Published in final edited form as:

Biometrics. 2015 June ; 71(2): 478–486. doi:10.1111/biom.12283.

Detecting Clinically Meaningful Biomarkers with Repeated Measurements: An Illustration with Electronic Health Records

Benjamin A. Goldstein[1,*], **Themistocles Assimes**[2], **Wolfgang C. Winkelmayer**[3], and **Trevor Hastie**[4]

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, U.S.A.

[2]Division of Cardiovascular Medicine, Stanford University School of Medicine, Palo Alto, California, U.S.A.

[3]Section of Nephrology, Baylor College of Medicine, Houston, Texas, U.S.A.

[4]Department of Statistics, Stanford University, Palo Alto, California, U.S.A.

Summary

Data sources with repeated measurements are an appealing resource to understand the relationship between changes in biological markers and risk of a clinical event. While longitudinal data present opportunities to observe changing risk over time, these analyses can be complicated if the measurement of clinical metrics is sparse and/or irregular, making typical statistical methods unsuitable. In this article, we use electronic health record (EHR) data as an example to present an analytic procedure to both create an analytic sample and analyze the data to detect clinically meaningful markers of acute myocardial infarction (MI). Using an EHR from a large national dialysis organization we abstracted the records of 64,318 individuals and identified 4769 people that had an MI during the study period. We describe a nested case-control design to sample appropriate controls and an analytic approach using regression splines. Fitting a mixed-model with truncated power splines we perform a series of goodness-of-fit tests to determine whether any of 11 regularly collected laboratory markers are useful clinical predictors. We test the clinical utility of each marker using an independent test set. The results suggest that EHR data can be easily used to detect markers of clinically acute events. Special software or analytic tools are not needed, even with irregular EHR data.

Keywords

Biological markers; Dialysis; Longitudinal data; Myocardial infarction; Risk prediction; Splines

1. Introduction

Electronic health records (EHRs) constitute a relatively new data source that are being used to understand and predict near-term clinical events (Goldstein et al., 2014). They are characterized by having dense, serial information on patients receiving clinical care,

[*]ben.goldstein@duke.edu.

allowing for a granular view of a patient's evolving health status. A National Heart Lung and Blood Institute working group recently prioritized the assessment of near term risk of acute cardiac events (Eagle et al., 2010). Specifically the group focused on the use of biomarkers to make such assessments.

Before developing predictive models, it is first necessary to detect potentially useful markers. If the biomarker were measured once, a typical approach to detect such markers would be to perform a logistic regression, regressing the probability of an event onto the marker and other covariates, such as:

$$logit(P(\text{Cardiac Event}))=\alpha+\beta_1\text{Marker}+\beta_W\text{Covariates} \quad (1)$$

where β_1 would be the parameter of interest, representing whether changes in the marker change the probability of the event of interest. Of course, one of the key advantages of EHR data is that markers are measured over time. While this allows for a more sophisticated view of changes it also makes the analysis more challenging. The analysis can be simplified by averaging or summarizing laboratory values across time but this may result in a loss of information. Instead we ideally want to consider variation in the marker over time. We can reform model (1) as:

$$logit(P(\text{Cardiac Event}))=\alpha+\int\beta_1(t)\text{Marker}(t)\mathrm{d}t+\beta_W\text{Covariates} \quad (2)$$

where now we are integrating over multiple time points, *t*. To fit such a model we can consider discretizing time, however, depending on the number of time points, this may result in a very high dimensional model. Complicating matters further is that EHR measures are taken irregularly and sometimes sporadically, meaning patients generally do not have laboratory measures at comparable time points and frequencies. This makes standard analytic techniques, which often require well-aligned measurements, challenging.

The difficulty of estimating model (2) is reflected in the complex theoretical work and software that others have developed to fit it (James, 2002; Yao et al., 2005; Gertheiss et al., 2013). James used a two-stage errors in variable model with cubic splines to estimate individual curves where the dimension of the spline is larger than the observed observations. Gertheiss et al. took a modified imputation approach to get measurements on the same time scale. Yao et al. utilized functional PCA to estimate the curves. Furthermore, while analyses of the form of model (2) have appeal from a predictive standpoint, they do not necessarily address the specific question of interest: namely is a given measure a clinically useful biomarker of an impending event. Another way of phrasing the question is: does a given laboratory measure show different and detectable patterns among those that experience an event? Moreover, we may ask, can the biomarker history, through a given time point, stratify risk in clinically meaningful ways?

With these questions in mind, we suggest a relatively straightforward solution to detecting clinically useful biomarkers. The model we propose is flexible enough to not only answer a series of questions about the utility of a laboratory measure to serve as a predictive marker, but also to allow for the detection of these relationships using established statistical methods. We illustrate this approach using EHR data from patients with end stage renal disease (ESRD) undergoing hemodialysis (HD). Patients undergoing outpatient HD are at increased risk of cardiac events, particularly myocardial infarction (MI). Cardiac disease accounts for 43% of all cause mortality with approximately 20% due to MI (Herzog, 1999). Moreover, patients receiving outpatient HD typically have routine and regularly scheduled monitoring of several laboratory values for months or years at a time. Therefore, patients undergoing HD represent an ideal population to study the role of repeated laboratory measures in the detection of an impending MI.

The rest of the article is arranged as follows: in Section 2 we describe the available data. Since one of the challenges in working with EHR data is appropriately selecting an analytic cohort, we also describe a generally useful sampling design. In Section 3, we walk through our proposed analysis, which consists of a series of goodness-of-fit tests. In Section 4, we present the analytic results and conclude in Section 5.

2. Data Description & Sample Selection

Working with EHR data presents unique opportunities and challenges. We first note that EHR data are inherently observational, implying all of the caveats and limitations of non-experimental data. The primary strength and challenge of EHR data are its longitudinal nature, with individuals having multiple measurements over time. While presenting the opportunity to observe changes over time—the primary aspect of the present analysis—this can become complicated since measurements are often taken irregularly. In some EHRs—though not the current one—the presence of a measure may serve as a risk indicator itself, for example, a patient feeling ill and visiting a doctor, producing a measurement in the EHR.

The first challenge is how to appropriately sample an analytic cohort from the EHR. In the present study, we are interested in identifying potential markers of acute MI. This lends itself well to a retrospective analysis: identify those people with an MI and observe how different markers change before the event. The subtler question is who is the comparative group (i.e., controls) and at what point in time should they be analyzed. Below we describe the data available, how we define the cases and more importantly how we sample the controls.

2.1. Data Source

We used two data sources in the analysis: the United States Renal Data System (USRDS) and the EHR from DaVita, Inc. The USRDS is a national registry that includes almost all persons with ESRD (USRDS, 2013). It is created from medical claims submitted to Medicare, which is mandated by law to pay for the healthcare of the majority of patients with ESRD, regardless of the age of patients at the start of their HD treatments. DaVita Inc. is the second largest chain of outpatient dialysis centers in the country. Their EHR contains detailed session level information on patient dialysis session, laboratory values, hemodynamic metrics, and more. We used an anonymous crosswalk provided by the USRDS Coordinating Center to link the two datasets. This was conducted under a Data use Agreement between the National Institute of Diabetes and Digestive and Kidney Diseases and one of the authors (WCW).

2.2. Selecting the Sample

One can consider an EHR as analogous to a large prospective cohort where only a small fraction of the cohort will experience an event, each at different time points. With this in mind, we describe a sampling approach motivated from nested case-control designs to sample appropriate controls along with eligible cases (Wacholder et al., 1992). We illustrate this process in Figure 1.

2.2.1. Eligible sample—Any individual who initiated HD between January 1, 1995 and December 31st, 2008 and was a patient at a DaVita, Inc., dialysis facility between January 1st, 2004 and December 31st, 2008 was eligible for study. Using the USRDS payer history file, we retained only those patients who were aged ≥67 at the initiation of dialysis and had at least 2 years of uninterrupted fee-for-service Medicare coverage before their reported first dialysis (first service date). Selecting this subset of individuals has two advantages. First we can observe the health-care claims and associated diagnoses and procedures before the onset of ESRD. This provides us with increased confidence that we are detecting an incident MI and not a claim related to a previous MI. Second, we can be near-certain that all health claims are recorded at the time of initiation of dialysis, without having to apply an eligibility window. We excluded all individuals with a history of an MI, defined through the presence of any of the following ICD-9 codes: 410.** and 412. To be as sensitive as possible, patients with any inpatient code or outpatient codes were removed from analysis.

2.2.2. Cases—Cases were subjects who developed incident MI between 2004 and 2008 while receiving ongoing dialysis treatment at DaVita, Inc. We defined a case as "active" if a laboratory measurement was recorded within 14 days of the qualifying event. Events were identified from either (a) the presence of an ICD-9 code of 410.** during a hospitalization (positive predictive value 96.9% (Petersen et al., 1999)) or (b) a primary cause of death being reported as due to MI (Code 2 or 23) on the death notification record to Medicare.

2.2.3. Controls—Sampling of controls is the primary challenge in designing retrospective, longitudinal analyses. For this analysis we suggest a nested case-control design (see below for other design considerations). For nested case-control designs, we want to sample a control whenever someone becomes a case, referred to as *incident density* sampling. To avoid potential bias, controls are sampled *with-replacement* meaning that it is possible for a control to be sampled more than once, or serve as both a case and control (Lubin and Gail, 1984; Robins, Gail, and Lubin, 1986). For example, a patient who was diagnosed with ESRD on 7/1/2006 and had an incident MI on 5/1/2008, would be eligible to serve as a control during the period preceding the MI.

In the EHR setting, there are two potential time domains upon which to sample: calendar time and clinical time, that is, the time since start of maintenance/chronic dialysis treatment for ESRD (also called "vintage"). We decided to sample controls based on calendar time and adjust for vintage. For all cases during a calendar month, an equal number of controls were sampled, creating an index date. While it is typical in nested case-control design to sample *matched* controls we chose not to perform such matching to avoid the additional complications (Cai and Zheng, 2012), but instead simply adjusted for covariates.

2.2.4. Sample split—To assess the proposed procedure, we divided the sample into a discovery set consisting of incident events and corresponding controls between 2004 and 2007 and an independent validation set consisting of incident events and controls within 2008.

2.3. Selecting Variables

2.3.1. Predictors of interest—Through the DaVita EHR, data were abstracted on 11 regularly collected laboratory measures: albumin, calcium, CO_2, creatinine, ferritin, hemoglobin, iron saturation, phosphorous, platelet count, potassium, and white blood cell count. It is important to note that these laboratory measures are collected per-protocol and not based on a patient's clinical characteristics. Table 1 lists the predetermined acceptable ranges and approximate frequency of collection. Any laboratory measures that fell outside these ranges were removed.

In order to analyze changes in laboratory measures over time, laboratory values for up to 180 days preceding the index data were abstracted. Patients were not required to have a minimal number of laboratory measures.

2.3.2. Covariates—Since we are not interested in estimating the direct association of the given laboratory measure but simply its utility as a biomarker, a minimal number of covariates were included in the analysis. Specifically, analyses were adjusted for patients age at time of ESRD, gender, race (Caucasian, African American, Hispanic, Asian, and other), and vintage (time since ESRD).

3. Analytic Approach

3.1. The Statistical Model

The goal of this study is to present a means of detecting clinically relevant laboratory markers of an impending clinical event. Therefore, in contrast to model (2) we are not interested in estimating the probability of MI given a sequence of laboratory measures, but instead modeling how the sequence of laboratory measures may differ between cases and controls. We consider that person i has n_i measurements of a given laboratory measure, at times t_{i1}, t_{i2}, …, t_{in_i}. We can fit a general model of the form:

$$Lab(t)=\alpha+1_{MI}\beta_1+1_{MI}f(t)+1_{1-MI}f^*(t)+W(t)\beta_W+\varepsilon. \quad (3)$$

The outcome variable is the laboratory measure, measured at multiple time points t. 1_{MI} is an indicator for whether the person has an MI with 1_{1-MI} the complement (i.e., case or control). f represents a general function to flexibly estimate changes in laboratory measures over time. Therefore, cases and controls are allowed to have different patterns over time. Finally, additional covariates (potentially time varying) are represented by W.

The primary analytic question is how to represent the function, f. Following the work of others we use regression splines, using a q-dimensional vector of basis functions $s(t)$, and hence $f(t) = s(t)' \gamma$. In our representation $s(t)$ is specified using $k = q - 1$ knots. $s(t)$ would be evaluated n_i times, filling the rows of a $n_i \times q$ basis matrix, where n_i is the number of observed laboratory measurements as above. These are produced for each person, and combined into an overall spline model matrix. To fit the model we estimate the parameter vector γ, a *q*-dimensional coefficient vector. Different spline formulations can be used, we consider truncated cubic power splines with basis functions:

$$\left(t, \left\{(t-\xi_k)_+^3\right\}_1^K\right)$$

evaluated at each knot, ξ_k. We note the lack of intercept. While natural splines are more commonly used over truncated power splines, the truncated power spline basis has the advantage of being linear to the left of the leftmost boundary knot while non-linear to the right. This is a feature we illustrate and exploit below. We placed $K = 5$ knots at, 150, 90, 60, 30, 14 days prior to the index date. By placing more knots closer to the index date we are able to capture more subtle changes directly prior to that date. Therefore, the final model is:

$$Lab_{ij}=\alpha_i+1_{MI_i}\beta_1+S'(t_{ij})\gamma+\left(1_{MI_i}\times S'(t_{ij})\right)\gamma^*+W_{ij}\beta_W+\varepsilon_{ij} \quad (4)$$

Here, we have indexed by person *i* for record *j*. Each individual has multiple observations so we include a random intercept, α_i. Since 1_{MI_i} is an indicator function, the spline basis for the controls is represented by $S'(t_{ij})\gamma$ and the basis for the cases $S'(t_{ij})\gamma + S'(t_{ij})\gamma^*$, allowing for two separate functional representations for cases and controls. Model (4) is easily estimated as a linear regression with a random intercept, a spline basis for the timing of the laboratory measurements, and an interaction term between the spline basis and case-control status.

3.2. Criterion for Clinically Meaningful Differences

Using model (4) as a general form, we conduct a series of goodness of fit tests to assess a set of clinical questions. To motivate these criterion we consider the prospective scenario where one is tracking a patient's laboratory measure over time and wants to determine whether the pattern indicates a risk of MI. Therefore, the goal of the analysis is to detect those laboratory measures that can be so used.

The first question is whether the trajectory of laboratory values differs between cases and controls. For this assessment the primary parameter of interest is the vector γ^*, which represents the difference between the curves for those that experience an MI compared to those that do not. To formally test whether the two curves are different we perform a likelihood ratio test comparing the full model to a nested model that does not contain γ^*, that is, a model where the only difference in laboratory measures between those that experience an MI and those that do not is represented as a shift through β_1. A rejection of the null hypothesis that the fits are equivalent, indicates that the laboratory measures differ over time between cases and controls.

A second consideration is the trajectory of a marker over time. Specifically, for a measure to have clinical utility, we would expect that those not experiencing an event (controls) should present predictable and stable patterns. Conversely, the values among those about to experience an event (cases) should show a deviation from this stable pattern. While we could hypothesize various "stable" patterns, for simplicity we consider linearity to imply stability. Therefore, the laboratory measures for controls should be linear and for cases non-linear (i.e., curved). To assess this, we can fit model (4) among cases and controls separately. Therefore, β_1 and γ^* are removed from the model and the parameter of interest is the spline vector γ:

$$Lab_{ij}=\alpha_i+S'(t_{ij})\gamma+W_{ij}\beta_W+\varepsilon_{ij}. \quad (5)$$

This fit is compared to a reduced model that only includes a linear term for time. To call a laboratory measure a potentially good marker we want to reject the null hypothesis among the cases and *fail* to reject the null hypothesis among the controls, that is, cases should be non-linear and controls should be linear.

This establishes three criterion to declare a laboratory marker clinically useful:

1. The patterns over time should be different between cases and controls
2. Cases should show non-linearity over time
3. Controls should be linear over time

For each of the laboratory tests we considered a p-value less than 0.05 to indicate significance and performed a Bonferroni correction across the set of three tests. This was repeated separately for each of the 11 markers.

Among the laboratory measures that passed these criterion, a second question of interest is: how long before an event can changes be detected? We note that truncated power splines are linear to the left of the left-most (earliest in time) knot. We illustrate this concept using simulated toy data in Figure 2. We are fitting a non-linear function (in black) placing successive knots along the *x*-axis. We note, that to the left of the first knot (indicated by a dashed line) the estimated fit is linear.

Using this property, we can consider the optimal placement of the first knot to be the point at which the laboratory measures are linear before, that is, do not change over time. This can give us an indication as to when a laboratory measure for those that will experience an event begins to change.

To assess this, we fit a series of models of the form of model (5) among those that experience the event. We started with a simple linear model. Next we added a knot at 14 days before the event. Then we fit a third model adding a knot at 14 and 28 day before the event, etc. until we had a model with 12 knots up to 168 days. While we could have used a likelihood ratio tests to pick the optimal fit we ultimately did not view this as a specific hypothesis test and instead chose the model with the minimal AIC as the one with the best fit.

Finally, we visually inspected the fits from model (4) for each laboratory measure. We calculated and plotted predicted values for laboratory measures over time with pointwise 95% confidence bands.

3.3. Additional Design & Analytic Considerations

The proposed design and analysis is essentially a retrospective analysis. We also could have sampled the data prospectively. This would consist of analyzing all available patients (a full cohort) or sampling controls in a case-cohort design. The advantage of a full cohort approach is that we utilize all available data, potentially improving efficiency. Given the relative rarity of the outcome, the efficiency gain may not be noticed, and a case-cohort design could be preferable. In this sampling design, one needs to use proper sample weights, but there is improved efficiency over nested case-control designs (Barlow et al., 2014). Since these sampling strategies would involve prospectively following changes, either design would be most appropriate for analyses of the form of model (2), a prospective analysis where we are modeling disease status as a function of labs. As discussed in the outset, such a model may prove undesirable. Moreover, in addition to the analytic challenges, there is also a sampling challenge. These approaches require as designation of time 0, which would naturally be date of ESRD. However, one of the complications of EHR data is that people move in-and-out of the EHR, leading to the potential for high rates of missing data. In the present design we have tried to account for that by requiring patients to have been an “active” patient but additional care would be needed for such prospective analyses.

Within the nested case-control framework, one potential for concern is the lack of independence among observations. Even though the cases and controls are frequency matched (as opposed to pair matched) they may exhibit correlation—particularly if there are strong secular trends. Additionally, an individual may serve as both a case and control or a control more than once—particularly if the ratio of controls to cases is low. To account for this we added an additional random effect terms into model (4). The results were unchanged, likely due both to the minimal correlation between observations and the nature of the likelihood ratio test, so we provide the more basic models.

An additional concern is skewness in the data. The above models, model the mean laboratory level. Many markers can be highly skewed, making it more appropriate to log-transform the values. In our present data, while there were relatively long tails, the degree of skewness was quite low and we report the untransformed results.

3.4. Assessment

After detecting biomarkers, one typically desires to use them to develop a predictor. To validate the proposed method, we assessed how well the “discovered” markers predicted MI. As discussed, modifications to model (2) have been proposed to directly estimate the probability of an event given a vector of time varying measures. However, few have been implemented in regularly available software. We fit the procedure of Goldsmith et al. (2009) as applied in the refund package in R. Using the independent confirmatory set (data from 2008), we calculated the probability of MI for each individual based on the 11 separate laboratory measures. We assessed the improvement in prediction by comparing the misclassification rate (via McNemar’s test) and area under the ROC curve (*c*-statistic). We considered the marker to be “validated” if there was a significant improvement ($p < 0.05$) upon inclusion of the laboratory measure to a model containing only demographic factors.

All analyses were performed in R 3.0.1 using the lme4 packages to calculate the mixed models (R Core Team 2012), and our own function to calculate the truncated power spline basis (see Appendix).

4. Results

A total of 64,318 people were available for study between 2004 and 2008. After removal of individuals with a history of MI, we abstracted 3677 individuals with an incident MI between 2004 and 2007 and an additional 1092 individuals with an incident MI in 2008 to serve as a validation set. An equal number of controls were selected during the same time period. Of the 4769 cases, 516 (11%) had served as a control at an earlier point in time. Additionally, 491 (10%) controls were randomly selected to serve as controls more than once. There were similarities between those experiencing events in age and gender but meaningful differences in regards to race (Table 2). Those experiencing an event tended to have spent less time on dialysis.

Using model (4) described above we estimated the differences in trends of laboratory measures among those that experienced an MI and those that did not. A likelihood ratio test with a Bonferroni correction was performed to test whether the two curves differed (Table 3). Overall, 7 of the 11 tests showed significant differences between those that experienced an MI and those that did not. In our second analysis, we assessed whether the 11 markers were linear over time among those that ultimately have an MI and those that do not. Using model (5) we again performed a likelihood ratio test comparing nested models. This resulted in six laboratory measures that were clinically useful based on our predefined criterion of significance.

Among the six laboratory measures that met all three of the above criteria we examined the point at which the laboratory measures for those experiencing an MI began to depart from linearity. A series of models were fit, with each one adding an additional knot over time. The model with the minimal AIC was chosen as the best fit. Table 3 also shows the optimal fit for each of the six laboratory measures. Albumin, hemoglobin, phosphorous, and platelet-count showed optimal departure within 4 weeks of the event, suggesting that changes could be detected 1-month before an event occurs.

We visually inspected the patterns of change for each of the 11 markers (Figure 3a–k). Using the estimates from model (4) we predicted the laboratory measure for a dialysis patient about to experience an MI and a similar control, with 95% point-wise prediction intervals. Visual inspection confirms the analytic results. Of the laboratory measures that were not identified as useful markers, all but ferritin, did not visually show differences between those about to experience and MI and those who did not. Most of the successful laboratory markers showed departures from linearity immediately preceding the MI, as suggested by analysis 3. The one exception was iron saturation which visually appears to have it's greatest departure at about 14 days but analytically was identified at 168 days.

Finally, we assessed the predictive performance of each measure among an independent validation set of events. Table 3 contains the misclassification rate and *c*-statistic for each of the 11 laboratory measures. Using only the baseline covariates of age, sex, race, and vintage, the misclassification rate was 0.470 and the *c*-statistic was 0.541—suggesting minimal predictive value based off these baseline characteristics alone. Of the six laboratory measures that met the suggested criteria, four had a significant (p <0.05) improvement in their predictive performance, with albumin, hemoglobin and white blood cell showing the most predictive improvement. Conversely, of the five measurements that did not meet our criteria none yielded a significant improvement in misclassification or discrimination. Using a model with all six variables yielded a misclassification rate of 0.411 and *c*-statistic of 0.620, suggesting that in combination the markers do not provide additional information.

5. Discussion

In this article, we suggest a straightforward procedure to detecting clinically meaningful markers of an impending clinical event within an EHR. The irregular and longitudinal nature of the data can make analyzing EHRs challenging. While some theoretical work has been developed to address these challenges, these methods are not all readily accessible. Instead, we suggest an approach that utilizes regular statistical methods and software. While we have focused here on outpatient HD, we note that there are many other comparable scenarios within typical hospital settings where patients get serial measurements, such as inpatient Intensive Care Units, monitoring during surgery, and cancer treatment where this approach should also prove useful. Moreover, these methods extend beyond EHRs, but to any longitudinal dataset with biomarkers serially collected at different time points.

The two steps in such an analysis are to first appropriately select a study sample from the EHR and second to analyze the data. To select the sample, we utilized a nested case-control study. Others have used nested case-control designs noting both their suitability and advantages for prediction with EHR data (Irizarry et al., 2012; Wu, Roy, and Stewart, 2010). While such designs are common in epidemiological studies, they are less common in traditional statistical analyses. However, they provide a useful means to sample complex longitudinal data.

Using the proposed mixed model with spline basis functions we illustrate a variety of analytic questions one can ask to asses the clinical utility of a laboratory measure. These include: Do those that experience the event show a different pattern over time? Are the laboratory measures linear over time among the controls and non-linear among the cases? How far out can we detect non-linearity in the cases? Undoubtedly, given a specific clinical question one could imagine that different comparisons could be drawn. We consider this flexibility to be one of the strengths of the proposed procedure. For example, we easily could have constrained the controls to show constant laboratory values over time, or if it were known that a laboratory measure changed via circadian rhythms (e.g., blood pressure) and was continuously measured over a single day, stability for controls could be proposed to have a sinusoidal pattern.

We assessed this approach using data from an EHR system of patients undergoing hemodialysis. We identified 4769 people with an incident MI and abstracted 11 regular laboratory measures over a 6 month period before the event. Of the 11 measure, 6 met our criteria. We evaluated the results both qualitatively via visualization and quantitatively through fitting a prediction model on an independent set of data. Four of the measurements showed strong utility as a predictor, with the three most promising measures for assessing risk of MI were a drop in albumin and hemoglobin and a rise in white blood cell count. Not surprisingly these markers have previously been associated with risk of MI (Friedman, Klatsky, and Siegelaub, 1974; Ensrud and Grimm 1992; Djousse et al., 2002; Bassand et al., 2010).

There are areas of future work that can be considered. In our assessment we heuristically identified where the curve departs from linearity, suggesting when a biomarker can be used to detect an event. One can imagine setting that time point as an estimable parameter, through which one can draw proper inference. Additional work can also consider the situation where the collection of a biomarker is informative, for example, a blood test when a patient is not feeling well. This is essentially a missing data problem, where the lack of missing data is itself informative. In our setting, since labs are taken per-protocol we implicitly assume that any data is missing completely at random. Finally, more work is needed that estimate the prospective model (2), particularly with multiple markers. While our approach presents a heuristic for detecting markers, the gold-standard definition of clinical utility is through clinical metrics such as risk-difference (Wentzensen and Wacholder, 2013).

Overall, we consider this a useful screening procedure to select markers to track either quantitatively through algorithms embedded in the EHR predicting the probability of an event or more qualitatively through clinical observation. The appeal of this procedure is its simplicity and intuitiveness and the use of standard statistical methodology and theory. We believe this approach is easily transferable to analyzing other types of serially collected EHR data that may be changing over time.

Acknowledgements

We thank the referee, associate editor and editor for their valuable comments. Dr. Goldstein is funded by National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) career development award K25 DK097279. Dr. Assimes is funded by NIDDK K23 DK088942. Dr. Hastie is partially supported by grant DMS-1407548 from the National Science Foundation, and grant RO1-EB001988-15 from the National Institutes of Health. This work was supported by grants R01DK090181 and R01DK095024 to Dr. Winkelmayer. Dr. Winkelmayer enjoys salary and research support through the endowed Gordon A. Cain Chair in Nephrology at Baylor College of Medicine

References

Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case cohort designs. Journal of Clinical Epidemiology. 1999; 52:1165–1172. [PubMed: 10580779]

Bassand JP, Afzal R, Eikelboom J, Wallentin L, Peters R, Budaj A, Fox KA, Joyner CD, Chrolavicius S, Granger CB, Mehta S, Yusuf S. OASIS 5 and OASIS 6 Investigators. Relationship between baseline haemoglobin and major bleeding complications in acute coronary syndromes. European Heart Journal. 2010; 31:50–58. [PubMed: 19825809]

Cai T, Zheng Y. Evaluating prognostic accuracy of biomarkers in nested case-control studies. Biostatistics. 2012; 13:89–100. [PubMed: 21856652]

Djouss L, Rothman KJ, Cupples LA, Levy D, Ellison RC. Serum albumin and risk of myocardial infarction and all-cause mortality in the Framingham Offspring Study. Circulation. 2002; 106:2919–2924. [PubMed: 12460872]

Eagle KA, Ginsburg GS, Musunuru K, Aird WC, Balaban RS, Bennett SK, Blumenthal RS, Coughlin SR, Davidson KW, Frohlich ED, Greenland P, Jarvik GP, Libby P, Pepine CJ, Ruskin JN, Stillman AE, Van Eyk JE, Tolunay HE, McDonald CL, Smith SC Jr. Identifying patients at high risk of a cardiovascular event in the near future: Current status and future directions: Report of a National Heart, Lung, and Blood Institute working group. Circulation. 2010; 121:1447–1454. [PubMed: 20351302]

Ensrud K, Grimm RH Jr. The white blood cell count and risk for coronary heart disease. American Heart Journal. 1992; 124:207–213. [PubMed: 1615807]

Friedman GD, Klatsky AL, Siegelaub AB. The leukocyte count as a predictor of myocardial infarction. New England Journal of Medicine. 1974; 290:1275–1278. [PubMed: 4827627]

Gawaz M. Role of platelets in coronary thrombosis and reperfusion of ischemic myocardium. Cardiovascular Research. 2004; 61:498–511. [PubMed: 14962480]

Gertheiss J, Goldsmith J, Crainiceanu C, Greven S. Longitudinal scalar-on-functions regression with application to tractography data. Biostatistics. 2013; 14:447–461. [PubMed: 23292804]

Goldsmith J, Crainiceanu CM, Caffo B, Reich D. Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. Journal of the Royal Statistical Society, Series C. 2012; 61:453–469.

Goldstein BA, Chang TI, Mitani AA, Assimes TL, Winkelmayer WC. Near-term prediction of sudden cardiac death in older hemodialysis patients using electronic health records. Clinical Journal of the American Society of Nephrology. 2014; 9:82–91. [PubMed: 24178968]

Herzog CA. Acute myocardial infarction in patients with end-stage renal disease. Kidney International Supplements. 1999; 71:S130–S133.

Ikizler TA. The use and misuse of serum albumin as a nutritional marker in kidney disease. Clinical Journal of the American Society of Nephrology. 2012; 7:1375–1377. [PubMed: 22904120]

Irizarry MC, Webb DJ, Boudiaf N, Logie J, Habel LA, Udaltsova N, Friedman GD. Risk of cancer in patients exposed to gabapentin in two electronic medical record systems. Pharmacoepidemiolgy Drug Safety. 2012; 21:214–225.

James G. Generalized linear models with functional predictor variables. Journal of the Royal Statistical Society, Series B. 2002; 64:411–432.

Lubin JH, Gail MH. Biased selection of controls for case-control analyses of cohort studies. Biometrics. 1984; 40:63–75. [PubMed: 6375751]

Petersen LA, Wright S, Normand SL, Daley J. Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database. Journal of General Intern Medicine. 1999; 14:555–558.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. Available at: www.R-project.org.

Robins JM, Gail MH, Lubin JH. More on biased selection of controls for case-control analyses of cohort studies. Biometrics. 1986; 42:293–299. [PubMed: 3741971]

U.S. Renal Data System. USRDS 2013 Annual Data Report: Atlas of chronic kidney disease and end-stage renal disease in the United States. Bethesda, MD: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 2013.

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. III. Design options. American Journal of Epidemiology. 1992; 135:1042–1050. [PubMed: 1595690]

Wentzensen N, Wacholder S. From differences in means between cases and controls to risk stratification: A business plan for biomarker development. Cancer Discovery. 2013; 3:148–157. [PubMed: 23299199]

Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. Medical Care. 2010; 48:S106–S113. [PubMed: 20473190]

Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association. 2005; 100:577–590.

Appendix

The following is R code for calculating a truncated power spline basis:

```
tps <- function(X, knots){
     k <- length(knots)
     b <- matrix(NA, nrow = length(X), ncol = k + 1)
     b[,1] <- X ###Add X to basis; no intercept
     for(i in 1:k){
                tp <- (X - knots[i])^3 ###Cubic polynomial
                tp <- ifelse(tp > 0, tp, 0) ###Truncate
                b[,(i+1)] <- tp
         }
     return(b)
}
```

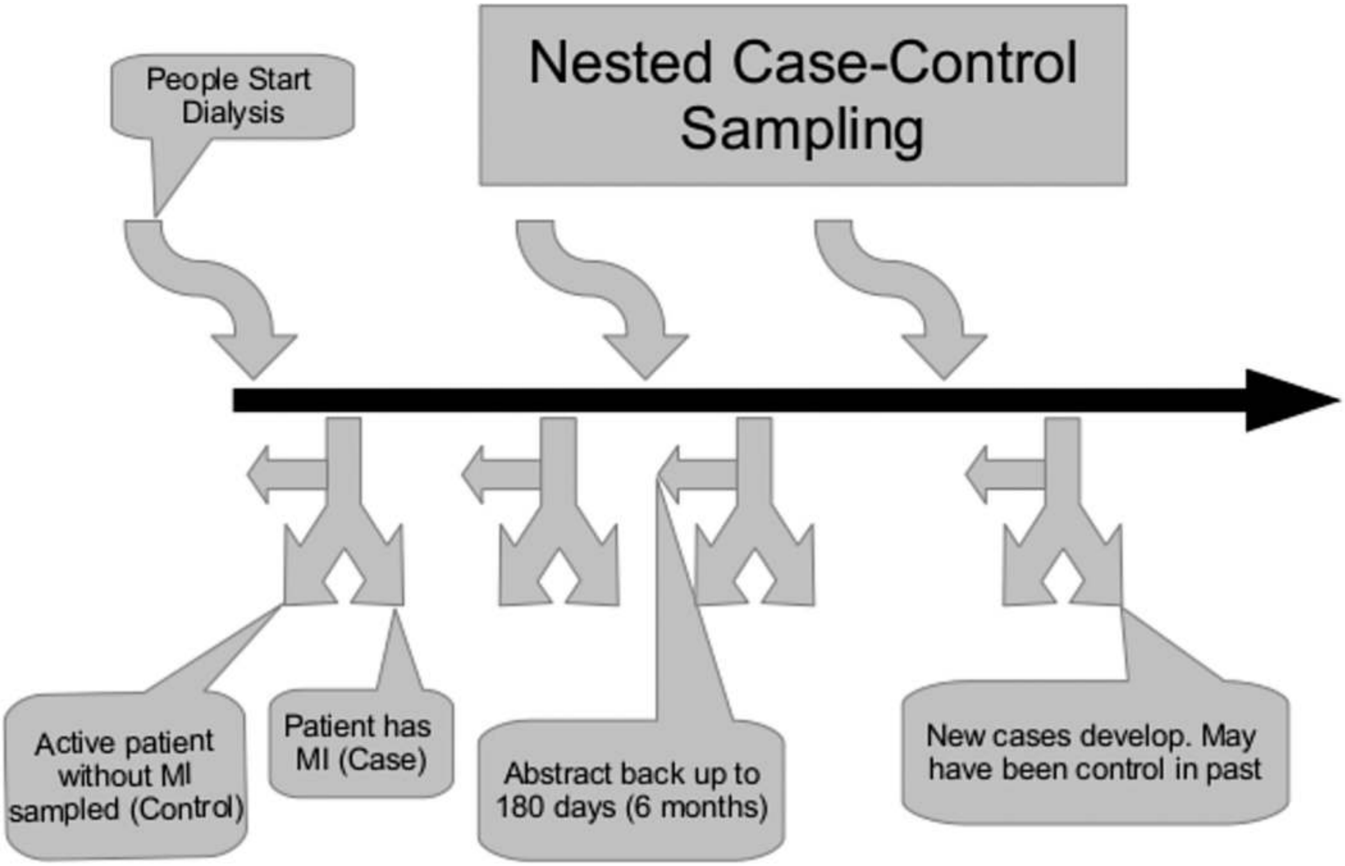
Nested Case-Control Sampling
People Start Dialysis
Active patient without MI sampled (Control)
Patient has MI (Case)
Abstract back up to 180 days (6 months)
New cases develop. May have been control in past

Figure 1.

Illustration of Nested Case-Control sampling design. Patients start dialysis at different time points. Whenever a case develops, an active control is sampled. Data are abstracted back up to 6 months. In the future these controls *may* become cases or serve as controls again.

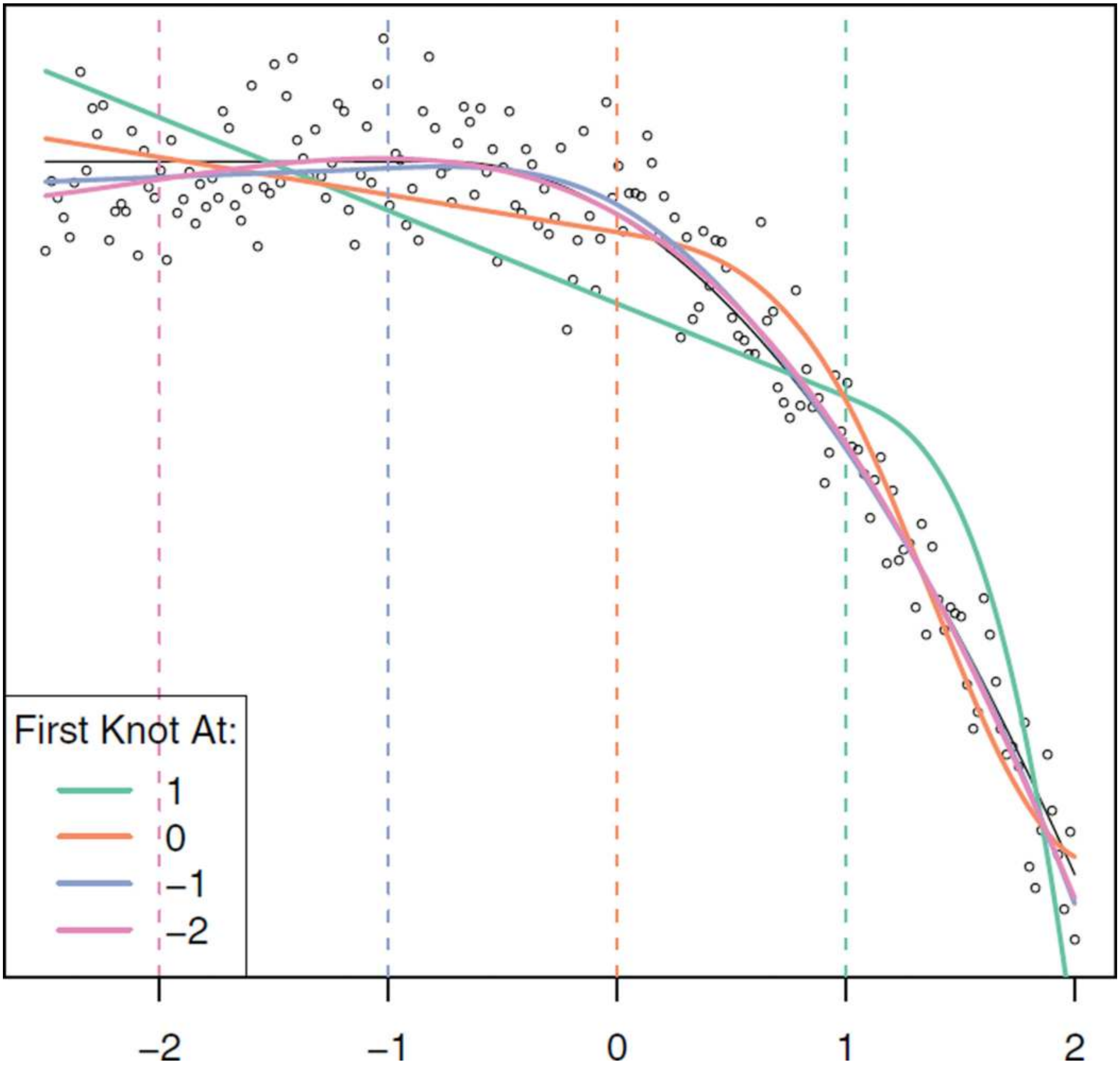

Figure 2.

Truncated power splines with different knot placements. The dashed line corresponds to the placement of the first (left-most) knot for the same colored line. We note that to the left of the first knot, the fit is linear. The black line shows the true function, with the dots the realized data. We note that adding the 4th knot at −2 does not change the functional fit. “This figure appears in color in the electronic version of this article.”

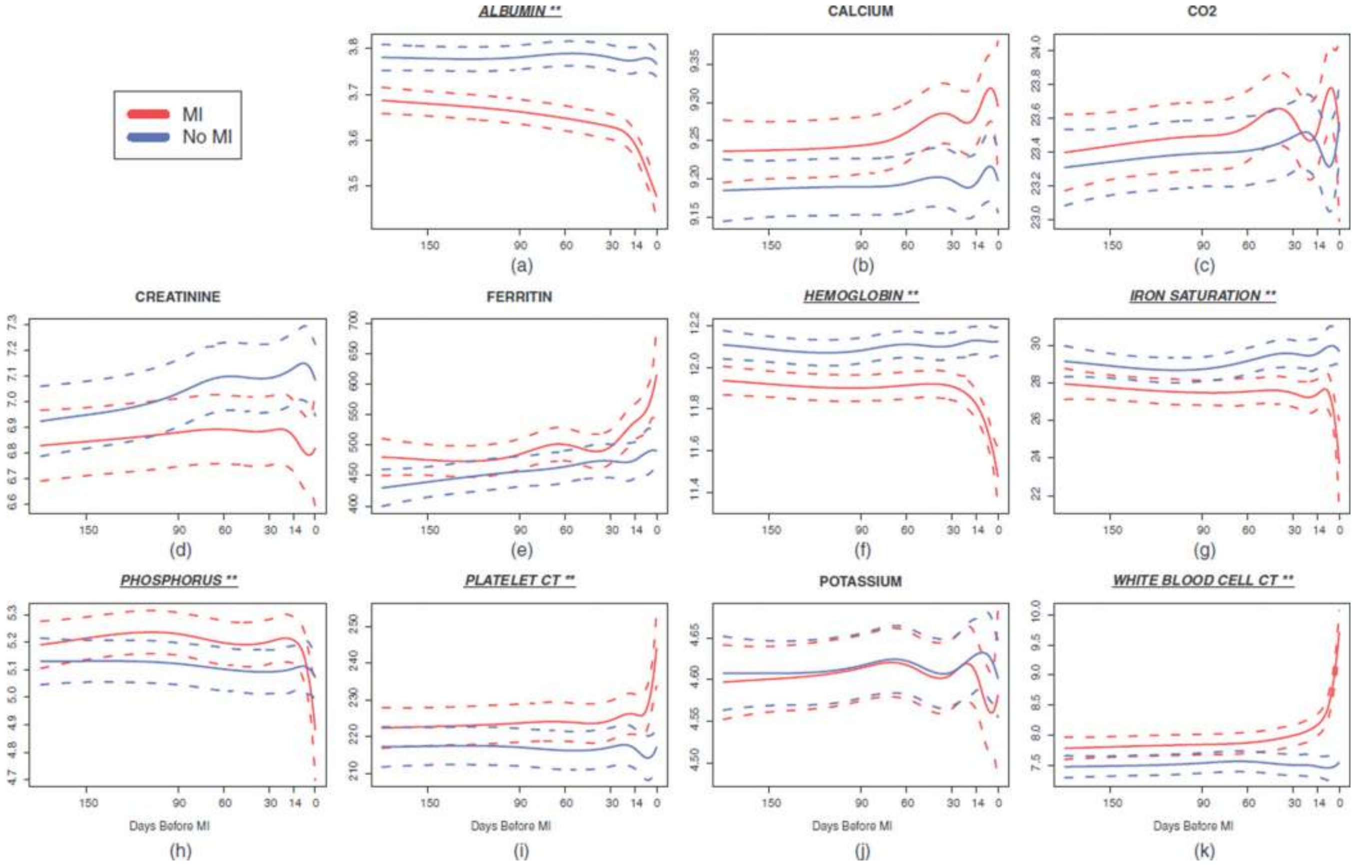

Figure 3.

Trajectory of laboratory measures preceding an MI.

**Six of the 11 markers (albumin [a], hemoglobin [f], iron saturation [g], iron saturation [h], platelet-count [i], and white blood cell count [k]) show clinically meaningful changes before an MI. “This figure appears in color in the electronic version of this article.”

Table 1

Frequency of collection and acceptable ranges for laboratory tests assessed

Laboratory test	Frequency collected	Acceptable range
Albumin	~30 days	0.1–6 g/dL
Calcium	~7 days	5–20 mg/dL
CO_2	~30 days	2–50 meq/L
Creatinine	~30 days	0.1–30 mg/dL
Ferritin	~90 days	0–10,000 ng/mL
Hemoglobin	~7 days	2–20 g/dL
Iron transferring saturation	~30 days	0–100%
Phosphorous	~7 days	0.5–20 mg/dL
Platelet count	~30 days	0–5000 1000/μL
Potassium	~30 days	1–9 meq/L
White blood cell count	~30 days	0–100 1000/μL

Table 2

Demographics of sampled data

	MI	No MI	P-value
Sample size	4769	4769	
Age at start of dialysis	75 (71, 80)	75 (71,79)	<0.001
Gender (male)	2319 (50%)	2380 (50%)	0.22
Race			0.022
Caucasian	3414 (72%)	3311 (69%)	
African American	1152 (24%)	1285 (27%)	
Hispanic	136 (3%)	115 (2%)	
Asian	50 (1%)	41 (1%)	
Other/unknown	17 (<1%)	17 (<1%)	
Days on dialysis	533 (188, 1088)	553 (245, 1081)	<0.001

Table 3

The first three columns show Bonferroni corrected p-values (across three tests) for each of the tested metrics. For those labs that met the above criteria we assessed at what point the cases differentiated themselves from the controls. Using the validation data the c-statistic for predicting MI.

Lab test[a]	LRT Overall	LRT Among cases	LRT Among controls	Optimal fit	Misclassification Rate[b]	*c*-Statistic
Albumin[a]	<0.001	<0.001	0.084	28 days	0.384[b]	0.635[b]
Calcium	0.072	0.210	1.000	—	0.468	0.551
CO_2	0.964	1.000	1.000	—	0.474	0.547
Creatinine	<0.001	0.187	0.303	—	0.473	0.546
Ferritin	0.156	0.003	1.000	—	0.468	0.550
Hemoglobin[a]	<0.001	<0.001	0.425	28 days	0.430[b]	0.591[b]
Iron transferring saturation[a]	<0.001	0.024	0.703	168 days	0.459[b]	0.572[b]
Phosphorous[a]	0.033	<0.001	1.000	14 days	0.459	0.546
Platelet count[a]	<0.001	0.012	1.000	14 days	0.470	0.556
Potassium	1.000	0.240	1.000	—	0.475	0.541
White blood cell count[a]	<0.001	<0.001	0.933	56 days	0.434[b]	0.588[b]

[a] Met all three analytic criterion.

[b] Significant (p <0.05) improvement over model with just demographic factors (base misclassification rate=0.470; *c*=0.541).