Comparing lagged linear methods for uncovering associations in EHR data

Matthew Levine, David Albers, and George Hripcsak

Department of Biomedical Informatics, Columbia University

The problem: Developing reliable high-throughput methods for identifying adverse drug effects that are easy to implement and produce readily interpretable results.

- Many observational health research inquiries can be satisfied by considering whether a patient ever had a particular condition, and by considering relevant time windows with respect to conditions of interest.
- However, it can be useful to consider methods with the potential to reveal fine temporal structure in EHR data
- Many of these approaches to time-series analysis rely on assumptions of stationarity that are frequently broken by clinical data.
- In addition, clinicians often sample patients at rates proportional to their health variability, inducing stationarity by indexing the time series not by *clock-time*, but rather by *measurement sequence*

What we know from our past work

1. Temporal clinical and physiologic processes can be described through *lagged linear correlation of concepts* extracted from signout notes and laboratory values
2. Time series data, under some clinical circumstances, are *better parameterized by their raw sequence than their clock measurements*
3. Health-care process events such as inpatient admission are systematically correlated with concepts and laboratory values

Summary

Here, we define and use 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.

1. We outline methods for time series construction
2. We evaluate the effects of intra-patient normalization and differences
3. We compare univariate and multivariate lagged linear regression
4. We consider the impact of including autoregressive lab terms to the model
5. We observe how adding context-related variables to the multivariate lagged model provide a method for explicitly probing for confounding.

The data

NewYork-Presbyterian Hospital clinical data warehouse

- Available in OMOP CDMv4
- Over 3 million patients
- 27 years old

Extracted drug orders (timestamp and MED code)–binary values
–Amphotericin B, Ibuprofen, Simvastatin, Spironolactone, Warfarin Extracted lab values (timestamp, value, and MED code)–real numeric values
–Total Creatine Kinase, Creatinine, Potassium, Hemoglobin Extracted inpatient admission events (timestamp)–binary values

We acknowledge NLM R01 LM06910 and NIDDK RO1DK090372 for financial support.

Timeline Construction–Linear Temporal Interpolation

For every time point where there was a concept (lab, drug, inpatient admission), the values of other variables at that time point were interpolated as the weighted (by clock-time) mean of the two surrounding values. Thus, all concepts, whether from categorical or real-valued sources, took on real-valued pairs at each time point.

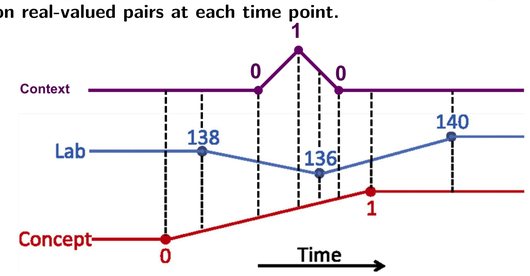


Figure 1: Laboratory values were continuous, and orders for the drug of interest were represented as 1 (present), whereas orders for other drugs were represented as 0 (absent). Inpatient admission timelines were defined with a 1 at the time of admission, and zeros at 24hrs before and after admission, effectively creating spikes at times of admission.

Pre-processing steps

Intra-patient normalization: Laboratory values were normalized within each patient to have mean=0, variance=1

$$y_t^* = \frac{y_t^{raw} - \bar{y}_t}{var(y_t^{raw})} \tag{1}$$

Differences: Each value was replaced with its difference from its preceding value

$$y_t = y_t^* - y_{t-1}^* \tag{2}$$

Lagged linear models–predicting laboratory values

Let x denote drug values, y denote lab values, and z denote admission values. Let t index the sequence of interpolated values and let L denote number of sequential lags ($L = 30$).

Univariate Lagged Linear Regression (ULLR):

$$y_t = c + \beta_\tau x_{t-\tau} + \epsilon \tag{3}$$

Multivariate Lagged Linear Regression (MLLR) Drug model:

$$y_t = c + \sum_{\tau=1}^{L} \beta_{x,\tau} x_{t-\tau} + \epsilon \tag{4}$$

MLLR Autoregressive drug and lab model:

$$y_t = c + \sum_{\tau=1}^{L} \beta_{y,\tau} y_{t-\tau} + \sum_{\tau=1}^{L} \beta_{x,\tau} x_{t-\tau} + \epsilon \tag{5}$$

MLLR Autoregressive drug, lab, and context model:

$$y_t = c + \sum_{\tau=1}^{L} \beta_{y,\tau} y_{t-\tau} + \sum_{\tau=1}^{L} \beta_{x,\tau} x_{t-\tau} + \sum_{\tau=1}^{L} \beta_{z,\tau} z_{t-\tau} + \epsilon \tag{6}$$

Temporal lab prediction with ULLR

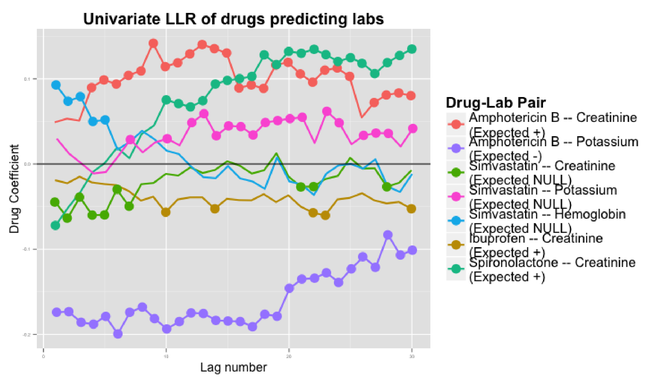
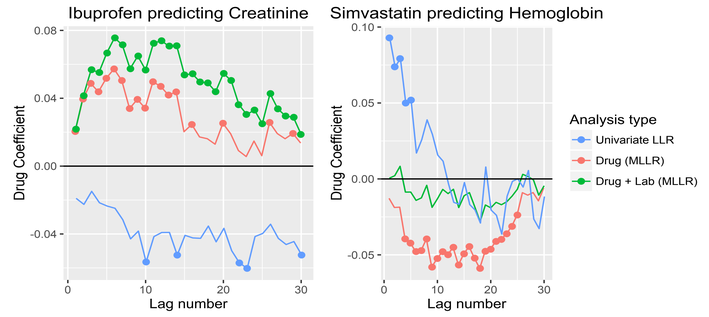


Figure 1: The ULLR model predicts expected directional effects of amphotericin B, but attributes undue significance to simvastatin and misdirects the effects of spironolactone and ibuprofen on creatinine.

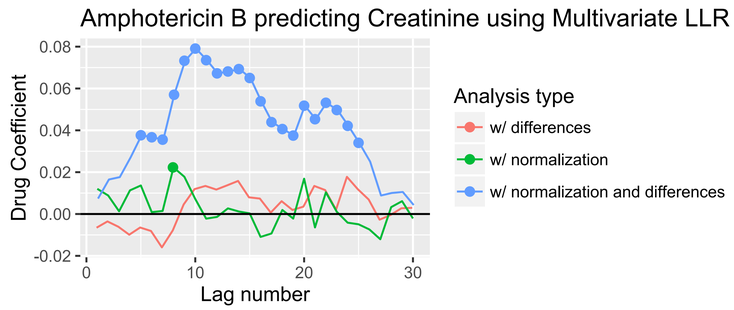
Adding autoregressive terms to MLLR

Adding AR terms reduces false positives and strengthens existing associations.



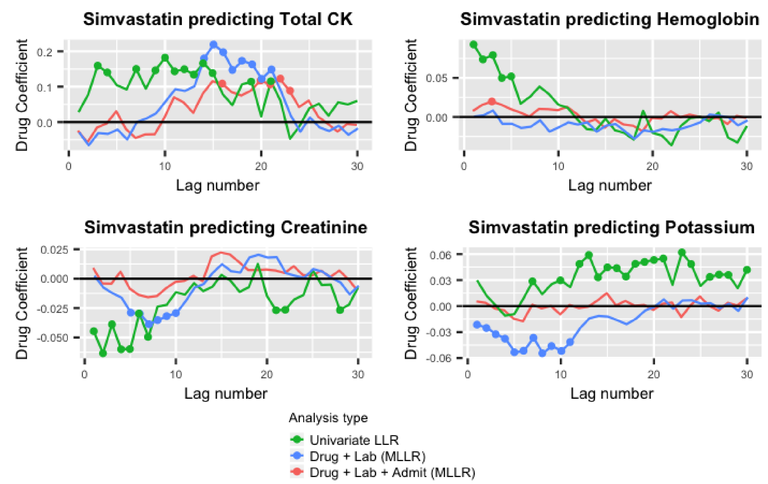
Importance of pre-processing steps

Intra-patient normalization and differences produce the most reliable results.



Adding context-related variables to MLLR

Inpatient admission explains away unexpected associations.



Conclusions and Future Directions

- Intra-patient normalization and differences are both beneficial in MLLR context.
- Multivariate lagged linear methods have better sensitivity and specificity than univariate models.
- Autoregressive terms of lab values improve predictions of drug effects.
- Context-related variables like inpatient-admission sometimes confound drug effects, and including them in the MLLR model can correct for this.