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