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.

It is particularly important to accurately estimate the independent effect of hospital treatment on readmission because it is being used to financially penalize hospitals.

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".

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.

2 Methods

2.1 Data

2.1.1 Cohort selection

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.

2.1.2 Hospital readmissions

The unit of analysis in all models 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.

2.1.3 Disease types

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 Confounders

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 study period), 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, 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.

2.3 Descriptive analysis

2.3.1 Choropleth

We believed that census tract of residence would strongly affect the probability of admission to the hospital nearest that census tract. We plotted choropleths of the rate of attendance at the twenty hospitals by census tract. The numerator was the number of live discharges at the hospital, and the denominator was the number of person-years accumulated in that census tract of residence by cohort members when their admissions would have been eligible (after the first continuous year within the cohort).

2.4 Models

2.4.1 Model g - probability of exposure

We developed a multinomial model g that predicted the probability of attending each of the twenty hospitals (the exposure) as a function of all the confounders. To fit g we used a random forest, a non-parametric model based on decision trees[1].

Our trees were decision stumps; we only used one splitting node at in each tree. We arbitrarily chose to grow 1200 trees (decision stumps), 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. 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 replicate did not include the discharge.

To assess the importance of the variables in predicting which hospital a patient will choose, for each tree, we calculate the increase in homogeneity of classes between the root of the tree and the leaves. To measure the homogeneity of classes, we repurpose a metric that is typically used to measure equality (homogeneity) of income, the Gini coefficient[2]. The homogeneity is defined as the sum of squared proportions in each class (hospital), with a maximum of 1 (all discharges at the same hospital) and a minimum of 1/20 (the discharges were evenly divided between the 20 hospitals). If the patients within the leaves of the tree made relatively homogeneous hospital choices, this variable predicts hospital choice well.

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. We multiplied each proportion used to calculate the Gini coefficient by this weight.

Random forest traditionally classifies each item by majority vote; we converted this into a probability by taking the proportion of votes for each hospital (using only out-of-bag trees for each discharge).

2.4.2 Q model - probability of outcome

For the Q model, we developed another random forest model, very similar to the g model, except that instead of predicting the choice of hospital, we directly predicted 30-day readmission. We also included a set of 19 indicator variables for the hospital attended. We also calibrated this model based on the inverse of the proportion of those readmitted.

We also fit a regularized logistic regression model for Q, using cyclical coordinate descent to estimate the parameters efficiently in our sparse but large matrix. We penalized the likelihood by the ℓ_2 norm, that is, the sum of the squares of the normalized regression parameters. The scale of the penalty was determined by the parameter λ. We optimized the selection of the penalty scale for the best partial likelihood using a nested a 10-fold cross-validation. Within each fold we assessed 100 λ-values spaced evenly between max(λ)×10-4 and max(λ), where max(λ) was the smallest λ-value that would result in a model with no non-zero coefficients.

2.4.3 Variable importance

For each tree, we calculated the decrease in Gini homogeneity when comparing the split nodes of the tree to the root. The Gini impurity at any node is, for all out-of-bag discharges.

For each hospital, the proportion of patients that chose this hospital p_h, multiplied by the probability of guessing that this patient would choose this hospital based on the distribution of patients $1 - p_h$. Summed over all hospitals, this is the Gini impurity. Gini impurity is 1 when all patients chose the same hospital.

Gini is defined as "inequity" when used in describing a society's distribution of income, or a measure of "node impurity" in tree-based classification. 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.

2.4.4 Model Q* - updated targeted maximum likelihood estimation

To avoid convergence problems for this one parameter model, we fit the parameter using a quasi-Newton method simultaneously discovered by Broyden[3], Fletcher[4], Goldfarb[5] and Shanno[6] (BFGS).

2.5 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 3.1.0)[7]. We implemented the random forest using the "bigrf" package (version 0.1.9)[8]. We implemented the GLM fitting using coordinate descent using the "glmnet" package (version 1.9.5)[9]. We plotted our figures using the "ggplot2" package (version 1.0.0)[10].

3 Results

4 Discussion

References

[1] 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.

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

[3] C. G. Broyden. ``The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations''. en. In: *IMA Journal of Applied Mathematics* 6.1 (1970), pp. 76–90. ISSN: 0272-4960, 1464-3634. DOI: 10.1093/imamat/6.1.76. URL: http://imamat.oxfordjournals.org/cgi/doi/10.1093/imamat/6.1.76 (visited on 06/02/2014).

[4] R. Fletcher. ``A new approach to variable metric algorithms''. en. In: *The Computer Journal* 13.3 (Mar. 1970), pp. 317–322. ISSN: 0010-4620, 1460-2067. DOI: 10.1093/comjnl/13.3.317. URL: http://comjnl.oupjournals.org/cgi/doi/10.1093/comjnl/13.3.317 (visited on 06/02/2014).

[5] Donald Goldfarb. ``A family of variable-metric methods derived by variational means''. en. In: *Mathematics of Computation* 24.109 (Jan. 1970), pp. 23–23. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-1970-0258249-6. URL: http://www.ams.org/jourcgi/jour-getitem?pii=S0025-5718-1970-0258249-6 (visited on 06/02/2014).

[6] D. F. Shanno. ``Conditioning of quasi-Newton methods for function minimization''. en. In: *Mathematics of Computation* 24.111 (Sept. 1970), pp. 647–647. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-1970-0274029-X. URL: http://www.ams.org/jourcgi/jour-getitem?pii=S0025-5718-1970-0274029-X (visited on 06/02/2014).

[7] 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/.

[8] 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.

[9] 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.

[10] 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.