1 Introduction

2 Methods

2.1 Data

2.1.1 Cohort selection

We sampled 18.60% of the *Régie de l'assurance maladie du Québec* (RAMQ) insured population that whose home address had a postal code within the Montreal 2006 Census Metropolitan Area (CMA) over the period of 1998-2006. The vast majority of the Montreal population is insured by RAMQ. We did so by randomly sampling 18.60% of the population that was ever insured in 1998. We then sampled 18.60% of patients that became newly insured in 1999 (they were born in 1999 in Montreal, they became RAMQ insured in 1999, or they were previously RAMQ insured but recently moved to Montreal). We repeated this for every year up to 2006, retaining each sampled person in the cohort until they no longer had an address in Montreal or they died. This sampling scheme resulted in an open, dynamic cohort, that was a random sample of 18.60% of the Montreal population during any year.

2.2 Error rate

For each tree, we calculated the out-of-bag decrease in Gini impurity 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.

To assess the importance of the variables in predicting which hospital a patient will choose, for each tree, we calculate the decrease in Gini impurity between the root of the tree and the leaves. If the patients within the leaves of the tree made relatively homogeneous hospital choices, this variable predicts hospital choice well, and the Gini impurity will decrease.

To calculate the Gini variable importance criterion, we summed the Gini impurity decrease over all trees.

2.3 Software

We implemented the random forest using the "bigr" [1] package . We implemented the GLM fitting using coordinate descent using the "glmnet" package [2]. We plotted our figures using the "ggplot2" package [3].

3 Results

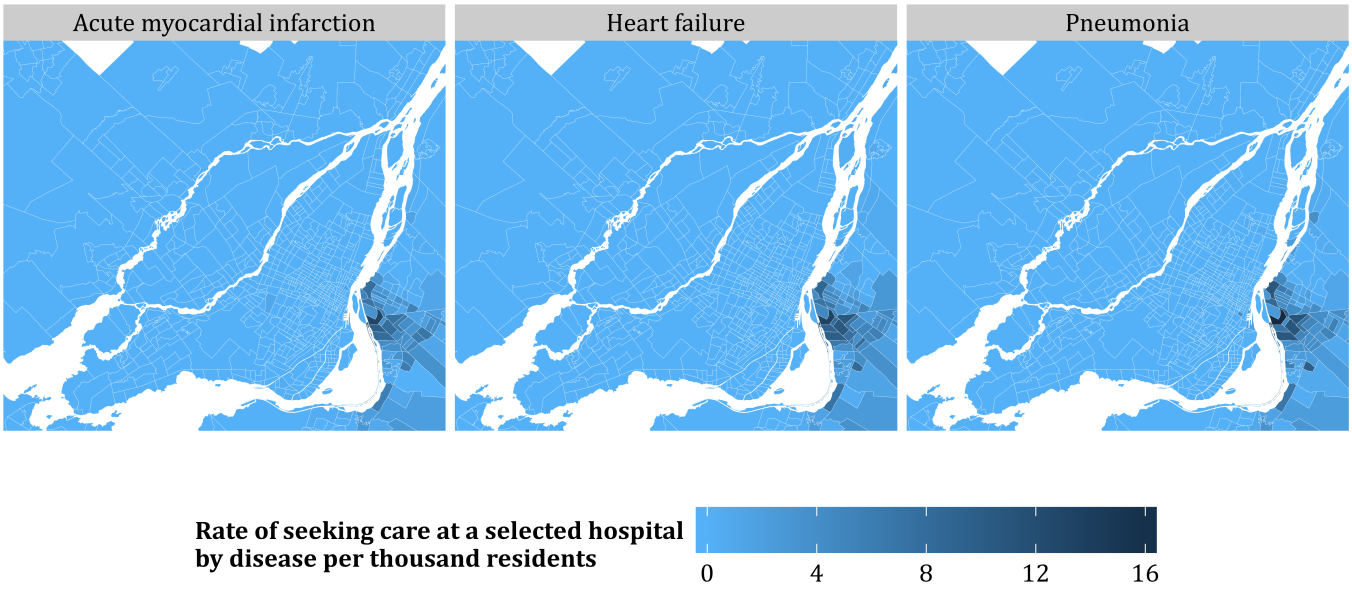

Figure 1: Rates of seeking care at a selected hospital by admission diagnosis. A descriptive sentence

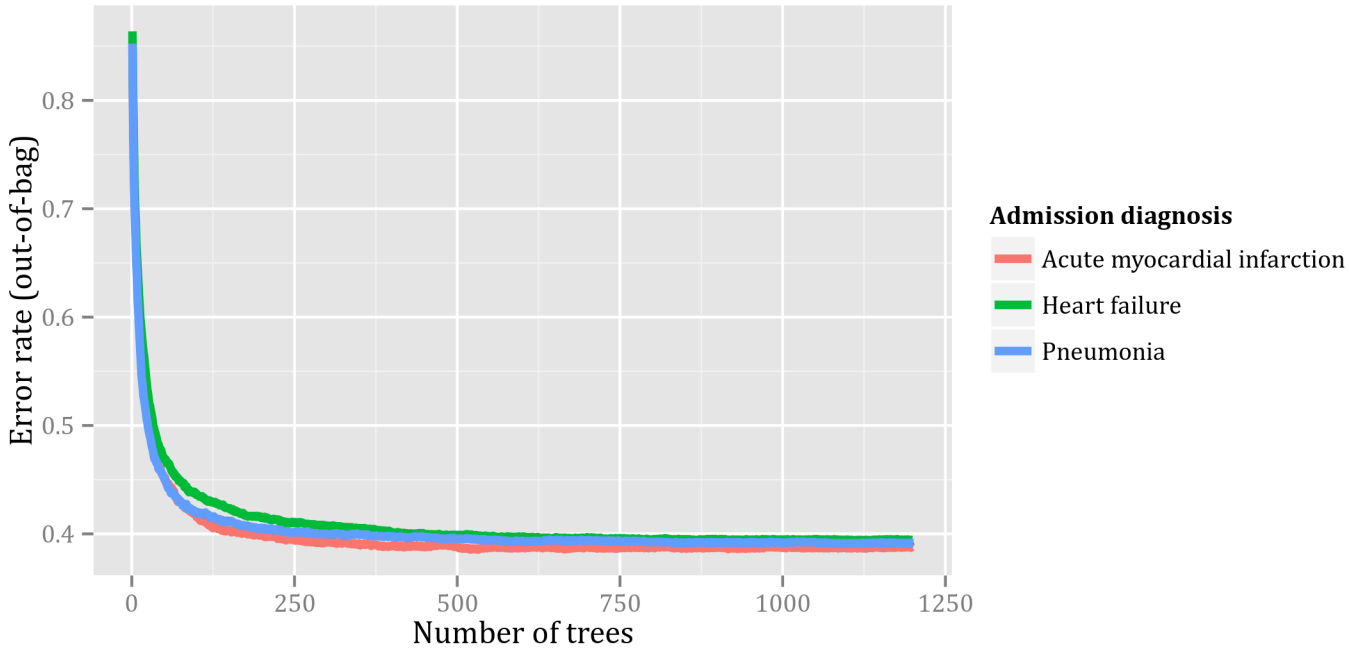

Figure 2: Error rate for random forest model of hospital choice. A descriptive sentence.

Figure 3: 10 most important variables for the G model, by disease. A descriptive sentence.

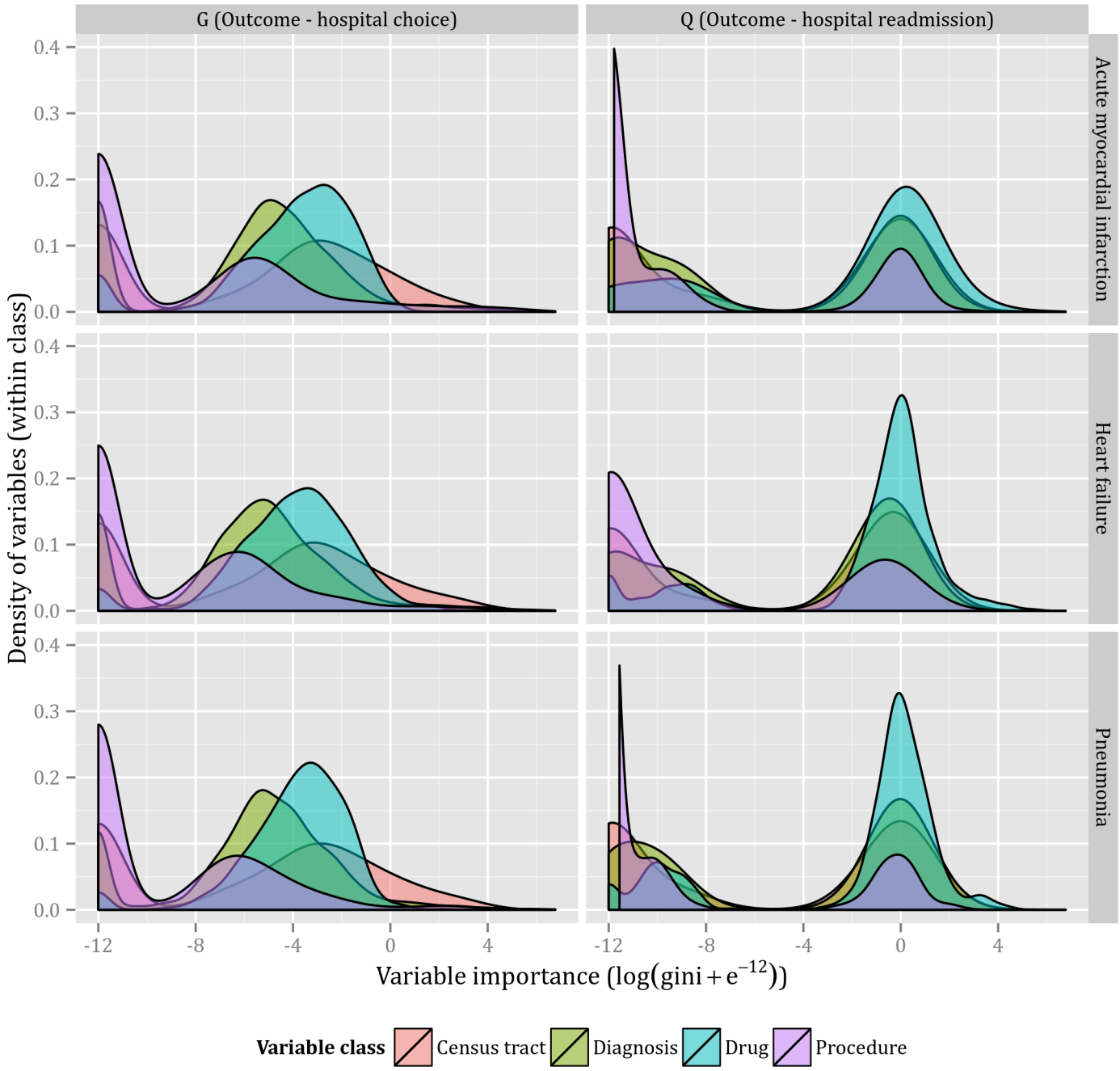

Figure 4: Variable importance by model and variable class. A descriptive sentence.

4 References

References

[1] 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`.

[2] 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), 1–22. ISSN: 1548-7660. URL: `http://www.jstatsoft.org/v33/i01`.

[3] 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`.