Abstract

Background: Hospitals in the US and other jurisdictions are being financially penalized based on estimates of their effect on 30-day readmission risk, after adjustment for confounders. Although hospital administrative data is information-rich, confounder adjustment in these models tends to be crude, risking residual confounding. Non-parametric machine learning techniques can take advantage of these rich data to predict readmission, but cannot isolate the independent effect of hospitals on readmission risk.

Research Design: To estimate the marginal effect of care at different hospitals on 30-day readmission risk, we used targeted maximum likelihood estimation (TMLE), which allowed us to use a non-parametric machine learning technique (random forest) to take advantage of the rich confounder data. We used an 11-year cohort of 65 year old patients from 20 hospitals in Montreal, Canada, and developed three models to estimate marginal readmission risk at each of the hospitals after hospitalization for heart failure, acute myocardial infarction (AMI), and pneumonia. To control for confounding, we modeled the effect of hundreds of types of pre-admission outpatient drug prescriptions, medical procedures, and diagnoses on readmission risk.

Results: Within 30 days of discharge, there were 5,520 / 24,847 (22%) heart failure readmissions, 3,183 / 20,421 (16%) pneumonia readmissions, and 2,525 / 15,746 (16%) AMI readmissions. Within each hospital, there was a wide variation in crude readmission risk across the twenty hospitals for pneumonia 3,183 / 20,421 (16%), heart failure 5,520 / 24,847 (22%), and AMI 2,525 / 15,746 (16%). However, after control for confounding, the marginal risk for readmission within all hospitals was nearly the same (within each admission category).

Conclusion: Although crude 30-day readmission risk estimates show variation between different hospitals, no difference was found in marginal risks after control for confounding.

1 Introduction

In the early eighties, hospital administrators in the US sought to reduce hospitalization costs by changing the reimbursement system. Instead of paying hospitals per day of hospitalization, hospitals were paid a fixed rate for the type of hospitalization and the procedures performed. Following implementation of this law, the length of stay at hospitals dropped dramatically, although evidence exists that is was already in decline.

Some worry that the new system created a perverse incentive to discharge patients early, and admit them again at a later date. To ensure proper quality of care in the hospitals while keeping costs controlled, administrators have sought to establish useful quality of care metrics. Hospital readmissions have been identified as a simple metric that can establish a baseline of care; if an abnormally high number of patients from a certain hospital are quickly readmitted, it could indicate poor quality of care.

Since hospitals admit patients with varying risk of readmission, it is important to accurately estimate the effect of hospital treatment on readmission independent of patient-level confounders. Without effectively controlling for confounding, we risk unfairly penalizing hospitals that treat sicker (more likely to be readmitted) patients. Fortunately, hospital and outpatient administrative data is information-rich. Drug prescriptions, diagnoses, and medical procedures can provide important information on how the effect of hospital care on readmission risk is confounded by patient health.

However, in most statistical models of readmission risk, the hospital administrative data is simplified to a few well-known confounders (age, sex, previous readmissions), and sometimes a summary "comorbidity score". If each drug, diagnosis and procedure was modeled with a separate covariate, the model would be very computationally expensive to fit. Such a model would also be very unwieldy to develop; analyzing how inclusion or exclusion of variables affects the model would be impossible to do effectively with hundreds of covariates.

By summarizing confounders into crude risk scores, we risk "residual confounding" leading to biased effect estimates. Furthermore, to compare hospitals, we are only interested in estimating one parameter, the independent effect of hospitals on readmissions. We are only interested in the other variables insofar as they confound the effect of hospitals on readmissions, estimating the individual effect of each of these variables is unnecessary, and statistically wasteful.

Non-parametric machine learning techniques are available that will let us accurately discriminate patient readmission risk using hundreds of variables in a computationally efficient way. Furthermore, because they are non-parametric, we avoid having to specify a functional form, and can find complex "interactions" between variables. On the other hand, these non-parametric techniques don't allow us to isolate (target) the effect of specific variables (such as care at a particular hospital) on readmission risk.

The targeted maximum likelihood estimator (TMLE), is a doubly-robust technique that uses propensity scores to estimate target parameters of interest from non-parametric (or parametric) models. To use TMLE, a model is developed to estimate the probability of exposure (the propensity score), and also fit another model to estimate the probability of outcome. These two probabilities are combined in a parametric model with only the parameter of interest, inversely weighted by the probability of exposure, and offset by the probability of the outcome. In this way, the discriminative power of non-parametric models can be used to extract estimates of parameters of interest.

In this study, we sought to estimate the marginal risk of readmission within 30 days of discharge for three different admission diagnoses (pneumonia, heart failure, and acute myocardial infarction) for each of twenty Montreal hospitals. We used a ten-year cohort of 65 year old people, for which we had outpatient and inpatient diagnoses and procedures, and outpatient drugs prescriptions. We used a non-parametric machine learning technique, (random forest), with TMLE to take advantage of the rich confounder data and provide less biased estimates of readmission risk at each hospital.

2 Methods

2.1 Study Design

We used a cohort extracted from a Canadian provincial (Quebec) administrative database of hospitalizations, obtained from the *Régie de l'assurance maladie du Québec* (RAMQ). We enrolled patients into this cohort on the month that two conditions were satisfied: 1) they had at least one diagnosis of a respiratory illness (the exact list of respiratory International Classification of Diseases, 9th Revision [ICD-9] codes is given in the Appendix) between January 1st, 1996 and March 31, 2006 (the study period), while living in the 2006 census metropolitan area of Montreal, and 2) were at least 65 years of age. We used this cohort because it represents the majority of 65-year olds who were hospitalized in the region during the study period.

From among this cohort, we selected hospital discharges for those who had accrued at least one continuous year in the cohort preceding the time of admission. We restricted our data to only the discharges from the twenty hospitals with the most discharges of patients 65 years or older within the study period; the twenty hospitals accounted for 75% of all such discharges. We only selected hospital discharges which resulted from hospital stays of at least one day. Therefore, the earliest possible hospital discharge was January 2, 1997.

From among the identified hospital discharges, we selected only those with one of three high-volume admission diagnoses with high rates of hospital readmissions: pneumonia, acute myocardial infarction (AMI), and heart failure, the three initial conditions selected by the Centers for Medicare and Medicaid Services (CMS) to implement the Hospital Readmissions Reduction Program mandated by the Affordable Care Act. We identified each of the admission diagnoses using ICD-9 codes; for pneumonia we used codes ranging from 480-487, for heart failure we used all 428 codes, and for AMI we used all 410 codes. The following methods were applied individually to all three disease subsets.

2.2 Hospital readmissions

The unit of analysis was the hospital discharge; a person could be discharged multiple times. A hospital readmission was defined as an emergency hospital admission to any Quebec hospital in the 30 days following a discharge. A person who died or had a non-emergency readmission in the 30 days following discharge was considered not readmitted. For the purposes of the survival analysis, "time zero" was the day of discharge; a person was considered right-censored if, before an emergency readmission, they died, had a non-emergency readmission, they changed residence to a location outside of Quebec, or the study ended.

2.3 Confounders and Risk Factors

For each hospital discharge, we colllected variables that measured states at the time of admission, or events that occurred prior to the hospital admission, and which may confound the relationship between hospital care and readmission. We used the demographic characteristics (age at time of admission (years), sex, birth year-month), the number of previous readmissions (within the preceding year), the admission diagnosis (as measured by the specific ICD-9 code). We also included the day of week of discharge, which has been previously shown to have an association with readmissions[1], and the month of discharge, because we hypothesized that readmission risk would vary by seasons in Montreal.

Additionally, for each discharge, we collected the Quebec hospital diagnoses, Quebec hospital procedures, and drugs dispensed outside of the hospital but inside Quebec, in the year preceding the admission. The hospital procedures were recorded in the Canadian Classification of Diagnostic, Therapeutic, and Surgical Procedures (CCP) system. Hospital diagnostic codes were coded using the ICD-9 system. Finally, drugs which were prescribed and dispensed outside the hospital, and were being taken on the day of admission were also recorded for each patient in the *code commune* system, which categorizes drugs based on the chemical compound. To ease computation, before fitting any model, we removed any diagnosis, procedure or drug that occurred less than 30 times among all discharges. We chose 30 because it appeared to be a natural breakpoint; if the number of variables included is a function f of the threshold, then the first derivative of f dropped at 30 for all three disease categories.

We believed that residential location would strongly affect the probability of admission to the hospital nearest that census tract. We included it in our models because we also expected it to crudely approximate a (expected) confounder: socio-economic status. We used the residential postal code at the time of admission to assign each patient in the cohort to a census tract, as defined by the 2006 Canadian census. (Census tracts contain between 2,500 and 8,000 people, and, at the time of their creation, are demarcated so as to maximize homogeneity of socioeconomic characteristics.)

2.4 30-day readmission

For each discharge i, we sought to estimate the effect of each of the twenty hospitals $A \in \{hosp_1, \ldots, hosp_{20}\}$ on 30-day readmission (Y), accounting for the vector of confounders (W). To estimate this risk, we used targeted maximum likelihood estimation, which consisted of several steps. We first estimated of a model of the propensity score $g = Pr(A|W)$ (using random forest described below). Next, we estimated of a model of readmission risk based on the confounders W and the variables for each of the hospitals $Q = Pr(Y = 1|A, W)$. We then calculated $h_a(A, W)$ (sometimes referred to as the clever covariate) described in equation 1

$$h_a(A, W) = \frac{I(A = a)}{g(a|W)} \tag{1}$$

(where I is the indicator function which evaluates to 1 when its argument is true, and 0 otherwise), and solved for all ϵ_a in fluctuation function described in equation 2.

$$Y_i = expit(logit(Q(Y_i|A_i, W_i)) + \sum_{a=hosp_1}^{hosp_{20}} \epsilon_a \times h_a(A_i, W_i)) \tag{2}$$

We solved for all twenty ϵ_a by fitting by regressing the 30-day readmission outcome Y (with a logit link function) onto $h_a(A, W_i)$ (with no intercept) offset by the inverse logit of the initial estimate of readmission risk $Q = (Y|A, W)$. Finally, for each discharge, we computed the estimated risk of 30-day readmission for all twenty counterfactual conditions (the risk of readmission for every discharge as if they had attended different hospital) using Equation 3.

$$Q^*_{ai} = expit(logit(Q(Y|a, W_i)) + \frac{\epsilon}{g(a|W)}) \tag{3}$$

For each hospital, we then cacluated the mean readmission risk (Q^*_a) and associated odds ratio.

To estimate both models $g(A_i|W_i)$ and $Q(Y_i|A, W_i)$, we used a random forest, a non-parametric model based on decision trees[2]. Decision trees use the independent variables (W_i) to repeatedly split data into partitions that are as homogenous as possible with respect to the outcome of interest (specifically measured with the Gini coefficient[3]). Complex interactions and nonlinearities are naturally modeled within this framework. Random forest improves decision trees by using bootstrap aggregation (bagging); multiple decision trees are grown on bootstrap replicates (sampled with replacement) to avoid overfitting. Additionally, within each tree, only a sample of the covariates is used (in our case we used a square root of the number of variables, rounded down, included in the model).

For both models $g(A_i|W_i)$ and $Q(Y_i|A, W_i)$, we arbitrarily chose to grow 1200 trees, and then measured the accuracy as a function of the number of trees to ensure that growing further trees would be unlikely to significantly improve accuracy. Because the model was used solely to estimate the *probability* of admission to to specific hospitals (and not to predict exactly which hospital was attended), we configured the model to favour calibration over discrimination: we weighted each of the twenty predicted hospitals by the inverse of the proportion of discharges at that hospital. When measuring the accuracy for each discharge, we only used trees for which the discharge was "out-of-bag", that is, we only used trees for which the bootstrap sample did not include the discharge.

To describe importance of the covariates in both models $g(A_i|W_i)$ and $Q(Y_i|A, W_i)$, for each variable, we measured the decrease in the Gini coefficient for each partition in which the variable was used, in every tree. A low Gini (i.e. higher decrease in Gini) means that a particular predictor variable plays a greater role in partitioning the data into the defined classes. We plotted the densities of variables with four different classes (census tract, procedure, diagnosis and drug) at different levels of Gini decreases.

Random forest classifies each item by majority vote. Although the vote proportion is in the scale of zero to one, it is not calibrated well as a probability; to calibrate the vote proportion, we used Platt scaling (logistic regression of the outcome (Y_i) on to the vote proportion).

2.5 Comparison to assessment with comorbidity scores

Finally, we compared our results of our analysis with 1) a logistic regression for 30-day readmission and 2) a Cox proportional hazards model for the time-to-readmission. In these models, we included only the age, sex, number of previous admissions, and the Charlson comoribidity score (Elixhauser version), along with indicator variables representing the hospitals themselves.

2.6 Software

The data were cleaned and prepared for statistical analysis using the Postgres relational database (version 9.2.6). We implemented our models using the R statistical package (version 3.1.1)[4]. We implemented the random forest using the "bigrf" package (version 0.1.11)[5]. We implemented the GLM fitting using coordinate descent using the "glmnet" package (version 1.9.8)[6].

We plotted our figures using the "ggplot2" package (version 1.0.0)[7]. All the code to develop used to fit clean our data, fit our models, and layout this paper is available for download at Github.

3 Results

Over the course of January 2, 1996 to March 31, 2006, 482,064 people were entered into our cohort. Among these, 16,521 were ever admitted for pneumonia, 13,884 were ever admitted for AMI, and 15,822 were ever admitted for heart failure. People ever admitted for pneumonia had a mean (median) 1.2 (1) pneumonia admissions, heart failure patients had a mean (median) 1.6 (1) heart failure admissions, and AMI patients had a mean (median) 1.1 (1) AMI admissions. In total, we analyzed 20,421 pneumonia discharges, 15,746 AMI discharges, and 24,847 heart failure discharges.

The accuracy of the random forest (for both models g and Q) did not appear to improve significantly beyond 125 trees (see figure 3 in the appendix). In figure 1 we plot the importance of variables (as measured by the Gini coefficient) in the random forest models for four variable classes, for all disease subsets for both the g and Q model. Although census tracts were found to be important in prediction of hospital choice, the other three variable classes were had a high density of important variables as well. The prescription drugs in particular had a high proportion of important variables, and generally the lowest proportion of unimportant variables. For the Q model, the variable density appeared bimodal within variable importance for all four variable classes. Additionally, the pre-admission drug prescriptions appeared to be strongly important in predicting readmission for all pneumonia, heart failure, and AMI admissions.

The predicted probability of admission to any particular hospital ($g = Pr(A = a|W)$) was less than 5% in 88% of cases (across all disease subsets and hosptials). We set δ (the lower bound of g when used to fit the ϵ values for $Q*$), to two different values, 10^{-2} and $10^{-2.5}$. Across all disease subsets and hospitals, 39% of discharge/hospital combinations and a g less than 10^{-2}, and 4% had a g less than $10^{-2.5}$. Figure 4 (in the Appendix) describes the histogram of g when it is below 0.05 for each disease/hospital combination separately.

The unadjusted proportion of patients readmitted in 30 days varied across hospitals for each disease subset (Tables 1-3). The linear correlation between the proportion of deaths during hospital stay and the proprtion readmitted was (0.19, -0.55, -0.28) among AMI, heart failure, and pneumonia admissions respectively. Using a model that adjusts for a few well-known confounders, for AMI, heart failure, and pneumonia respectively, one, three, and five hospitals had significantly different odds than the reference hospital. Notably, the significant odds ratios are all relatively small, with point estimates ranging from 0.92 - 1.04. In contrast, in the TMLE models, at both values of δ, for all admission diagnoses, nearly all of the hospitals had significantly different odds than the reference hospital.

In some hospitals and disease subsets, the parameter δ, (the lower bound on the probability of exposure $g(A|W)$) had a considerable effect on the marginal risk and the associated odds ratios. For example, for AMI (shown in Table 1), the marginal risk for hospital 17 increases by six percent when δ decreases from 10^{-2} to $10^{-2.5}$. In figure 2, we display the marginal risk for each of the twenty hospitals and disease subsets as a function of the parameter δ. For many hospitals, the effect was quite strong; for pneumonia admissions, hospital 16 went from having the second-lowest marginal risk when $\delta = 0.1$ to having the highest marginal risk when $\delta = 0.025$.

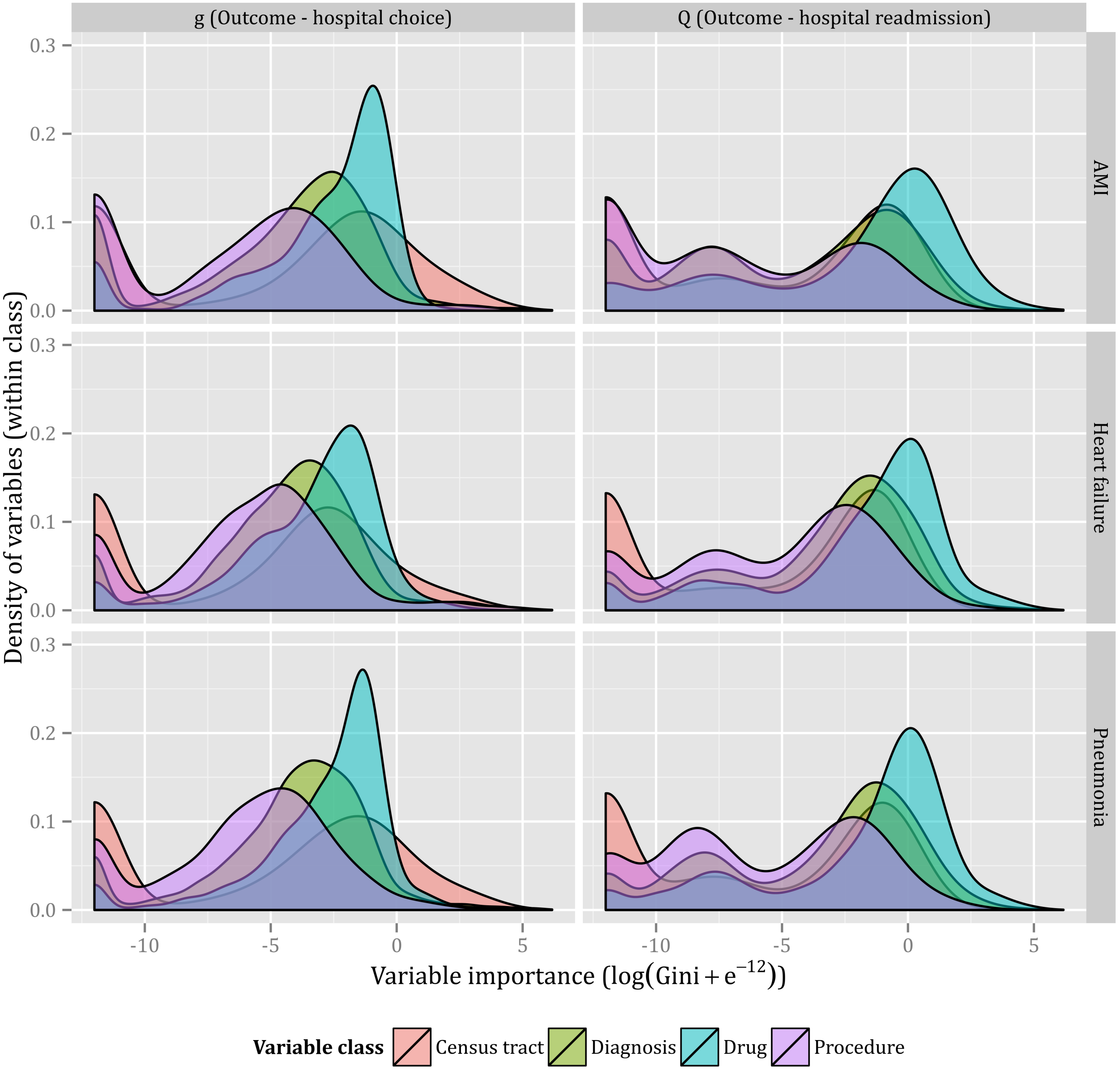

Figure 1: Variable importance by model and variable class. For each random forest classifier, the variable importance was measured by the decrease in the Gini coefficient when that variable splits a node. The horizontal axis within each panel is displayed on a log_e scale. Some variables had exactly zero importance; to avoid evaluating the logarithm of zero, we added a small constant (e^{-12}) to the measure of variable importance. The vertical axis in each panel represents the variable density at the corresponding level of variable importance. To transform the individual variable importances into a continuous density, we smoothed using a Gaussian kernel density estimator, using Silverman's 'rule-of-thumb'[8] to select the bandwidth. The density is measured separately for each class; the area under each variable class curve is exactly one.

Hsp.					**Logistic regression**		**TMLE ($\delta = 10^{-2}$)**		**TMLE ($\delta = 10^{-2.5}$)**	
	Admitted	Died	Discharged	Readmitted	Odds ratio	Marginal	Odds ratio	Marginal	Odds ratio	Marginal
		(%)		(%)	(95% CI)	risk	(95% CI)	risk	(95% CI)	risk
1	763	112 (15)	651	105 (16)	0.98 (0.95-1.01)	0.16	0.86 (0.83-0.89)	0.16	0.77 (0.74-0.81)	0.14
2	1557	148 (10)	1409	191 (14)	0.97 (0.95-1.00)	0.15	0.85 (0.82-0.87)	0.15	0.85 (0.82-0.87)	0.15
3	606	83 (14)	523	84 (16)	0.98 (0.95-1.02)	0.16	1.01 (0.97-1.05)	0.18	1.09 (1.04-1.14)	0.19
4	1022	125 (12)	897	136 (15)	0.97 (0.94-1.00)	0.15	0.72 (0.69-0.74)	0.13	0.72 (0.69-0.74)	0.13
5	729	150 (21)	579	98 (17)	0.98 (0.95-1.02)	0.16	0.75 (0.72-0.77)	0.14	0.73 (0.71-0.76)	0.14
6	826	119 (14)	707	106 (15)	0.98 (0.94-1.01)	0.15	0.57 (0.54-0.60)	0.11	0.57 (0.54-0.60)	0.11
7	1491	241 (16)	1250	216 (17)	0.99 (0.96-1.01)	0.16	1.04 (1.01-1.06)	0.18	1.03 (1.01-1.06)	0.18
8	1270	198 (16)	1072	138 (13)	0.95 (0.92-0.98)	0.13	0.69 (0.67-0.71)	0.13	0.69 (0.67-0.71)	0.13
9	780	152 (19)	628	130 (21)	1.01 (0.97-1.05)	0.19	0.54 (0.51-0.56)	0.10	0.52 (0.50-0.54)	0.10
10	778	124 (16)	654	123 (19)	1.01 (0.97-1.05)	0.19	1.19 (1.15-1.23)	0.20	1.27 (1.22-1.31)	0.21
11	705	125 (18)	580	97 (17)	0.99 (0.96-1.03)	0.17	0.89 (0.85-0.92)	0.16	0.90 (0.86-0.94)	0.16
12	1284	266 (21)	1018	166 (16)	0.99 (0.96-1.02)	0.16	0.90 (0.88-0.93)	0.16	0.90 (0.88-0.93)	0.16
13	739	86 (12)	653	110 (17)	0.99 (0.95-1.02)	0.16	1.19 (1.16-1.23)	0.20	1.22 (1.18-1.27)	0.21
14	1307	184 (14)	1123	210 (19)	(Reference)	0.18	(Reference)	0.18	(Reference)	0.18
15	1152	168 (15)	984	129 (13)	0.97 (0.95-1.01)	0.15	0.70 (0.68-0.73)	0.13	0.70 (0.68-0.73)	0.13
16	408	70 (17)	338	43 (13)	0.97 (0.93-1.01)	0.15	0.84 (0.80-0.88)	0.15	0.84 (0.80-0.89)	0.15
17	807	123 (15)	684	134 (20)	1.02 (0.99-1.06)	0.20	1.76 (1.72-1.81)	0.27	2.30 (2.23-2.37)	0.33
18	894	144 (16)	750	116 (15)	0.98 (0.95-1.01)	0.16	0.91 (0.87-0.94)	0.16	0.91 (0.87-0.95)	0.16
19	499	94 (19)	405	50 (12)	0.95 (0.91-0.99)	0.13	0.57 (0.53-0.61)	0.11	0.57 (0.53-0.61)	0.11
20	1025	184 (18)	841	143 (17)	0.99 (0.96-1.02)	0.17	1.05 (1.02-1.09)	0.18	1.05 (1.02-1.09)	0.18

Table 1: Risk of 30-day readmission after admission for acute myocardial infarction (AMI) in twenty Montreal hospitals. The proportion of those who were readmitted within 30 days is caluculated using the number discharged alive as the denominator. The confidence intervals for the odds ratios for the parameters in the logistic regression model were calculated using the profile likelihood method. The marginal risk for the odds ratios was calculated by using the regression model to calculate the mean predicted probability of readmission for every admission, except individually fixing the hospital attended to one hospital. The parameter δ represents the lower bound on the probability of exposure to that hospital (g); we display odds ratios and marginal risks for two versions of the TMLE model with varying levels of δ .

Hsp.					**Logistic regression**		**TMLE ($\delta = 10^{-2}$)**		**TMLE ($\delta = 10^{-2.5}$)**	
	Admitted	Died	Discharged	Readmitted	Odds ratio	Marginal	Odds ratio	Marginal	Odds ratio	Marginal
		(%)		(%)	(95% CI)	risk	(95% CI)	risk	(95% CI)	risk
1	1229	141 (11)	1088	248 (23)	1.00 (0.97-1.03)	0.22	0.61 (0.59-0.63)	0.11	0.50 (0.48-0.53)	0.09
2	2071	166 (8)	1905	441 (23)	1.02 (0.99-1.05)	0.24	1.13 (1.11-1.16)	0.19	1.13 (1.11-1.16)	0.19
3	1243	134 (11)	1109	285 (26)	1.03 (1.00-1.07)	0.25	0.71 (0.69-0.72)	0.13	0.52 (0.50-0.54)	0.10
4	1076	122 (11)	954	214 (22)	1.01 (0.97-1.04)	0.23	1.06 (1.04-1.09)	0.18	0.92 (0.89-0.96)	0.16
5	1550	181 (12)	1369	288 (21)	0.99 (0.96-1.02)	0.21	0.71 (0.69-0.72)	0.13	0.58 (0.56-0.60)	0.11
6	827	107 (13)	720	128 (18)	0.97 (0.94-1.00)	0.19	0.73 (0.70-0.75)	0.13	1.08 (1.03-1.14)	0.18
7	2917	386 (13)	2531	666 (26)	1.04 (1.02-1.07)	0.26	1.63 (1.61-1.66)	0.25	1.67 (1.64-1.71)	0.26
8	1456	197 (14)	1259	232 (18)	0.97 (0.94-1.00)	0.19	0.72 (0.70-0.74)	0.13	0.68 (0.66-0.70)	0.12
9	881	111 (13)	770	157 (20)	0.98 (0.95-1.02)	0.20	1.27 (1.25-1.29)	0.21	1.18 (1.16-1.20)	0.20
10	1410	149 (11)	1261	311 (25)	1.01 (0.99-1.05)	0.23	0.66 (0.65-0.68)	0.12	0.57 (0.55-0.60)	0.11
11	1297	153 (12)	1144	258 (23)	1.01 (0.98-1.04)	0.23	0.90 (0.88-0.92)	0.16	0.86 (0.83-0.88)	0.15
12	1323	162 (12)	1161	192 (17)	0.92 (0.89-0.95)	0.13	0.79 (0.76-0.81)	0.14	0.76 (0.74-0.78)	0.14
13	1231	102 (8)	1129	262 (23)	1.00 (0.97-1.03)	0.22	0.94 (0.93-0.96)	0.16	0.91 (0.87-0.95)	0.16
14	2110	234 (11)	1876	424 (23)	(Reference)	0.22	(Reference)	0.17	(Reference)	0.17
15	1389	190 (14)	1199	203 (17)	0.97 (0.94-1.00)	0.19	0.74 (0.72-0.77)	0.13	0.81 (0.79-0.84)	0.14
16	681	94 (14)	587	111 (19)	0.98 (0.94-1.01)	0.20	0.75 (0.73-0.78)	0.14	0.84 (0.80-0.87)	0.15
17	1438	139 (10)	1299	328 (25)	1.04 (1.01-1.07)	0.26	1.50 (1.48-1.53)	0.24	1.85 (1.80-1.90)	0.28
18	1984	212 (11)	1772	438 (25)	1.03 (1.00-1.06)	0.25	0.76 (0.74-0.77)	0.14	0.74 (0.72-0.76)	0.13
19	932	99 (11)	833	163 (20)	0.98 (0.95-1.01)	0.20	0.88 (0.86-0.90)	0.16	0.81 (0.79-0.84)	0.14
20	1048	167 (16)	881	171 (19)	0.99 (0.96-1.02)	0.21	1.25 (1.22-1.27)	0.21	1.20 (1.17-1.23)	0.20

Table 2: Risk of 30-day readmission after admission for heart failure in twenty Montreal hospitals. The columns in this table are described in Table 1.

Hsp.					**Logistic regression**		**TMLE ($\delta = 10^{-2}$)**		**TMLE ($\delta = 10^{-2.5}$)**	
	Admitted	Died	Discharged	Readmitted	Odds ratio	Marginal	Odds ratio	Marginal	Odds ratio	Marginal
		(%)		(%)	(95% CI)	risk	(95% CI)	risk	(95% CI)	risk
1	1184	176 (15)	1008	159 (16)	1.00 (0.98-1.03)	0.15	1.23 (1.18-1.27)	0.15	1.21 (1.17-1.26)	0.15
2	199	11 (6)	188	31 (16)	1.02 (0.97-1.08)	0.17	1.09 (1.07-1.12)	0.14	1.25 (1.17-1.34)	0.16
3	1085	132 (12)	953	160 (17)	1.01 (0.98-1.04)	0.16	0.83 (0.80-0.87)	0.11	0.82 (0.78-0.87)	0.11
4	863	91 (11)	772	113 (15)	1.00 (0.97-1.03)	0.15	0.85 (0.81-0.88)	0.11	0.84 (0.81-0.88)	0.11
5	923	147 (16)	776	143 (18)	1.04 (1.01-1.07)	0.19	0.96 (0.93-1.00)	0.12	0.95 (0.91-0.99)	0.12
6	788	136 (17)	652	89 (14)	1.00 (0.96-1.03)	0.14	0.89 (0.85-0.94)	0.12	0.91 (0.86-0.96)	0.12
7	2194	228 (10)	1966	328 (17)	1.03 (1.00-1.05)	0.17	1.33 (1.29-1.37)	0.16	1.33 (1.29-1.37)	0.16
8	1485	243 (16)	1242	173 (14)	0.99 (0.97-1.02)	0.14	0.97 (0.93-1.01)	0.12	0.97 (0.94-1.01)	0.13
9	990	166 (17)	824	158 (19)	1.04 (1.01-1.08)	0.19	1.30 (1.25-1.35)	0.16	1.28 (1.23-1.33)	0.16
10	1214	139 (11)	1075	181 (17)	1.01 (0.99-1.04)	0.16	1.45 (1.40-1.51)	0.18	1.46 (1.40-1.51)	0.18
11	892	147 (16)	745	119 (16)	1.02 (0.98-1.05)	0.16	1.39 (1.34-1.44)	0.17	1.40 (1.35-1.46)	0.17
12	1102	185 (17)	917	91 (10)	0.96 (0.93-0.98)	0.10	0.47 (0.44-0.50)	0.06	0.47 (0.44-0.50)	0.06
13	1914	204 (11)	1710	325 (19)	1.03 (1.00-1.05)	0.18	0.80 (0.77-0.83)	0.10	0.84 (0.79-0.89)	0.11
14	1980	278 (14)	1702	263 (15)	(Reference)	0.15	(Reference)	0.13	(Reference)	0.13
15	1365	179 (13)	1186	163 (14)	1.00 (0.97-1.03)	0.15	0.86 (0.83-0.90)	0.11	0.85 (0.81-0.89)	0.11
16	541	77 (14)	464	46 (10)	0.96 (0.93-1.00)	0.11	1.45 (1.39-1.52)	0.18	1.55 (1.46-1.65)	0.19
17	1338	163 (12)	1175	193 (16)	1.02 (0.99-1.05)	0.17	0.86 (0.83-0.89)	0.11	0.79 (0.76-0.82)	0.10
18	1356	168 (12)	1188	200 (17)	1.02 (0.99-1.04)	0.17	1.40 (1.36-1.44)	0.17	1.40 (1.35-1.44)	0.17
19	1020	123 (12)	897	122 (14)	0.99 (0.96-1.02)	0.14	0.94 (0.90-0.98)	0.12	0.98 (0.93-1.03)	0.13
20	1152	171 (15)	981	126 (13)	0.98 (0.95-1.01)	0.13	1.11 (1.07-1.16)	0.14	1.11 (1.07-1.16)	0.14

Table 3: Risk of 30-day readmission after admission for pneumonia in twenty Montreal hospitals. The columns in this table are described in Table 1.

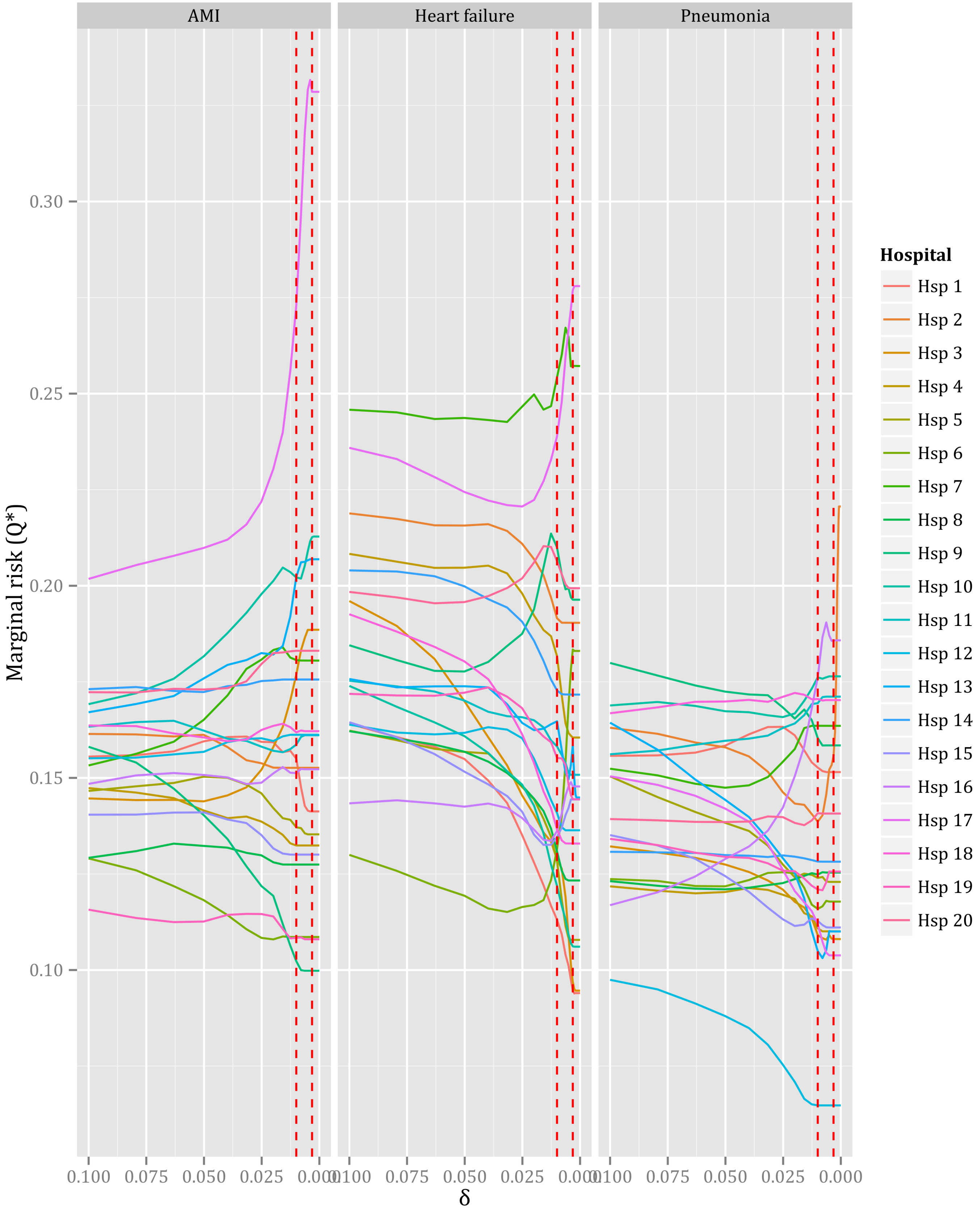

Figure 2: Effect of δ (the bound on $g(A|W)$) on the marginal risk ($Q*$). The vertical axis represents the marginal risk as calculated by the TMLE model. The marginal risk (Q^*) was evaluated at 31 levels of δ, from 10^{-1} to 10^{-5}, (the exponent decreasing by 0.1). Note that the scale of the horizontal axis decreases from left-to-right. The hatched vertical lines mark the two levels of delta displayed in tables 1,2 and 3.

4 Discussion

Using targeted maximum likelihood estimation (TMLE) to adjust precisely for measured confounders, we found that the differences in marginal risk of 30-day hospital readmission in twenty Montreal hospitals were much stronger than a model that only crudely adjusted for confounding readmission risk suggested. Additionally, our study revealed some practical positivity violations for some hospitals, suggesting that the relative readmission risk may not always be estimaable from observed data.

Our study have several strengths.

By using a machine learning technique (random forest) to adjust accurately for confounding, and a doubly-robust technique to estimate the parameters (TMLE), we could estimate less biased estimates

Our work suggests that the difference between the approaches is not only a practical one; in assessing the effect of hospitals on readmission, it has a substantive effect as well.

Some other studies have applied machine learning algorithms to readmission data to develop predictive models, including one using the data used in this study. No study to our knowledge has used machine learning algorithms to infer target parameters.

Some authors have suggested that

Our study demonstrates that this is not the case; despite a weak predictive accuracy, we can still draw inference on target parameters.

Some authors[9] claim that by using readmission rates as a quality metric, we assume that readmissions are both preventable and related to poor quality care. Some authors[10] have also claimed that most readmissions are not preventable.

Typically, people try to identify whether individual readmissions are preventable.

In this study, we define a preventable readmission as one that would not have occured if the patient was treated at a different hospital (the unobservable counterfactual).

To estimate this parameter, it is unnecessary to identify whether individual readmissions were preventable, or to identify the mechanism of readmission; we need only to measure the readmission rate at a population level, and adjust for confounding.

Several readmission risk models have been criticized on the basis that they failed to distinguish preventable readmissions from readmissions due to chance alone. Despite evidence that clincians cannot reliably distinguish preventable and non-preventable readmissions, pairs of diagnosis codes have been developed that are "potentially preventable". For example, following a hospitalization for surgery, a readmission for infection would be considered preventable, but a readmission for trauma would not.

We did not atempt to measure whether *individual* readmissions were preventable, we instead opted to place readmissions within a counterfactual framework, and only identify the difference in readmission risk between hospitals, after controlling for differences in case-mix.

The counterfactual model clarifies the study question: What is the difference in the proportion with an emergency readmission within 30 days if all patients had attended Hospital A vs Hospital B? The notion of a "preventable" readmission is implicit; if a patient would have been readmitted if treated at another hospital, then the readmission was preventable. Since we cannot directly observe the desired proportions, we estimate it by attempting to recreate exchangeability (control for confounding) among the populations that visited different hospitals.

Our results were sensitive to the bounding parameter; the more precise we allowed the g model to be, the more divergent the odds ratios.

This suggests a practical positivity violation; some hospitals may not be admitting certain types of patients with high readmission risk, making it difficult to estimate what the risk of those patients would be if they had attended that hospital.

We can either bound the probability predicted by g, or we can use the probabilities directly. If we use the probabilities directly, then the pseudopopulations created for certain hospitals will be dominated by patients with very low probability of attending that hospital. Since a probability of 0.0001 is not uncommon, that person would count for 100 times the weight of a person with only a 0.01 probability of attending that hospital.

Unfortunately, our model for g is simply not likely to be that precise: a very small absolute difference in probability will make a big difference in the pseudopopulation. However, this only strengthens the argument that we need very well calibrated and accurate predictive models of g to estimate the parameter of interest. Machine learning models like random forest are best suited to do that.

Additionally, it is good that we identify these positivity violations. It may be the case that certain groups of hospitals are simply not comparable, because the patient populations they receive have such varying risk.

Because readmissions are relatively common, a small difference on a relative scale can imply a large difference on an absolute scale.

Following other hospital readmission work, we did not include hospital length of stay as a risk factor for readmission. Although we have strong reason to believe that the length of stay may affect readmission risk (and varies between hospitals), it acts as a mediator between hospital care and readmission; if we included it in our models, we risk biasing our estimate of the hospital's effect on readmission. However, unlike many other hospital readmission studies, we also excluded all diagnoses and procedures that occurred during the hospital admission, because these covariates were also effectively mediators between hospital care and readmission. For example, hospital-acquired infections are one source of preventable readmissions; if we controlled for diagnosis codes indicating hospital-acquired infections, we would attenuate the effect that the hospital care was having on readmissions.

The major competing risk for hospital readmission is death, but others include moving outside the study area, or admission to a hospital for a non-emergency reason. In our analysis, we did not account for these competing risks. If patients died within 30 days of discharge more often at one hospital than another, we could have biased our estimate of readmission risk. Similarly, if patients died during the hospitalization more often at some hospitals than others, it could have created a selection bias (left censorship) in which hospitals with better care were discharging sicker (but still living) patients, who would be more likely to be readmitted. Also, there is no special significance of 30 days in readmission, except for the fact that it is (recently) widely used as a cutoff. In future work, we plan to account for both left censorship and competing risks in a model that estimates the effect of hospital care on time-to-readmission.

Differing *admission* practices can strongly affect the rates of readmission. Since most (89% in this cohort) patients who are readmitted within 30 days are readmitted to the same hospital that they were discharged from, a hospital that is more likely to admit patients will have a higher readmission rate. In some cases, like a major trauma, admission is certain, but in most cases, there is some variation in practice of how patients are admitted. In future work, we plan to study the effect of the probability of admission on readmission.

Entry to our cohort was dependent on having one diagnosis of a respiratory illness in an inpatient or outpatient setting. Respiratory illness was defined rather broadly, including extremely common diagnoses such as "cough". We expect that the the majority of 65-year old patients who would be hospitalized would have at least one respiratory illness diagnosis in an outpatient setting. We cannot, however, exclude the possibility that parameters were affected by selection bias with respect to the full population of 65 year old patients.

The effect of hospital care on readmissions is confounded by a vast spectrum of health-related states of the admitted patients. In the absence of a clear theoretical basis of the structure of that confounding, we can 1) identify relatively few, well-understood and measureable confounders to include in our model, or 2) forgo any theoretical understanding of the structure of confounding, and attempt to identify the broadest measureable set of even faintly plausible confounders.

The first option has some advantages: in a situation where data collection is expensive, we it may not be plausible to measure thousands of variables. Additionally, by reducing the confounders to a well-understood few, the model gains credibility because it can be shown that the confounders are having the expected effect. Non-parametric techniques such as random forest don't allow us to look (easily) at the individual effects of the confounders, and even in a parametric model it would be difficult to analyze thousands of variables. We summarized the densities of the effects a few classes of variables in figure 1, but this still does not allow the variable-by-variable analysis typical in an epidemiologic study. Also, by including many confounders we also risk *inducing* bias, such as the M-bias.

However, the recent availability of large scale healthcare administrative data has put us into the situation where the cost of data collection is relatively low. By using machine learning techniques like random forest, we also automatically fit multi-way interactions that we would be unlikely to explore in a model fit "by hand". Finally, because the structure of the confounding is unclear, we cannot assess M-bias is present. We argue that in this situation, where we have a large dataset, thousands of measureable confounders, and little understanding of the structure of confounding, the second option is more appropriate.

Despite a relatively standardized data collection process, some hospitals may have idiosyncratic code usage patterns, leading to differing specificity and sensitivity of some diagnostic and procedural codes. This possible differential misclassification could have biased our estimate of the parameters of interest.

The unit of analysis in this study was the discharge, but each discharge was "clustered" within a patient. The expected within-cluster homogeneity could have biased our estimates of variance, and even our parameter estimates. However, because the number of clusters (unique patients) was relatively high when compared to the sample size (the number of discharges), we do not expect that our parameter or variance estimates to be biased very strongly.

Beside random forest, we could have used many other machine learning techniques on these data, many of which we explored in other work. Also, some ensemble machine learning techinques, (in particular SuperLearner which is commonly used with TMLE), are available, that combine any number of other machine learning techniques. We found that in these data, ensemble learning techniques were too computationally expensive. We selected random forest because of its relative simplicity, and because our variables were nearly all binary, for which decision trees are particularly suitable.

Calibration of the random forest vote proportions in the Q model strongly affected estimates of our parameter of interest in the Q^* update step. Other articles using non-parametric techniques typically combined them with other models in an ensemble learner (like SuperLearner). The final step in (many) ensemble learners is to combine all the probability estimates in parametric model, which would effectively calibrate the probabilities. In our study, a single, non-parametric technique was used, so an additional, seperate calibration step was necessary to convert the ranking scores into a probability estimate.

References

[1] Carl van Walraven and Chaim M Bell. ``Risk of death or readmission among people discharged from hospital on Fridays''. In: *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne* 166.13 (June 2002), pp. 1672–1673. ISSN: 0820-3946.

[2] Leo Breiman. ``Random Forests''. English. In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[3] Corrado Gini. *Variabilità e Mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche.* Italian. Bologna: C. Cuppini, 1912.

[4] R. Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014. URL: http://www.R-project.org/.

[5] Aloysius Lim, Leo Breiman, and Adele Cutler. *bigrf: Big Random Forests: Classification and Regression Forests for Large Data Sets*. R package version 0.1-9. 2014. URL: https://github.com/aloysius-lim/bigrf.

[6] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. ``Regularization Paths for Generalized Linear Models via Coordinate Descent''. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. ISSN: 1548-7660. URL: `http://www.jstatsoft.org/v33/i01`.

[7] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN: 978-0-387-98140-6. URL: `http://had.co.nz/ggplot2/book`.

[8] Bernard W Silverman. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986.

[9] Devan Kansagara et al. ``Risk prediction models for hospital readmission: a systematic review''. In: *JAMA: The Journal of the American Medical Association* 306.15 (Oct. 2011), pp. 1688–1698. ISSN: 1538-3598. DOI: `10.1001/jama.2011.1515`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/22009101` (visited on 11/27/2011).

[10] Karen E. Joynt and Ashish K. Jha. ``Thirty-Day Readmissions — Truth and Consequences''. In: *New England Journal of Medicine* 366.15 (Mar. 2012), pp. 1366–1369. ISSN: 0028-4793. DOI: `10.1056/NEJMp1201598`. URL: `http://dx.doi.org/10.1056/NEJMp1201598` (visited on 10/01/2012).

5 Appendix

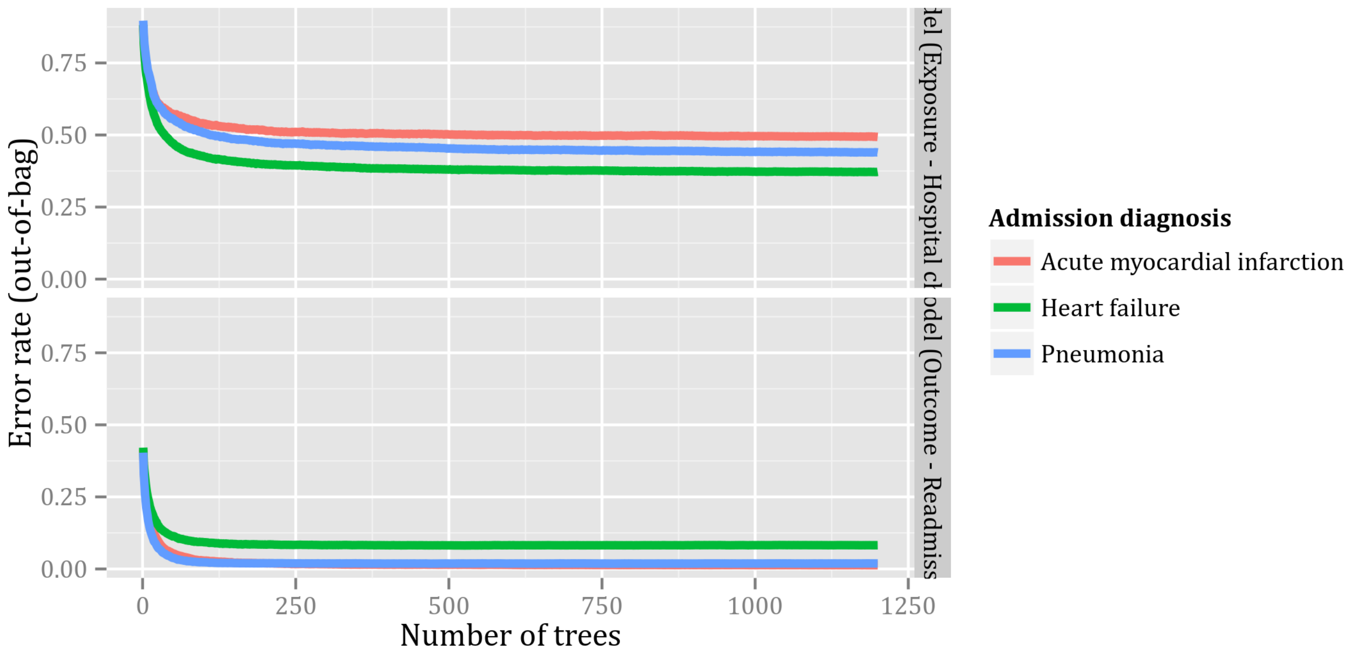

Figure 3: Error rate for both random forest models of hospital choice (g) and readmission (Q) as a function of the number of trees grown. For each admission, only out-of-bag trees were used to predict the given outcome.

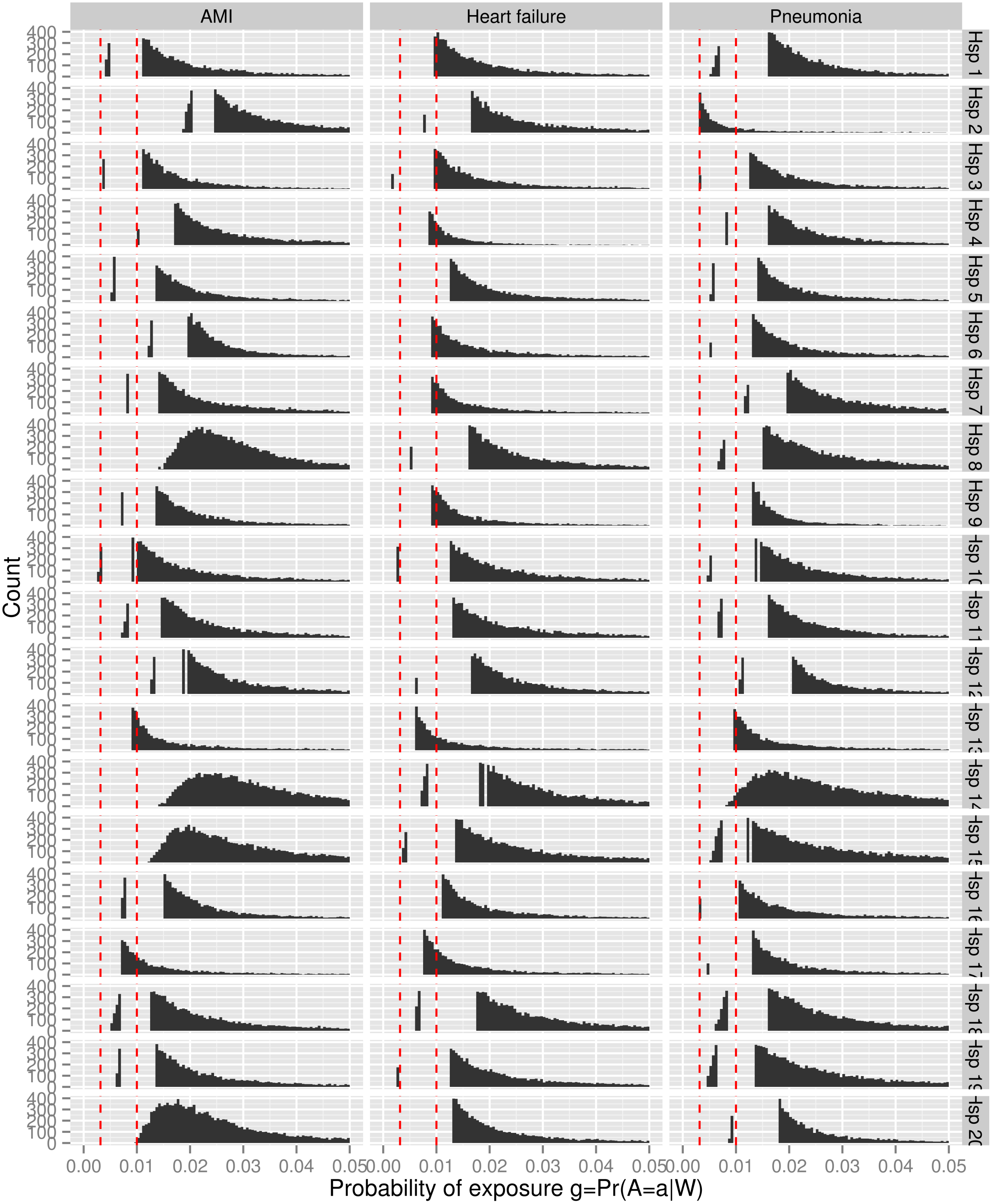

Figure 4: Histogram of the probability of exposure (g), restricted to the range of (0,0.05). The bin width is 0.005. The dotted red lines indicate the two values of δ used in tables 1, 2, and 3.