Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

A HIERARCHICAL MODEL FOR ASSOCIATION RULE MINING OF SEQUENTIAL EVENTS: AN APPROACH TO AUTOMATED MEDICAL SYMPTOM PREDICTION

By Tyler H. McCormick*, Cynthia Rudin† and David Madigan*

*Columbia University** *and Massachusetts Institute of Technology*†

In many healthcare settings, patients visit healthcare professionals periodically and report multiple medical conditions, or *symptoms*, at each encounter. We propose a statistical modeling technique, called the Hierarchical Association Rule Model (HARM), that predicts a patient's possible future symptoms given the patient's current and past history of reported symptoms. The core of our technique is a Bayesian hierarchical model for selecting predictive association rules (such as "*symptom 1 and symptom 2* $\rightarrow$ *symptom 3*") from a large set of candidate rules. Because this method "borrows strength" using the symptoms 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 symptoms 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 symptoms. 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.*, 2010). Specifically we propose the Hierarchical Association Rule Mining 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 symptoms 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 symptom that a patient will experience.
- Given basic demographic information, predict the first symptom that a patient will report.
- Given partial data from an encounter (and possibly prior encounters) predict the next symptom.

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 symptoms. This patient-level sparsity presents a challenge for predictive modeling. Our hierarchical modeling approach attempts to address this challenge by borrowing strength across patients.

Applications of association rules usually do not usually concern supervised learning problems (though there exist some exceptions, e.g. Veloso *et al.*, 2008). The sequential event prediction problem is a supervised learning problem, that as far as we know, has been formalized only here and by Rudin *et al.* (2010). 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.

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.*, 2010) 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 symptom 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 symptom 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 symptom b, given a set of current symptoms, 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 symptom predictions.

2.1. *Definitions.* An *association rule* in our context is an implication $a \to b$ where the left side is a subset of symptoms that the patient has experienced, and b is a single symptom 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 $a \to b$ for a patient is the empirical probability:

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

The support of set a is:

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

where $\hat{P}(a)$ is the empirical proportion of times that symptoms a were experienced. When a patient has experienced a particular set of symptoms 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 symptoms. 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 often not even 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 symptom that occurs rarely is strongly linked with another rare symptom, it is essential not to exclude the rules characterizing these symptoms. In our data, the distribution of symptoms has a long tail, where the vast majority of events happen rarely: out of 1800 possible symptoms, 1400 occur less than 10 times. These 1400 symptoms 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.*, 2010):

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

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 $P(b|a)$, rules are extremely interpretable, 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:

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

where $\#x$ is the support of symptom 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 symptom that has occurred only once, and also whether a particular symptom is more likely to repeat than another.

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

from common prior causes such as other symptoms 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 $\cap$ rhs) in n_{ir} relevant previous encounters (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 0 (patient i has never reported the symptoms on the left hand side of r before). 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 symptoms occur very infrequently. In some cases these symptoms may still be highly associated with certain other conditions. For instance, compared to some symptoms, migraines are relatively rare. Patients who have migraines however typically also experience nausea. A minimum support threshold algorithm might easily exclude this rule if migraines 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.* (2010), 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 $\mu_{\boldsymbol{\beta}}$ and variance $\sigma_{\boldsymbol{\beta}}^2$. 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 μ_τ, σ_τ^2, $\mu_{\boldsymbol{\beta}}$, and $\sigma_{\boldsymbol{\beta}}^2$. 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 $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_R$. The prior assumptions yield the following

posterior:

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

HARM produces draws from the (approximate) posterior distribution for each probability. In the context of symptom 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 mode. We carry out our computations using a Gibbs sampling algorithm, provided in Figure 1.

2.3. *Online updating.* Given a batch of data, HARM makes predictions based on the posterior distributions of p_{ir}. Since the posterior is not available in closed form, predictions using HARM requires iterating the algorithm in Figure 1 to convergence. The next time the patient visits the physician, p_{ir} could be updated by again iterating the algorithm in Figure 1 to convergence. In some applications new data continue arrive frequently, making it impractical to compute approximate posterior distributions using the algorithm in Figure 1 for each new encounter. In this section we provide an online updating scheme which incorporates new patient data after an initial batch of encounters has already been processed.

Beginning with an initial batch of data, we run the algorithm in Figure 1 to obtain $\hat{\tau}_i$ and $\hat{\pi}_{ir}$, which are defined to be posterior mean of the estimated distributions for τ_i and π_{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:

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

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,

For a suitably initialized chain, at step v:

1. Update p_{ir} from the conjugate Beta distribution given $\pi, \tau, \mathbf{Y}, \mathbf{N}, \mathbf{M}$.
2. Update τ_i using a Metropolis step with proposal τ_i^* where
$$\log(\tau_i^*) \sim \mathrm{N}(\tau_i^{(v-1)}, (\text{scale of jumping dist})).$$
3. For each rule, update the vector $\boldsymbol{\beta}_r$ using a Metropolis step with
$$\log(\boldsymbol{\beta}_r^*) \sim \mathrm{N}(\boldsymbol{\beta}_r^{(v-1)}, (\text{scale of jumping dist})).$$
4. Update γ_i using a Metropolis step with
$$\log(\gamma_i^*) \sim \mathrm{N}(\gamma_i^{(v-1)}, (\text{scale of jumping dist})).$$
5. Update $\mu_{\boldsymbol{\beta}} \sim \mathrm{N}(\hat{\mu}_{\boldsymbol{\beta}}, \sigma_\beta^2)$ where
$$\hat{\mu}_{\boldsymbol{\beta}} = \Big(\frac{1}{D+R}\Big) \sum_{r=1}^{R} \sum_{d=1}^{D} \beta_{r,d}.$$
6. Update $\sigma_{\boldsymbol{\beta}}^2 \sim \text{Inv-}\chi^2(d-1, \hat{\sigma}_{\boldsymbol{\beta}}^2)$ where
$$\hat{\sigma}_{\boldsymbol{\beta}}^2 = \Big(\frac{1}{D+R}\Big) \sum_{r=1}^{R} \sum_{d=1}^{D} (\beta_{r,d} - \mu_{\boldsymbol{\beta}})^2.$$
7. Update $\sigma_\tau^2 \sim \text{Inv-}\chi^2(I-1, \hat{\sigma}_\tau^2)$ where $\hat{\sigma}_\tau^2 = \frac{1}{I}\sum_{i=1}^{I}(\tau_i - \mu_\tau)^2$.
8. Update $\mu_\gamma \sim \mathrm{N}(\hat{\mu}_\gamma, \sigma_\gamma^2)$ where $\hat{\mu}_\gamma = \frac{1}{I}\sum_{i=1}^{I}\gamma_i$.
9. Update $\sigma_\gamma^2 \sim \text{Inv-}\chi^2(I-1, \hat{\sigma}_\gamma^2)$ where $\hat{\sigma}_\gamma^2 = \frac{1}{I}\sum_{i=1}^{I}(\gamma_i - \mu_\gamma)^2$.

FIG 1. *Gibbs sampling algorithm for hierarchical bayesian association rule mining for sequential event prediction (HARM).*

$P(y_{ir}^{(\text{newobs.})}|n_{ir}^{(\text{newobs.})}, p_{ir})$, 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 online updating scheme in Section 2.3, on co-prescribing data from a large clinical trial. 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 symptom/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 symptoms 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. We fit age using a series of indicator variables corresponding to four groups (40-49, 50-59, 60-69, 70+).

Our experiments consider only the marginal probabilities (support) and probabilities conditional on one previous symptom. Thus, the left hand side of each rule contains either 0 items or 1 item.

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 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 encounters 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 the top 50 most frequently reported conditions; these conditions represent 60% of all conditions reported.

The algorithm was used to iteratively predict the condition revealed at each encounter. For each selected patient, starting with their first test encounter, and prior to that encounters' 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 modes 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 current condition. This process was repeated for the patient's remaining encounters. 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.

Figures 2, 3, and 4 show the results. Figure 2 presents the distribution of scores for the entire collection of partially observed, fully observed, and new

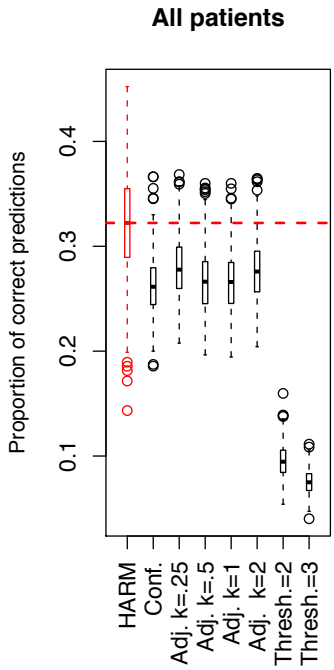

FIG 2. *Predictive performance for all patients. Each boxplot represents the distribution of scores over 500 runs. These plots include data from both partially observed and new patients.*

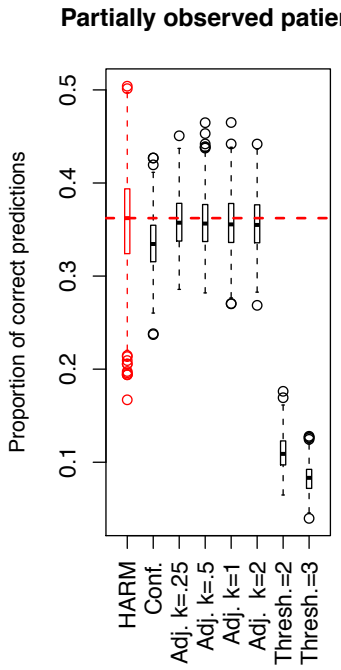

FIG 3. *Predictive performance for partially observed patients. These are patients for which there are training encounters.*

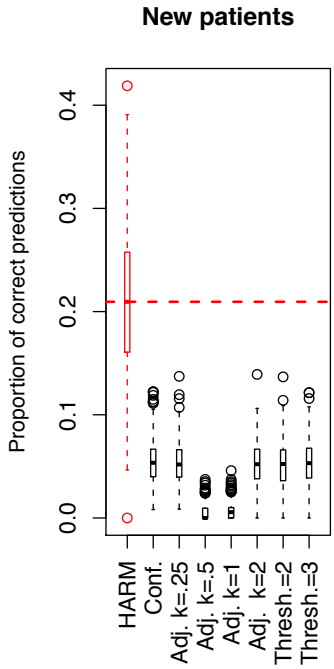

FIG 4. *Predictive performance for new patients.*

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 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. We consider the predictive performance of HARM with respect to partially observed (Figure 3) and new (Figure 4) patients. Though our method overall has a higher frequency of correct predictions than the other algorithms in both cases, the advantage is more pronounced for new patients; in cases where there is no data for each patient, there is a large advantage to sharing information.

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 → myocardial infarction: Figure 5 plot (a) shows the distribution of posterior median 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

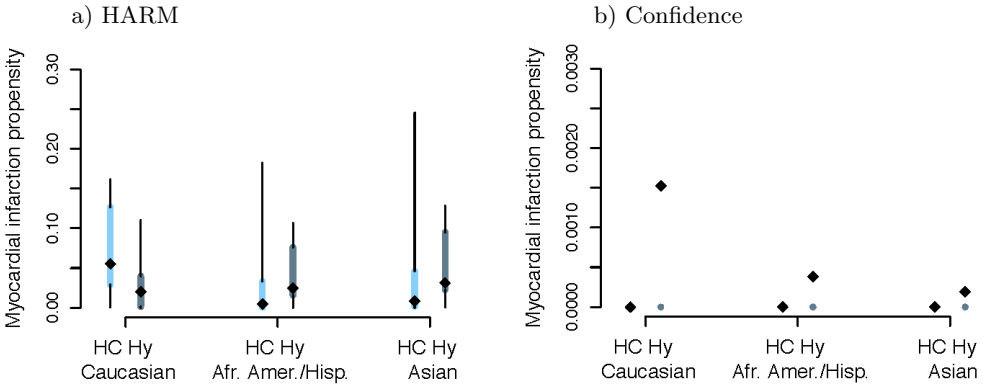

FIG 5. *Propensity of myocardial infarction in patients who have reported high cholesterol or hypertension using HARM (plot (a)) and (unadjusted) confidence (plot (b)). For each demographic group, high cholesterol (HC) is on the left and hypertension (Hy) is on the right. Thick lines represent the middle half of the posterior median 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.*

respondents who are observably similar. Some other specific characteristics of the 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 5 (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. The colored, smaller dots represent the rest of the distribution (quartiles), which 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 → 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 6 and 7 display the posterior medians of propensity of myocardial infarction for respondents separated by age and treatment condition. Figure 6 considers patients who have reported hypertension, Figure 7 considers patients who have reported high cholesterol. In both Figure 6 and Figure 7, it appears that the propensity of myocar-

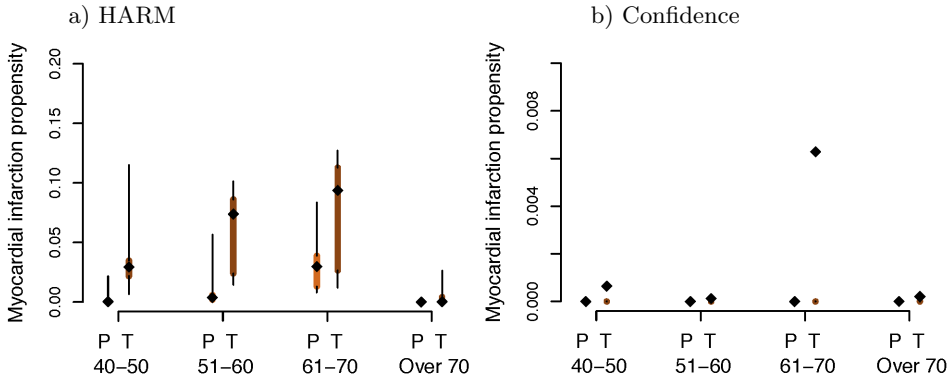

FIG 6. *Propensity of myocardial infarction in patients who have reported hypertension, estimated by HARM (plot (a)) and (unadjusted) confidence (plot (b)). For each demographic group, the placebo (P) is on the left and the treatment medication (T) is on the right. Thick lines represent the middle half of the posterior median 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 that those who do not.*

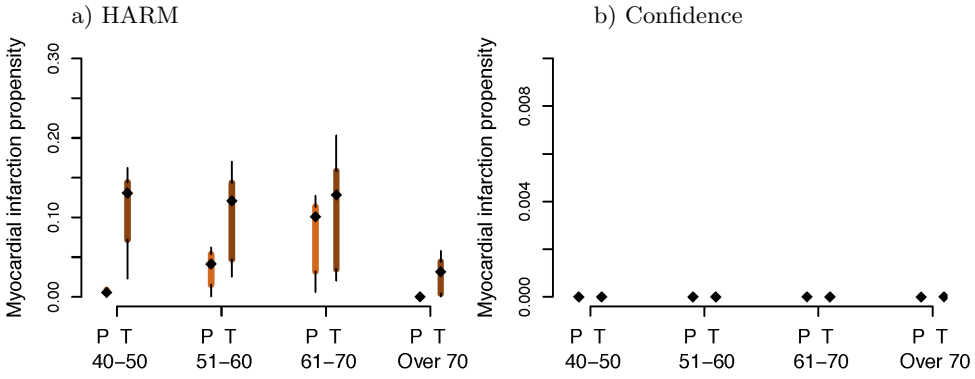

FIG 7. *Propensity of myocardial infarction in patients who have reported high cholesterol, estimated by HARM (plot (a)) and (unadjusted) confidence (plot (b)). For each demographic group, the placebo (P) is on the left and the treatment medication (T) is on the right. Thick lines represent the middle half of the posterior median propensities for respondents in the indicated demographic group. Outer lines represent the middle 90% and dots represent the mean.*

dial infarction predicted by HARM is greatest for individuals between 50 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 6 and Figure 7. 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 demonstrated by Figures 6(b) and 7(b).

4. Related Works. As far as we know, the line of work by Davis *et al.* (2009) is the first to use an approach from recommender systems to predict medical symptoms, though in a completely different way than ours; it is based on vector similarity, in the same way as Breese, Heckerman and Kadie (1998a) (also see references in Davis *et al.* (2009) for background on collaborative filtering).

Three relevant works on Bayesian hierarchical modeling and recommender systems are those of DuMouchel and Pregibon (2001, "D&P"), Breese, Heckerman and Kadie (1998b), and Condliff, Lewis and Madigan (1999). D&P 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 Breese et al. nor Condliff et al. aim to model repeat purchases (recurring symptoms). Breese et al. 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.

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, $a \to b$ indicates symptoms a are followed by symptoms 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 symptoms 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 symptom prediction, where experiencing two symptoms in succession may have different clinical implications than experiencing either symptom in isolation.

collaborative nature of medical symptoms make the sequential predidtion

There are several possible directions for future work. One possibility is to investigate whether expanding the set of rules has an influence on prediction accuracy. In the case that the set of rules is too large, it may be important to develop parsimonious representations of these associations, potentially through a method similar to model-based clustering (Fraley and Raftery, 2002). Another direction is to incorporate higher-order dependence, along the line of work by Berchtold and Raftery (2002). 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 directly updates the hyperparameters as more data arrive.

Acknowledgements. Tyler McCormick is supported by a Google PhD Fellowship in Statistics.

References.

Agrawal, R., Imieliński, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 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. (1998a). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. 43–52. Morgan Kaufmann.

Breese, J. S., Heckerman, D. and Kadie, C. M. (1998b). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proc. Uncertainty in Artificial Intelligence* 43-52.

Condliff, M. K., Lewis, D. D. and Madigan, D. (1999). Bayesian Mixed-Effects Models for Recommender Systems. In *ACM SIGIR 99 Workshop on Recommender Systems: Algorithms and Evaluation*.

Davis, D. A., Chawla, N. V., Christakis, N. A. and Barabsi, A.-L. (2009). 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 *Proc. ACM SIGKDD international conference on Knowledge discovery and data mining* 67–76.

Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* **97** 611-631.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis*, 2 ed. Chapman and Hall/CRC.

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, United States, 1999-2006. *Annals of Family Medicine* **8** 327-333.

Piatetsky-Shapiro, 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., Salleb-Aouissi, A., Kogan, E. and Madigan, D. (2010). A Framework for Supervised Learning with Association Rules. Submitted.

Tan, P. N., Kumar, V. and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *KDD '02: Proc. Eighth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Veloso, A. A., Almeida, H. M., Gonçalves, M. A. and Jr., W. M. (2008). Learning to rank at query-time using association rules. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* 267–274. ACM, New York, NY, USA.

Address of the First and Third authors
Department of Statistics
Columbia University
1255 Amsterdam Ave. New York, NY 10027, USA
E-mail: tyler@stat.columbia.edu
madigan@stat.columbia.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