

BAYESIAN HIERARCHICAL RULE MODELING FOR PREDICTING MEDICAL CONDITIONS

BY TYLER H. MCCORMICK^{*}, CYNTHIA RUDIN[†] AND DAVID MADIGAN[‡]

University of Washington^{}, Massachusetts Institute of Technology[†] and
Columbia University[‡]*

We propose a statistical modeling technique, called the Hierarchical Association Rule Model (HARM), that predicts a patient’s possible future medical conditions given the patient’s current and past history of reported conditions. The core of our technique is a Bayesian hierarchical model for selecting predictive association rules (such as “*condition 1 and condition 2* \rightarrow *condition 3*”) from a large set of candidate rules. Because this method “borrows strength” using the conditions of many similar patients, it is able to provide predictions specialized to any given patient, even when little information about the patient’s history of conditions is available.

1. Introduction. The emergence of large-scale medical record databases presents exciting opportunities for data-based personalized medicine. Prediction lies at the heart of personalized medicine and in this paper we propose a statistical model for predicting patient-level sequences of medical conditions. We draw on new approaches for predicting the next event within a “current sequence,” given a “sequence database” of past event sequences (Rudin *et al.*, 2011a,b). Specifically we propose the Hierarchical Association Rule Model (HARM) that generates a set of *association rules* such as *dyspepsia and epigastric pain* \rightarrow *heartburn*, indicating that dyspepsia and epigastric pain are commonly followed by heartburn. HARM produces a ranked list of these association rules. Both patients and caregivers can use the rules to guide medical decisions. Built-in explanations represent a particular advantage of the association rule framework—the rule predicts heartburn *because* the patient has had dyspepsia and epigastric pain.

In our setup, we assume that each patient visits a healthcare provider periodically. At each encounter, the provider records time-stamped medical conditions experienced since the previous encounter. In this context, we address several prediction problems such as:

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35

Keywords and phrases: Association rule mining, healthcare surveillance, hierarchical model, machine learning

- Given data from a sequence of past encounters, predict the next condition that a patient will report.
- Given basic demographic information, predict the first condition that a patient will report.
- Given partial data from an encounter (and possibly prior encounters) predict the next condition.

Though medical databases often contain records from thousands or even millions of patients, most patients experience only a handful of the massive set of potential conditions. This patient-level sparsity presents a challenge for predictive modeling. Our hierarchical modeling approach attempts to address this challenge by borrowing strength across patients.

The sequential event prediction problem is new a supervised learning problem that has been formalized here and by [Rudin *et al.* \(2011a,b\)](#). [DuMouchel and Pregibon \(2001\)](#) presented a Bayesian analysis of association rules. Their approach, however, does not apply in our context because of the sequential nature of our data. Rules are particularly useful in our context: rules yield very interpretable models, and their conditional probabilities involve few variables and are thus more reliable to estimate.

The experiments this paper presents indicate that HARM outperforms several baseline approaches including a standard “maximum confidence, minimum support threshold” technique used in association rule mining, and also a non-hierarchical version of our Bayesian method (from [Rudin *et al.*, 2011a,b](#)) that ranks rules using “adjusted confidence.”

More generally, HARM yields a prediction algorithm for sequential data that can potentially be used for a wide variety of applications beyond condition prediction. For instance, the algorithm can be directly used as a recommender system (for instance, for vendors such as Netflix, amazon.com, or online grocery stores such as Fresh Direct and Peapod). It can be used to predict the next move in a video game in order to design a more interesting game, or it can be used to predict the winners at each round of a tournament (e.g., the winners of games in a football season). All of these applications possess the same basic structure as the condition prediction problem: a database consisting of sequences of events, where each event is associated to an individual entity (medical patient, customer, football team). As future events unfold in a new sequence, our goal is to predict the next event.

In [Section 2](#) we provide basic definitions and present our model. In [Section 3](#) we evaluate the predictive performance of HARM, along with several baselines through experiments on clinical trial data. [Section 4](#) provides related work, and [Section 5](#) provides a discussion and offers potential exten-

sions.

2. Method. This work presents a new approach to association rule mining by determining the “interestingness” of rules using a particular (hierarchical) Bayesian estimate of the probability of exhibiting condition b , given a set of current conditions, a . We will first discuss association rule mining and its connection to Bayesian shrinkage estimators. Then we will present our hierarchical method for providing personalized condition predictions.

2.1. *Definitions.* An *association rule* in our context is an implication $a \rightarrow b$ where the left side is a subset of conditions that the patient has experienced, and b is a single condition that the patient has not yet experienced since the last encounter. Ultimately, we would like to rank rules in terms of “interestingness” or relevance for a particular patient at a given time. Using this ranking, we make predictions of subsequent conditions. Two common determining factors of the “interestingness” of a rule are the “confidence” and “support” of the rule (Agrawal, Imieliński and Swami, 1993; Piatetsky-Shapiro, 1991).

The confidence of a rule $a \rightarrow b$ for a patient is the empirical probability:

$$\begin{aligned} \text{Conf}(a \rightarrow b) &:= \frac{\text{Number of times conditions } a \text{ and } b \text{ were experienced}}{\text{Number of times conditions } a \text{ were experienced}} \\ &:= \hat{P}(b|a). \end{aligned}$$

The support of set a is:

$$\begin{aligned} \text{Support}(a) &:= \text{Number of times conditions } a \text{ were experienced} \\ &\propto \hat{P}(a), \end{aligned}$$

where $\hat{P}(a)$ is the empirical proportion of times that conditions a were experienced. When a patient has experienced a particular set of conditions only a few times, a new single observation can dramatically alter the confidence $\hat{P}(b|a)$ for many rules. This problem occurs commonly in our clinical trial data, where most patients have reported fewer than 10 total conditions. The vast majority of rule mining algorithms address this issue with a minimum support threshold to exclude rare rules, and the remaining rules are evaluated for interestingness (reviews of interestingness measures include those of Tan, Kumar and Srivastava, 2002; Geng and Hamilton, 2007). The definition of interestingness is often heuristic, and is not necessarily a meaningful estimate of $P(b|a)$.

It is well-known that problems arise from using a minimum support threshold. For instance, consider the collection of rules meeting the minimum support threshold condition. Within this collection, the confidence

alone should not be used to rank rules: among rules with similar confidence, the rules with larger support should be preferred. More importantly, “nuggets,” which are rules with low support but very high confidence, are often excluded by the threshold. This is problematic, for instance, when a condition that occurs rarely is strongly linked with another rare condition, it is essential not to exclude the rules characterizing these conditions. In our data, the distribution of conditions has a long tail, where the vast majority of events happen rarely: out of 1800 possible conditions, 1400 occur less than 10 times. These 1400 conditions are precisely the ones in danger of being excluded by a minimum support threshold.

Our work avoids problems with the minimum support threshold by ranking rules with a shrinkage estimator of $P(b|a)$. These estimators directly incorporate the support of the rule. One example of such an estimator is the “adjusted confidence” (Rudin *et al.*, 2011a,b):

$$\text{AdjConf}(a \rightarrow b, K) := \frac{\text{Number of times conditions } a \text{ and } b \text{ were experienced}}{\text{Number of times conditions } a \text{ were experienced} + K}.$$

The effect of the penalty term K is to pull low-support rules towards the bottom of the list; any rule achieving a high adjusted confidence must overcome this pull through either a high enough support or a high confidence. Using the adjusted confidence avoids the problems discussed earlier: “interestingness” is closely related to the conditional probability $P(b|a)$, and among rules with equal confidence the higher support rules are preferred, and there is no strict minimum support threshold.

In this work, we extend the adjusted confidence model in an important respect, in that our method shares information across similar patients to better estimate the conditional probabilities. The adjusted confidence is a particular Bayesian estimate of the confidence. Assuming a Beta prior distribution for the confidence, the posterior mean is:

$$\tilde{P}(b|a) := \frac{\alpha + \#(a\&b)}{\alpha + \beta + \#a},$$

where $\#x$ is the support of condition x , and α and β denote the parameters of the (conjugate) Beta prior distribution. Our model allows the parameters of the Binomial to be chosen differently for each patient and also for each rule. This means that our model can determine, for instance, whether a particular patient is more likely to repeat a condition that has occurred only once, and also whether a particular condition is more likely to repeat than another.

We note that our approach makes no explicit attempt to infer causal relationships between conditions. The observed associations may in fact arise

from common prior causes such as other conditions or drugs. Thus a rule such as *dyspepsia* \rightarrow *heartburn* does not necessarily imply that successful treatment of dyspepsia will change the probability of heartburn. Rather the goal is to accurately predict heartburn in order to facilitate effective medical management.

2.2. *Hierarchical Association Rule Model (HARM)*. For a patient i and a given rule, r , say we observe y_{ir} co-occurrences (support for lhs \cup rhs), where there were a total of n_{ir} encounters that include the lhs (n_{ir} is the support for lhs). We model the number of co-occurrences as Binomial(n_{ir}, p_{ir}) and then model p_{ir} hierarchically to share information across groups of similar individuals. Define \mathbf{M} as a $I \times D$ matrix of static observable characteristics for a total of I individuals and D observable characteristics, where we assume $D > 1$ (otherwise we revert back to a model with a rule-wise adjustment). Each row of \mathbf{M} corresponds to a patient and each column to a particular characteristic. We define the columns of \mathbf{M} to be indicators of particular patient categories (gender, or age between 30 and 40, for example), though they could be continuous in other applications. Let \mathbf{M}_i denote the i^{th} row of the matrix \mathbf{M} . We model the probability for the i^{th} individual and the r^{th} rule p_{ir} as coming from a Beta distribution with parameters π_{ir} and τ_i . We then define π_{ir} through the regression model $\pi_{ir} = \exp(\mathbf{M}'_i \boldsymbol{\beta}_r + \gamma_i)$ where $\boldsymbol{\beta}_r$ defines a vector of regression coefficients for rule r and γ_i is an individual-specific random effect. More formally, we propose the following model:

$$\begin{aligned} y_{ir} &\sim \text{Binomial}(n_{ir}, p_{ir}) \\ p_{ir} &\sim \text{Beta}(\pi_{ir}, \tau_i) \\ \pi_{ir} &= \exp(\mathbf{M}'_i \boldsymbol{\beta}_r + \gamma_i). \end{aligned}$$

Under this model,

$$E(p_{ir} | y_{ir}, n_{ir}) = \frac{y_{ir} + \pi_{ir}}{n_{ir} + \pi_{ir} + \tau_i},$$

which is a more flexible form of adjusted confidence. This expectation also produces non-zero probabilities for a rule even if n_{ir} is zero (patient i has never reported the conditions on the left hand side of r before). This could allow rules to be ranked more highly even if n_{ir} is zero. The fixed effect regression component, $\mathbf{M}'_i \boldsymbol{\beta}_r$, adjusts π_{ir} based on the patient characteristics in the \mathbf{M} matrix. For example, if the entries of \mathbf{M} represented only gender, then the regression model with intercept $\beta_{r,0}$ would be $\beta_{r,0} + \beta_{r,1} \mathbf{1}_{\text{male}}$ where $\mathbf{1}_{\text{male}}$ is one for male respondents and zero for females. Being male, therefore,

has a multiplicative effect of $e^{\beta_{r,1}}$ on π_{ir} . In this example, the $\mathbf{M}'_i\boldsymbol{\beta}_r$ value is the same for all males, encouraging similar individuals to have similar values of π_{ir} . For each rule r , we will use a common prior on all coefficients in $\boldsymbol{\beta}_r$; this imposes a hierarchical structure, and has the effect of regularizing coefficients associated with rare characteristics.

The π_{ir} 's allow rare but important “nuggets” to be recommended. Even across multiple patient encounters, many conditions occur very infrequently. In some cases these conditions may still be highly associated with certain other conditions. For instance, compared to some conditions, migraines are relatively rare. Patients who have migraines however typically also experience nausea. A minimum support threshold algorithm might easily exclude the rule “migraines \rightarrow nausea” if a patient hasn’t experienced many migraines in the past. This is especially likely for patients who have few encounters. In our model, the π_{ir} term balances the regularization imposed by τ_i to, for certain individuals, increase the ranking of rules with high confidence but low support. The τ_i term reduces the probability associated with rules that have appeared few times in the data (low support), with the same effect as the penalty term (K) in the adjusted confidence. Unlike the cross-validation or heuristic strategies suggested in [Rudin *et al.* \(2011a,b\)](#), we estimate τ_i as part of an underlying statistical model. Within a given rule, we assume τ_i for every individual comes from the same distribution. This imposes additional structure across individuals, increasing stability for individuals with few observations.

It remains now to describe the precise prior structure on the regression parameters and hyperparameters. We assign Gaussian priors with mean 0 and variance σ_τ^2 to the τ on the log scale. Since any given patient is unlikely to experience a specific medical condition, the majority of probabilities are close to zero. Giving τ_i a prior with mean zero improves stability by discouraging excessive penalties. We assign all elements $\beta_{r,d}$ of vectors $\boldsymbol{\beta}_r$ a common Gaussian prior on the log scale with mean μ_β and variance σ_β^2 . We also assume each γ_i comes from a Gaussian distribution on the log scale with common mean μ_γ and variance σ_γ^2 . Each individual has their own γ_i term, which permits flexibility among individuals; however, all of the γ_i terms come from the same distribution, which induces dependence between individuals. We assume diffuse uniform priors on the hyperparameters σ_τ^2 , μ_β , and σ_β^2 . Denote \mathbf{Y} as the matrix of y_{ir} values, \mathbf{N} as the matrix of n_{ir} values, and $\boldsymbol{\beta}$ as the collection of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_R$. The prior assumptions yield the following

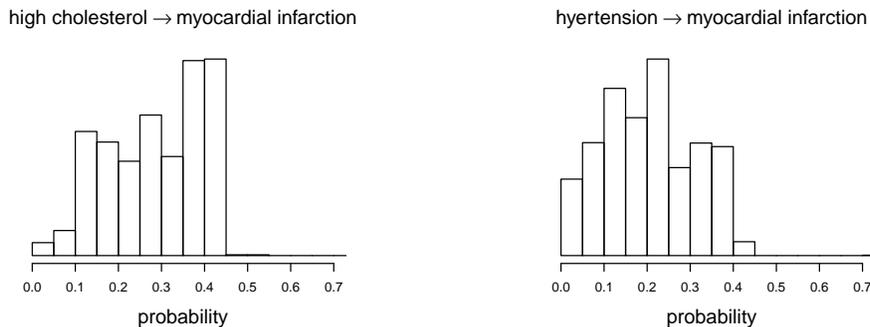


FIG 1. Approximate posterior of two rules. These are histograms of the posterior means for the set of patients.

posterior:

$$\begin{aligned}
 p, \pi, \tau, \beta | \mathbf{Y}, \mathbf{N}, \mathbf{M} &\propto \prod_{i=1}^I \prod_{r=1}^R p_{ir}^{y_{ir} + \pi_{ir}} (1 - p_{ir})^{n_{ir} - y_{ir} + \tau_i} \\
 &\times \prod_{r=1}^R \prod_{d=1}^D \text{Normal}(\log(\beta_{r,d}) | \mu_{\beta}, \sigma_{\beta}^2) \\
 &\times \prod_{i=1}^I \text{Normal}(\log(\gamma_i) | \mu_{\gamma}, \sigma_{\gamma}^2) \text{Normal}(\log(\tau_i) | 0, \sigma_{\tau}^2).
 \end{aligned}$$

HARM produces draws from the (approximate) posterior distribution for each probability. Figure 1 shows estimates of the posterior probabilities for *high cholesterol* \rightarrow *myocardial infarction* and *hypertension* \rightarrow *myocardial infarction*. Comparing the distributions of related rules can often provide insights into associations in the data, as we demonstrate in Section 3.2. In the context of medical condition prediction, these probabilities are of interest and we analyze our estimates of their full posterior distributions in Section 3.2. To rank association rules for the purpose of prediction, however, we need a single estimate for each probability (rather than a full distribution), which we chose as the posterior mean. In practice, we suggest evaluating the mean as well other estimators for each probability (the mode or median, for example) and selecting the one with the best performance in each particular application. We carry out our computations using a Gibbs sampling algorithm, provided in Figure 2.

2.3. *Approximate updating.* Given a batch of data, HARM makes predictions based on the posterior distributions of the p_{ir} 's. Since the posteriors

For a suitably initialized chain, at step v :

1. Update p_{ir} from the conjugate Beta distribution given $\pi_{ir}, \tau_i, \mathbf{Y}, \mathbf{N}, \mathbf{M}$.
2. Update τ_i using a Metropolis step with proposal τ_i^* where

$$\log(\tau_i^*) \sim N(\tau_i^{(v-1)}), \text{ (scale of jumping dist)}).$$

3. For each rule, update the vector β_r using a Metropolis step with

$$\log(\beta_r^*) \sim N(\beta_r^{(v-1)}), \text{ (scale of jumping dist)}).$$

4. Update γ_i using a Metropolis step with

$$\log(\gamma_i^*) \sim N(\gamma_i^{(v-1)}), \text{ (scale of jumping dist)}).$$

5. Update $\pi_{ir} = \exp(\mathbf{M}_i^t \beta_r + \gamma_i)$.
6. Update $\mu_\beta \sim N(\hat{\mu}_\beta, \sigma_\beta^2)$ where

$$\hat{\mu}_\beta = \left(\frac{1}{D+R} \right) \sum_{r=1}^R \sum_{d=1}^D \beta_{r,d}.$$

7. Update $\sigma_\beta^2 \sim \text{Inv-}\chi^2(d-1, \hat{\sigma}_\beta^2)$ where

$$\hat{\sigma}_\beta^2 = \left(\frac{1}{D+R-1} \right) \sum_{r=1}^R \sum_{d=1}^D (\beta_{r,d} - \mu_\beta)^2.$$

8. Update $\sigma_\tau^2 \sim \text{Inv-}\chi^2(I-1, \hat{\sigma}_\tau^2)$ where $\hat{\sigma}_\tau^2 = \frac{1}{I-1} \sum_{i=1}^I (\tau_i - \mu_\tau)^2$.

9. Update $\mu_\gamma \sim N(\hat{\mu}_\gamma, \sigma_\gamma^2)$ where $\hat{\mu}_\gamma = \frac{1}{I} \sum_{i=1}^I \gamma_i$.

10. Update $\sigma_\gamma^2 \sim \text{Inv-}\chi^2(I-1, \hat{\sigma}_\gamma^2)$ where $\hat{\sigma}_\gamma^2 = \frac{1}{I-1} \sum_{i=1}^I (\gamma_i - \mu_\gamma)^2$.

FIG 2. *Gibbs sampling algorithm for hierarchical bayesian association rule modeling for sequential event prediction (HARM).*

are not available in closed form, we need to iterate the algorithm in Figure 2 to convergence in order to make predictions. Each time the patient visits the physician, each p_{ir} could be updated by again iterating the algorithm in Figure 2 to convergence. In some applications, new data continue arrive frequently, making it impractical to compute approximate posterior distributions using the algorithm in Figure 2 for each new encounter. In this section we provide an approximate updating scheme to incorporate new patient data after an initial batch of encounters has already been processed. The approximate scheme can be used for real-time online updating.

Beginning with an initial batch of data, we run the algorithm in Figure 2 to convergence in order to obtain $\hat{\tau}_i$ and $\hat{\pi}_{ir}$, which are defined to be the posterior mean of the estimated distributions for τ_i and π_{ir} . The approximate updating scheme keeps τ_i and π_{ir} fixed to be $\hat{\tau}_i$ and $\hat{\pi}_{ir}$. Given that up to encounter $e - 1$, we have observed $y_{ir}^{(e-1)}$ and $n_{ir}^{(e-1)}$, we are presented with new observations that have counts $y_{ir}^{(\text{newobs.})}$ and $n_{ir}^{(\text{newobs.})}$ so that $y_{ir}^{(e)} = y_{ir}^{(e-1)} + y_{ir}^{(\text{newobs.})}$ and $n_{ir}^{(e)} = n_{ir}^{(e-1)} + n_{ir}^{(\text{newobs.})}$. In order to update the probability estimates to reflect our total current data, $y_{ir}^{(e)}$, $n_{ir}^{(e)}$, we will use the following relationship:

$$P(p_{ir} | y_{ir}^{(e)}, n_{ir}^{(e)}, \hat{\tau}_i, \hat{\pi}_{ir}) \propto P(y_{ir}^{(\text{newobs.})} | n_{ir}^{(\text{newobs.})}, p_{ir}) \times P(p_{ir} | y_{ir}^{(e-1)}, n_{ir}^{(e-1)}, \hat{\tau}_i, \hat{\pi}_{ir}).$$

The expression $P(p_{ir} | y_{ir}^{(e-1)}, n_{ir}^{(e-1)}, \hat{\tau}_i, \hat{\pi}_{ir})$ is the posterior up to encounter $e - 1$ and has a Beta distribution. The likelihood of the new observations, $P(y_{ir}^{(\text{newobs.})} | n_{ir}^{(\text{newobs.})}, p_{ir})$, is Binomial. Conjugacy implies that the updated posterior also has a Beta distribution. In order to update the probability estimates for our hierarchical model, we use the expectation of this distribution, that is

$$E(p_{ir} | y_{ir}^{(e)}, n_{ir}^{(e)}, \hat{\tau}_i, \hat{\pi}_{ir}) = \frac{y_{ir}^{(e-1)} + y_{ir}^{\text{newobs.}} + \hat{\pi}_{ir}}{n_{ir}^{(e-1)} + n_{ir}^{\text{newobs.}} + \hat{\pi}_{ir} + \hat{\tau}_i}.$$

3. Application to repeated patient encounters. We present results of HARM, with the approximate updating scheme in Section 2.3, on co-prescribing data from a large clinical trial. In the trial, each patient visits a healthcare provider periodically. At each encounter, the provider records time-stamped medical conditions (represented by MedDRA terms) experienced since the previous encounter. Thus, each encounter is associated with a sequence of medical conditions. These data are from around 42,000 patient encounters from about 2,300 patients, all at least 40 years old. The matrix of

observable characteristics encodes the basic demographic information: gender, age group (40-49, etc.), ethnicity. For each patient we have a record of each medication prescribed and the condition/chief complaint (back pain, asthma, etc) that warranted the prescription. We chose to predict patient complaints rather than prescriptions since there are often multiple prescribing options (medications) for the same complaint. Some patients had pre-existing conditions that continued throughout the trial. For these patients, we include these pre-existing conditions in the patient's list of conditions at each encounter. Other patients have recurrent conditions for which we would like to predict the occurrences. If a patient reports the same condition more than once during the same thirty day period we only consider the first occurrence of the condition at the first report. If the patient reports the condition once and then again more than thirty days later, we consider this two separate incidents.

As covariates, we used age, gender, race and drug/placebo (an indicator of whether the patient was in the treatment or control group for the clinical trial). We fit age using a series of indicator variables corresponding to four groups (40-49, 50-59, 60-69, 70+). We included all available covariates in our simulation studies. In practice, model selection will likely be essential to select the best subset of covariates for predictive performance. We discuss covariate selection in further detail in the supplement article ([McCormick, Rudin and Madigan, 2011](#)).

Our experiments consider only the marginal probabilities (support) and probabilities conditional on one previous condition. Thus, the left hand side of each rule contains either 0 items or 1 item. In our simulations, we used chains of 5,000 iterations keeping every 10th iteration to compute the mean we used for ranking and discarding the first thousand iterations.

In Section 3.1 we present experimental results to compare the predictive performance of our model to other rule mining algorithms for this type of problem. In Section 3.2 we use the probability estimates from the model to demonstrate its ability to find new associations; in particular, we find associations that are present in the medical literature but that may not be obvious by considering only the raw data.

3.1. Predictive performance. We selected a sample of patients by assigning each patient a random draw from a Bernoulli distribution with success probability selected to give a sample of patients on average around 200. For each patient we drew uniformly an integer t_i between 0 and the number of encounters for that patient. We ordered the encounters chronologically and used encounters 1 through t_i as our training set and the remaining encoun-

ters as the test set. Through this approach, the training set encompasses the complete set of encounters for some patients (“fully observed”), includes no encounters for others (“new patients”), and a partial encounter history of the majority of the test patients (“partially-observed patients”). We believe this to be a reasonable approximation of the context where this type of method would be applied, with some patients having already been observed several times and other new patients entering the system for the first time. We evaluated HARM’s predictive performance using a combination of common and rare conditions. For each run of the simulation, we use the 25 most popular conditions, then randomly select and additional 25 conditions for a total of 50.

The algorithm was used to iteratively predict the conditions revealed at each encounter. For each selected patient, starting with their first test encounter, and prior to that encounter’s first condition being revealed, the algorithm made a prediction of c possible conditions, where $c = 3$. Note that to predict the very first condition for a given patient when there are no previous encounters, the recommendations come from posterior means of the coefficients estimated from the training set. The algorithm earned one point if it recommended the current condition before it was revealed, and no points otherwise. Then, y_{ir} and n_{ir} were updated to include the revealed condition. This process was repeated for the patient’s remaining conditions in the first encounter, and repeated for each condition within each subsequent encounter. We then moved to the next patient and repeated the procedure.

The total score of the algorithm for a given patient was computed as the total number of points earned for that patient divided by the total number of conditions experienced by the patient. The total score of the algorithm is the average of the scores for the individual patients. Thus, the total score is the average proportion of correct predictions per patient. We repeated this entire process (beginning with selecting patients) 500 times and recorded the distribution over the 500 scores. We compared the performance of HARM (using the same scoring system) against an algorithm that ranks rules by adjusted confidence, for several values of K . We also compared with the “max confidence minimum support threshold” algorithm for different values of the support threshold θ , where rules with support below θ are excluded and the remaining rules are ranked by confidence. For both of these algorithms, no information across patients is able to be used.

Figure 3 shows the results, as boxplots of the distribution of scores for the entire collection of partially-observed, fully observed, and new patients. Paired t-tests comparing the mean proportion of correct predictions from HARM to each of the alternatives had p-values for a significant difference in

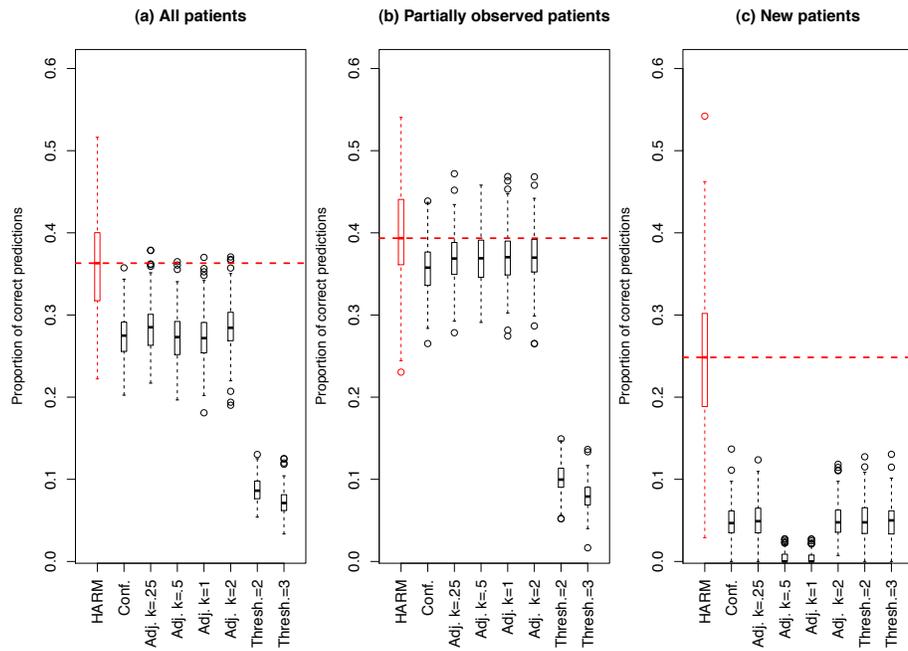


FIG 3. Predictive performance for (a) all patients, (b) partially-observed patients, (c) new patients. Each boxplot represents the distribution of scores over 500 runs. For (a), each run's score (an individual point on a boxplot) is based on a sample of approximately 200 patients. For (b) and (c), each point is based on a subset of these ~ 200 patients.

our favor less than 10^{-15} . In other words, HARM has statistically superior performance over all K and θ ; *i.e.*, better performance than either of the two algorithms even if their parameters K and θ had been tuned to the best possible value. For all four values of K for the adjusted confidence, performance was slightly better than for the plain confidence ($K = 0$). The “max confidence minimum support threshold” algorithm (which is a standard approach to association rule mining problems) performed poorly for minimum support thresholds of 2 and 3. This poor performance is likely due to the sparse information we have for each patient. Setting a minimum support threshold as low as even two eliminates many potential candidate rules from consideration.

The main advantage of our model is that it shares information across patients in the training set. This means that in early stages where the observed y_{ir} and n_{ir} are small, it may still be possible to obtain reasonably accurate probability estimates, since when patients are new, our recommendations depend heavily on the behavior of previously observed similar patients. This advantage is shown explicitly through Figures 3(b) and 3(c), which pertain to partially-observed and new patients, respectively. The advantage of HARM over the other methods is more pronounced for new patients: in cases where there are no data for each patient, there is a large advantage to sharing information. We performed additional simulations which further illustrate this point and are presented in the supplement (McCormick, Rudin and Madigan, 2011).

3.2. *Association mining.* The conditional probability estimates from our model are also a way of mining a large and highly dependent set of associations.

Ethnicity, high cholesterol or hypertension \rightarrow myocardial infarction: Figure 4(a) shows the distribution of posterior mean propensity for myocardial infarction (heart attack) given two conditions previously reported as risk factors for myocardial infarction: high cholesterol and hypertension (see Kukline, Yoon and Keenan, 2010, for a recent review). Each bar in the figure corresponds to the set of respondents in a specified ethnic group. For Caucasians, we typically estimate a higher probability of myocardial infarction in patients who have previously had high cholesterol. In African Americans / Hispanics and Asian patients, however, we estimate a generally higher probability for patients who have reported hypertension. This distinction demonstrates the flexibility of our method in combining information across respondents who are observably similar. Some other specific characteristics of the

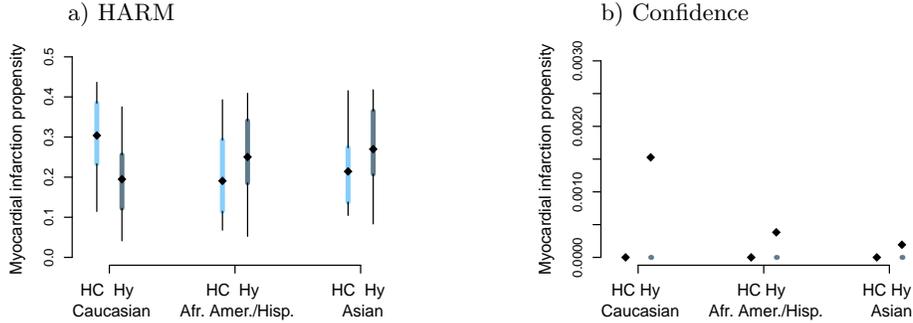


FIG 4. Propensity of myocardial infarction in patients who have reported high cholesterol or hypertension using (a) HARM and (b) (unadjusted) confidence. For each demographic group, high cholesterol (HC) is on the left and hypertension (Hy) is on the right. Thick lines represent the middle half of the posterior mean propensities for respondents in the indicated demographic group. Outer lines represent the middle 90% and dots represent the mean. The vast majority of patients did not experience a myocardial infarction, which places the middle 90% of the distribution in plot (b) approximately at zero.

estimated distributions vary with ethnicity, for instance, the propensity distribution for Caucasians who have had high cholesterol has a much longer tail than those of the other ethnic groups.

As a comparison, we also included the same plot using (unadjusted) confidence, in Figure 4(b). In both Figure 4(a) and Figure 4(b), the black dots are the mean across all the patients, which are not uniformly at zero because there were some cases of myocardial infarction and hypertension or high cholesterol. In Figure 4(b), the colored, smaller dots represent the rest of the distribution (quartiles), which all appear to be at zero in plot (b) since the vast majority of patients did not have a myocardial infarction at all, so even fewer had a myocardial infarction after reporting hypertension or high cholesterol.

Age, high cholesterol or hypertension, treatment or placebo \rightarrow myocardial infarction: Since our data come from a clinical trial, we also included an indicator of treatment vs. placebo condition in the hierarchical regression component of HARM. Figures 5 and 6 display the posterior means of propensity of myocardial infarction for respondents separated by age and treatment condition. Figure 5 considers patients who have reported hypertension, Figure 6 considers patients who have reported high cholesterol. In both Figure 5 and Figure 6, it appears that the propensity of myocardial infarction predicted by HARM is greatest for individuals between 50

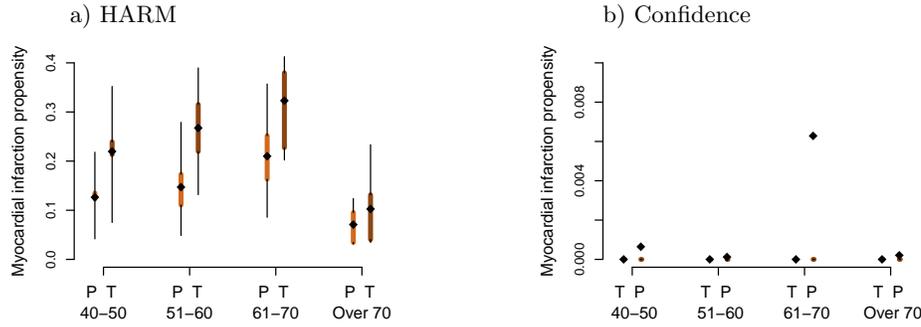


FIG 5. Propensity of myocardial infarction in patients who have reported hypertension, estimated by (a) HARM and (b) (unadjusted) confidence. For each demographic group, the placebo (P) is on the left and the treatment medication (T) is on the right. Thick lines represent the middle half of the posterior mean propensities for respondents in the indicated demographic group. Outer lines represent the middle 90% and dots represent the mean. Overall the propensity is higher for individuals who take the study medication than those who do not.

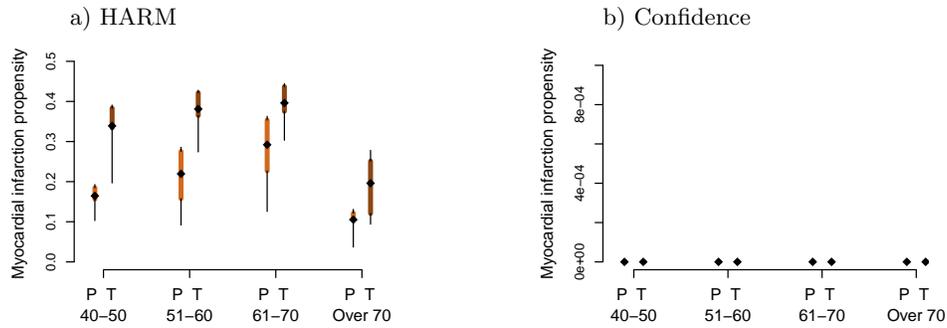


FIG 6. Propensity of myocardial infarction in patients who have reported high cholesterol, estimated by (a) HARM and (b) (unadjusted) confidence.

and 70, with the association again being stronger for high cholesterol than hypertension.

For both individuals with either high cholesterol or hypertension, use of the treatment medication was associated with increased propensity of myocardial infarction. This effect is present across nearly every age category. The distinction is perhaps most clear among patients in their fifties in both Figure 5 and Figure 6. The treatment product in this trial has been linked to increased risk of myocardial infarction in numerous other studies. The product was eventually withdrawn from the market by the manufacturer

because of its association with myocardial infarctions.

The structure imposed by our hierarchical model gives positive probabilities even when no data are present in a given category; in several of the categories, we observed no instances of a myocardial infarction, so estimates using only the data cannot differentiate between the categories in terms of risk for myocardial infarction, as particularly illustrated through Figure 6(b).

4. Related Works. Four relevant works on Bayesian hierarchical modeling and recommender systems are those of [DuMouchel and Pregibon \(2001\)](#), [Breese, Heckerman and Kadie \(1998\)](#), [Condliff, Lewis and Madigan \(1999\)](#) and [Agarwal, Zhang and Mazumder \(2011\)](#). [DuMouchel and Pregibon \(2001\)](#) deal with the identification of interesting itemsets (rather than identification of rules). Specifically, they model the ratio of observed itemset frequencies to baseline frequencies computed under a particular model for independence. Neither [Condliff, Lewis and Madigan \(1999\)](#) nor [Breese, Heckerman and Kadie \(1998\)](#) aim to model repeat purchases (recurring conditions). [Breese, Heckerman and Kadie \(1998\)](#) uses Bayesian methods to cluster users, and also suggests a Bayesian network. [Condliff, Lewis and Madigan \(1999\)](#) present a hierarchical Bayesian approach to collaborative filtering that “borrows strength” across users. [Agarwal, Zhang and Mazumder \(2011\)](#) also build a personalized recommender system that models item-item similarities. Their model uses logistic regression for estimating p_{ir} rather than using π_{ir} and τ_i . This has the advantage of being a simpler model, but loses the interpretability our model has through using association rules. It also loses the potential advantage of estimating only conditional probabilities involving few variables.

As far as we know, the line of work by [Davis *et al.* \(2010\)](#) is the first to use an approach from recommender systems to predict medical conditions, though in a completely different way than ours; it is based on vector similarity, in the same way as [Breese, Heckerman and Kadie \(1998\)](#). (Also see references in [Davis *et al.* \(2010\)](#) for background on collaborative filtering.)

5. Conclusion and Future Work. We have presented a hierarchical model for ranking association rules for sequential event prediction. The sequential nature of the data is captured through rules that are sensitive to time order, that is, $a \rightarrow b$ indicates conditions a are followed by conditions b . HARM uses information from observably similar individuals to augment the (often sparse) data on a particular individual; this is how HARM is able to estimate probabilities $P(b|a)$ before conditions a have ever been reported. In the absence of data, hierarchical modeling provides structure. As

more data become available, the influence of the modeling choices fade as greater weight is placed on the data. The sequential prediction approach is especially well suited to medical condition prediction, where experiencing two conditions in succession may have different clinical implications than experiencing either condition in isolation.

Model selection is important for using our method in practice. There are two types of model selection required for HARM: the choice of covariates encoded by the matrix \mathbf{M} , and the collection of available rules. For the choice of covariates in \mathbf{M} , standard feature selection methods can be used, for instance, a forward stagewise procedure where one covariate at a time is added as performance improves, or a backward stagewise method where features are iteratively removed. Another possibility is to combine covariates, potentially through a method similar to model-based clustering (Fraley and Raftery, 2002). To perform model selection on the choice of rules, it is possible to construct analogous “rule selection” methods as one might use for a set of covariates. A forward stagewise procedure could be constructed, where the set of rules is gradually expanded as prediction performance increases. Further, it is possible to combine a set of rules into a single rule as in model-based clustering; *e.g.*, rather than separate rules where the left side is either “dorsal pain,” “back pain,” “low back pain,” or “neck pain,” we could use simply “back or neck pain” for all of them.

Another direction for future work is to incorporate higher-order dependence, along the line of work by Berchtold and Raftery (2002). An algorithm for sequential event prediction is presented in ongoing work (Letham, Rudin and Madigan, 2011), which is loosely inspired by the ideas of Berchtold and Raftery (2002), but does not depend on association rules. A third potential future direction is to design a more sophisticated online updating procedure than the one in Section 2.3. It may be possible to design a procedure that approximately updates all of the hyperparameters as more data arrive.

Acknowledgements. This work was partially supported by a Google PhD fellowship in statistics. We would like to acknowledge support for this project from the National Science Foundation under grant IIS-1053407. We would like to thank the editor, associate editor and referee for helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplement A: Additional simulation results

(<http://lib.stat.cmu.edu/aoas/???/???>). In this supplement, we present additional simulation results which speak to the performance of HARM.

In the main text, we consider a combination of popular and randomly selected rules. In Figure 7, we present results for a similar simulation set-up with the to 50 most common rules. While HARM still outperforms the other methods we consider, the distinction is less than with the randomly selected rules. This difference reflects the benefits of the hierarchical structure in HARM for rare rules.

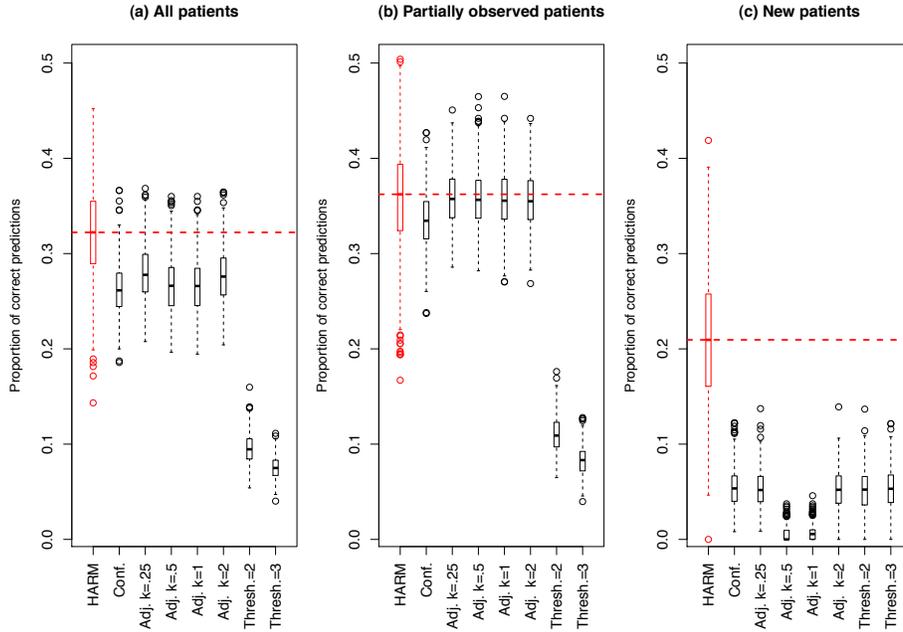


FIG 7. *Simulation experiment with common rules. Each boxplot represents the performance of HARM and other methods for 500 simulation iterations. In each simulation, 200 patients are randomly selected and evaluated using the 50 most common rules.*

The second set of supplementary material concerns model selection. In the text, we use a model which contains all of the available covariates. Using this model highlights the flexibility of our approach, but is not necessarily the model which will give the best predictions. We fit a series of models with an intercept and one demographic factor (age, race, gender, or treatment). We also fit a model with only an intercept term. Figures 8 and 9 present boxplots of the predictive performance of several alternative models. The overall out of sample performance, generated by combining new and partially observed patients, is displayed in Figure 10. The out-of-sample performance of the more complicated model (the full model) is not as good as the out-of-sample performance on some of the simpler models with fewer co-

variates. Since this often indicates a problem with overfitting, we looked at the in-sample performance to see whether it agreed with the out-of-sample performance. Figure 11 presents boxplots of the in-sample predictive performance of the models. As it turns out, the in-sample and out-of-sample performance were similar, so there is no problem with generalization. To determine why the simpler models performed better in-sample, we note that the evaluation metric for the algorithm’s performance is different from the quantity it actually optimizes. This means we are not directly optimizing performance, and this mismatch could potentially cause a substantial difference in performance between the different algorithms.

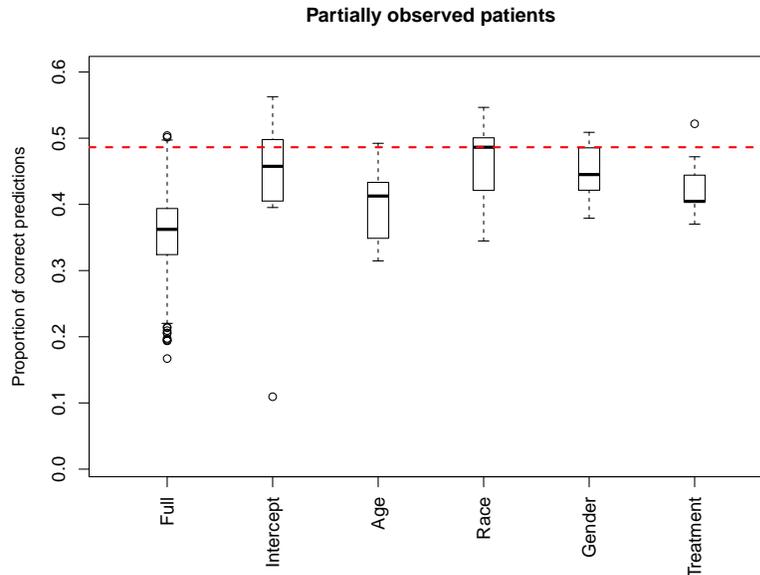


FIG 8. *Simulation experiment for model selection using partially observed patients. Each boxplot represents the performance of HARM with a subset of the available covariates. The model using just race covariates performed the best out of the models we considered.*

In practice, we suggest additional model selection in cases where prediction is the main goal. The results we present could be conceived as the beginning of a forward stagewise procedure based on the simulation strategy outlined in the paper.

In many cases, the best subset for prediction may change. In such cases the computationally intensive strategies we have proposed will likely not be pragmatic. Instead, an approach based on Bayes Factors or one-step-ahead prediction could be implemented.

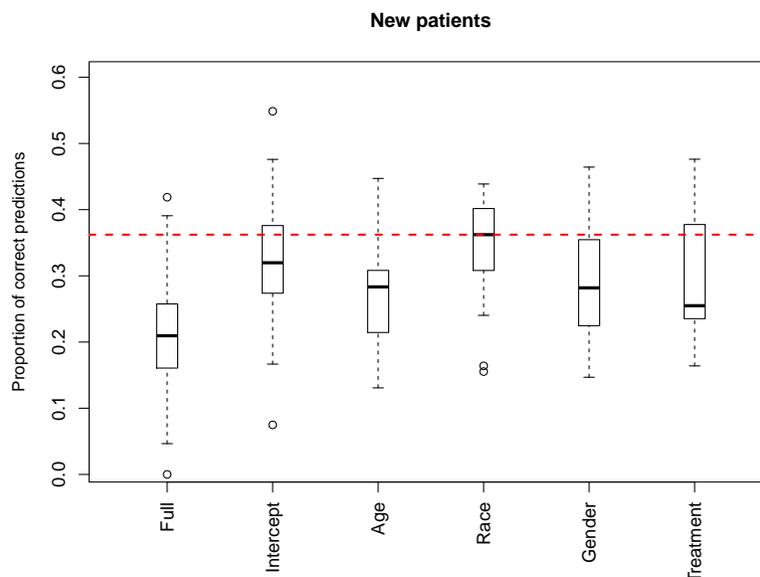


FIG 9. Simulation experiment for model selection using new patients. Each boxplot represents the performance of HARM with a subset of the available covariates. The model using just race covariates performed the best out of the models we considered.

References.

- AGARWAL, D., ZHANG, L. and MAZUMDER, R. (2011). Modeling item-item similarities for personalized recommendations on Yahoo! front page. *Annals of Applied Statistics*. Forthcoming.
- AGRAWAL, R., IMIELIŃSKI, T. and SWAMI, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* 207–216. ACM, New York, NY, USA.
- BERCHTOLD, A. and RAFTERY, A. E. (2002). The Mixture Transition Distribution Model for high-order Markov Chains and non-Gaussian time series. *Statistical Science* **17** pp. 328–356.
- BREESE, J. S., HECKERMAN, D. and KADIE, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty and Artificial Intelligence* 43–52.
- CONDLIFF, M. K., LEWIS, D. D. and MADIGAN, D. (1999). Bayesian Mixed-Effects Models for Recommender Systems. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation*.
- DAVIS, D. A., CHAWLA, N. V., CHRISTAKIS, N. A. and BARABASI, A.-L. (2010). Time to CARE: a collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery* **20** 388–415.
- DUMOUCHEL, W. and PREGIBON, D. (2001). Empirical Bayes screening for multi-item associations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 67–76.

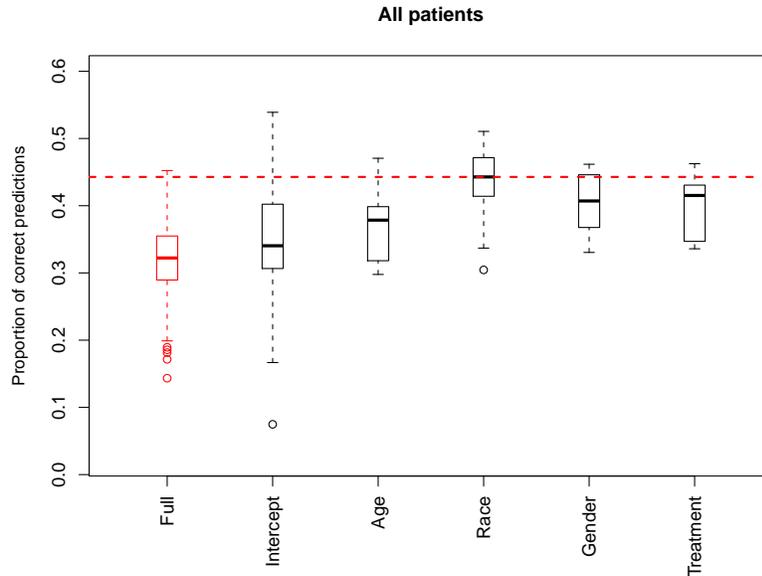


FIG 10. Simulation experiment for model selection using all patients (both partially observed and new patients). Each boxplot represents the performance of HARM with a subset of the available covariates. The model using just race covariates performed the best out of the models we considered.

- FRALEY, C. and RAFTERY, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* **97** 611-631.
- GENG, L. and HAMILTON, H. J. (2007). Choosing the right lens: Finding what is interesting in data mining. In *Quality Measures in Data Mining* 3-24. Springer.
- KUKLINE, E., YOON, P. W. and KEENAN, N. L. (2010). Prevalence of Coronary Heart Disease risk factors and screening for high Cholesterol levels among young adults in the United States, 1999-2006. *Annals of Family Medicine* **8** 327-333.
- LETHAM, B., RUDIN, C. and MADIGAN, D. (2011). Sequential event prediction. In preparation.
- MCCORMICK, T., RUDIN, C. and MADIGAN, D. (2011). Supplement to “Bayesian hierarchical rule modeling for predicting medical conditions”. DOI: .
- PIATETSKY-SHAPIO, G. (1991). Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases* (G. Piatetsky-Shapiro and W. J. Frawley, eds.) 229-248. AAAI Press.
- RUDIN, C., LETHAM, B., KOGAN, E. and MADIGAN, D. (2011a). A learning theory framework for association rules and sequential events. *SSRN eLibrary*.
- RUDIN, C., LETHAM, B., SALLES-AOUISSI, A., KOGAN, E. and MADIGAN, D. (2011b). Sequential event prediction with association rules. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*.
- TAN, P. N., KUMAR, V. and SRIVASTAVA, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

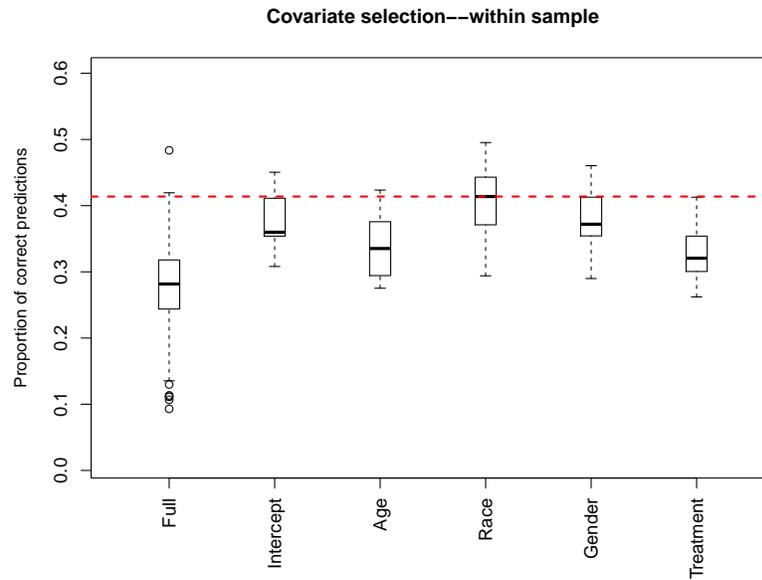


FIG 11. *Simulation experiment for in-sample prediction. Each boxplot represents the performance of HARM for a specific with a subset of the available covariates. For each run of the simulation, patients we evaluated the performance of HARM in predicting the sequence of encounters used for training.*

ADDRESS OF THE FIRST AUTHOR
 DEPARTMENT OF STATISTICS
 UNIVERSITY OF WASHINGTON
 BOX 354320 SEATTLE, WA 98105, USA
 E-MAIL: tylermc@u.washington.edu

ADDRESS OF THE SECOND AUTHOR
 MIT SLOAN SCHOOL OF MANAGEMENT, E62-576
 MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 CAMBRIDGE, MA 02139, USA
 E-MAIL: rudin@mit.edu

ADDRESS OF THE THIRD AUTHOR
 DEPARTMENT OF STATISTICS
 COLUMBIA UNIVERSITY
 1255 AMSTERDAM AVE. NEW YORK, NY 10027, USA
 E-MAIL: madigan@stat.columbia.edu