in South-East Queensland, Australia: an observational study, *BMJ Open*. 6 (2016) e010204. doi:10.1136/bmjopen-2015-010204.
- [29] K. Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, B. Armstrong, Time series regression studies in environmental epidemiology, *Int J Epidemiol*. 42 (2013) 1187–1195. doi:10.1093/ije/dyt092.
- [30] M.T. Bahadori, Y. Liu, An examination of practical granger causality inference, in: *Proceedings of the 2013 SIAM International Conference on Data Mining*, SIAM, 2013. <http://epubs.siam.org/doi/abs/10.1137/1.9781611972832.52> (accessed March 10, 2016).
- [31] M. Ghassemi, M.A.F. Pimentel, T. Naumann, T. Brennan, D.A. Clifton, P. Szolovits, et al., A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data, in: *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. <http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9393> (accessed March 11, 2016).
- [32] R. Joshi, P. Szolovits, Prognostic Physiology: Modeling Patient Severity in Intensive Care Units Using Radial Domain Folding, *AMIA Annu Symp Proc*. 2012 (2012) 1276–1283.
- [33] X. Wang, D. Sontag, F. Wang, Unsupervised Learning of Disease Progression Models, n.d.
- [34] K. Jung, N.H. Shah, Implications of non-stationarity on predictive modeling using EHRs, *J Biomed Inform*. 58 (2015) 168–174. doi:10.1016/j.jbi.2015.10.006.
- [35] T.A. Lasko, Nonstationary Gaussian Process Regression for Evaluating Clinical Laboratory Test Sampling Strategies, *Proc Conf AAAI Artif Intell*. 2015 (2015) 1777–1783.
- [36] T.A. Lasko, Efficient Inference of Gaussian-Process-Modulated Renewal Processes with Application to Medical Event Data, in: *Uncertainty in Artificial Intelligence: Proceedings of The... Conference*. Conference on Uncertainty in Artificial Intelligence, NIH Public Access, 2014: p. 469. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4278374/> (accessed March 10, 2016).
- [37] A.D. Luber, L. Maa, M. Lam, B.J. Guglielmo, Risk factors for amphotericin B-induced nephrotoxicity, *Journal of Antimicrobial Chemotherapy*. 43 (1999) 267–271.
- [38] E. Usami, M. Kimura, T. Kanematsu, S. Yoshida, T. Mori, K. Nakashima, et al., Evaluation of hypokalemia and potassium supplementation during administration of liposomal-amphotericin₃, *Experimental and Therapeutic Medicine*. (2014). doi:10.3892/etm.2014.1534.
- [39] T.I. Poirier, Reversible renal failure associated with ibuprofen: case report and review of the literature, *Drug Intell Clin Pharm*. 18 (1984) 27–32.
- [40] T. Ulinski, V. Guigonis, O. Dunan, A. Bensman, Acute renal failure after treatment with non-steroidal anti-inflammatory drugs, *Eur. J. Pediatr*. 163 (2004) 148–150. doi:10.1007/s00431-003-1392-7.
- [41] M.J. Robbins, A. Iqbal, R. Hershman, Lovastatin-induced hemolytic anemia: not a class-specific reaction, *The American Journal of Medicine*. 99 (1995) 328–329.
- [42] P.H. Jones, M.H. Davidson, E.A. Stein, H.E. Bays, J.M. McKenney, E. Miller, et al., Comparison of the efficacy and safety of rosuvastatin versus atorvastatin, simvastatin, and pravastatin across doses (STELLAR* Trial), *The American Journal of Cardiology*. 92 (2003) 152–160. doi:10.1016/S0002-9149(03)00530-7.
- [43] B. Williams, T.M. MacDonald, S. Morant, D.J. Webb, P. Sever, G. McInnes, et al., Spironolactone versus placebo, bisoprolol, and doxazosin to determine the optimal treatment for drug-resistant hypertension (PATHWAY-2): a randomised, double-blind, crossover trial, *The Lancet*. 386 (2015) 2059–2068.
- [44] G.C. Roush, M.E. Ernst, J.B. Kostis, S. Yeasmin, D.A. Sica, Dose doubling, relative potency, and dose equivalence of potassium-sparing diuretics affecting blood pressure and serum potassium: systematic review and meta-analyses, *J. Hypertens*. 34 (2016) 11–19. doi:10.1097/HJH.0000000000000762.