Chaire de recherche Industrielle Alliance
sur les enjeux économiques
des changements démographiques
CEDIA.ca

17-05

# Hospital Readmission is Highly Predictable from Deep Learning

CAHIER DE RECHERCHE
WORKING PAPER

Damien Échevin, Qing Li and Marc-André Morin

Décembre / December 2017

UNIVERSITÉ LAVAL
Faculté des sciences sociales
HEC MONTRÉAL
ESG UQÀM

La Chaire de recherche Industrielle Alliance sur les enjeux économiques des changements démographiques est une chaire multi-institutionnelle qui s'appuie sur un partenariat avec les organisations suivantes :

**- Centre interuniversitaire de recherche en analyse des organisations (CIRANO)**
**- iA Groupe financier**
**- Retraite Québec**





Dépôt légal : Bibliothèque et Archives nationales du Québec et Bibliothèque et Archives Canada, 2017.
ISSN 2368-7207

# Hospital Readmission is Highly Predictable from Deep Learning

Damien Échevin[*†], Qing Li[†] and Marc-André Morin[†]

December 2017

**Abstract**

Hospital readmission is costly and existing models are often poor or moderate in predicting readmission. We sought to develop and test a method that can be applied generally by hospitals. Such a tool can help clinicians identify patients who are more likely to be readmitted, either at early stages of hospital stay or at hospital discharge. Relying on state-of-the art machine learning algorithms, we predict probability of 30-day readmission at hospital admission and at hospital discharge using administrative data on 1,633,099 hospital stays from Quebec between 1995 and 2012. We measure performance of the predictions with the area under receiver operating characteristic curve (AUC). Deep Learning produced excellent prediction of readmission province-wide, and Random Forest reached very similar level. The AUC for these two algorithms reached above 78% at hospital admission and above 87% at hospital discharge, and the diagnostic codes are among the most predictive variables. The ease of implementation of machine learning algorithms, together with objectively validated reliability, brings new possibilities for cost reduction in the health care system.

**Keywords:** Machine learning; Logistic regression; Risk of re-hospitalisation; Healthcare costs.
**JEL codes**: I10; C52; C55.

[*]CRCHUS, Université de Sherbrooke and Université Laval; e-mail: damien.echevin@usherbrooke.ca. (Corresponding author).
[†]Apexmachina.

# 1 Introduction

Hospital readmissions contribute to a significant proportion of total inpatient spending among the many cost drivers of healthcare. Within the literature, readmission often refers to hospital admissions within 30 days following the initial discharge and can occur at either the same hospital or a different hospital (Yu et al. 2015; Stone and Hoffman 2010). It has been documented that readmission rates are associated with age, patient comorbidities, and many other factors such as diagnostics or length of stay in the hospital (see e.g. Wolff 2002, Kind 2007, Pham 2007, Krumholz, Normand, and Keenan 2008abc, Au et al. 2012, Wang et al. 2014, Yu et al. 2015).

Despite the importance of predicting readmission, most existing works have poor or moderate predictive results that prevent more general application of the methods (Kansagara et al. 2011). As an example, the LACE index is sometimes applied to score the risk of readmission in clinical settings (Length of stay, Acuity of the admission, Charlson comorbidity index score and Emergency department use; Walraven et al. 2010; Gruneir et al. 2011); yet this index is not strong in predicting 30-day readmission (Cotter et al. 2012). The Area Under receiver operating characteristic Curve (AUC, Hanley and McNeil 1982) is a standard measure of prediction accuracy. Generally speaking, an AUC of 0.5 indicates that the model is no better than chance; an AUC of 0.7 to 0.8 indicates modest or acceptable discriminative ability, and a threshold of greater than 0.8 indicates good discriminative ability (Kansagara et al. 2011, Schneeweiss et al. 2001, Ohman et al. 2000). The probability of readmission can be estimated at early stages during a hospitalization in order to identify high-risk patients for intervention. Similarly, the risk can be estimated at hospital discharge. Kansagara et al. (2011) reviewed the literature systematically and found seven studies whose results could potentially be used to predict hospital readmission at early stages during a hospitalization (AUC from 0.56 to 0.72). They also found five studies that could be used to predict readmission at the discharge from hospital (AUC from 0.68 to 0.83). The highest AUC of 0.83 was obtained in Coleman et al. (2004) with a relatively small dataset where the authors combined administrative information with self-reported health information from survey data; meanwhile, the AUC decreased to around 0.77 without the information from survey data. The AUC increased a little bit in more recent works, but the increase is often resulted from more specific approaches. For instance, Yu et al. (2015) applied institution-specific prediction models on three different hospitals using supporting vector machine and Cox regression algorithms; the AUC reached 0.85 for hospital one but was lower for hospital two and hospital three (around 0.67). They concluded following the experiments that a possible way to implement the prediction of readmission is to use hospital-specific models with hospital specific information.

The goal of this paper is to explore possibilities to increase predictability of hospital readmission for practical use. With a big administrative database that is representative at least for Quebec, we try to derive a general tool that can be implemented easily by hospitals. Machine learning is a fast-growing research area. Performance of machine learning algorithms can be tested objectively by cross-validation. In our context, medical codes (e.g., DRG) are often factor variables with many different categories, making the feature space large and sparse. Classical statistical models such as the logistic regression are not particularly designed to handle this type of data (see, e.g., Tan et al. 2010, Conway and White 2012), and state-of-the-art machine learning algorithms provide another possibility to predict readmission with all this information.

The paper is organized as follows. In part 2 we describe data and variables. In part 3 we present study design and methodological approach. Part 4 presents the results and discuss the implications. Finally we conclude in part 5.

# 2 Data and Variables

## 2.1 Datasets

Predicting readmissions requires the use of two administrative files. We use the MED-ECHO and RAMQ databases between 1995 and 2012. We have access to a subsample of all patients in Quebec. Individuals are included only if they were born on an odd year in April or October, which accounts for almost one-twelfth of the hospitalized population in Quebec. This constraint is imposed by the RAMQ, which does not allow the creation of data files with more than 135,000 cases in any given year. The available RAMQ sample includes patients hospitalized at least once between 1995 and 2012. The RAMQ merged the two data sets based on the health insurance number of each individual. Each individual is followed during the entire period (1995-2012). For this study, we delete the hospital stays if the patient died within the hospital, since dead patients cannot be readmitted. There remain 1,633,099 hospital stays after deleting such entries.

### 2.1.1 MED-ECHO

The MED-ECHO database includes all the information related to hospital stays and day surgeries in Quebec. Therefore, the only type of outpatient care included in our data set is day surgeries performed in hospital. All other types of consultations, procedures and physical examinations that

are performed outside hospitals are not included in our data set. It also does not include events related to psychiatric hospitals, rehabilitation hospitals, long-term care facilities and physicians operating in those settings. For each hospitalisation, the entry and exit dates of the patient (and therefore the length of stay) are available, as well as an indicator of death within the hospital. Matched to MED-ECHO is a database containing all diagnostics (using International Classification of Diseases, ICD-9 or ICD-10) observed during the hospitalisation. Each hospitalisation is associated with an APR-DRG, which is a classification using all available information during the stay to group patients who use a similar amount of resources. This APR-DRG measure includes a DRG (diagnostic-related group) code, a gravity code indicating the severity of the condition within the DRG, as well as a mortality code which indicates the probability of death. DRG can also be grouped in MDC (Major Diagnostic Categories).

### 2.1.2 RAMQ Services

Data regarding medical services billed to the RAMQ allow us to capture the remuneration of the vast majority of Quebec physicians. The RAMQ databases include all reimbursement demand forms filled in by health professionals who receive a fee for each service provided. Physicians paid through fee-for-service or blended payments must fill out a form for each act, which includes the service code, amount of reimbursement demanded (according to the Health Ministry guidelines), moment at which the act was executed, identity of the patient receiving the act and diagnostic code associated with the act. Therefore, the RAMQ database covers costs related to physicians paid through fee-for-service as well as the fee-for-service part of costs attributed to physicians who receive blended payments. The information from this database represents the functional system of physicians in Quebec. Here we use this database to determine the costs for the patients before and during the hospitalisation. All costs are converted to 2012 Canadian dollars using the consumer price index from Statistics Canada.

## 2.2 Variables

The 30-day readmission is measured as a binary (0-1) variable where 1 indicates that the patient was readmitted within 30 days after the discharge and 0 otherwise. To illustrate methodological and practical issues, we predict readmission separately at hospital admission and hospital discharge. Available variables at hospital admission include the region code of the patients and the region code of the hospitals, the type of healthcare, the specialty of the treating physician at admission, the department of admission, the year of admission, the total amount in Canadian dollars billed by physicians during one year before the hospitalization and the total amount during two

years before the hospitalization, as well as age and gender of each patient. The prediction at hospital discharge further adds the length of each stay, MDC, DRG, major disease/condition of the patient, gravity level, mortality level, destination after the end of the stay, chronological number of the stay within the studied period, specialty of the treating physician at the discharge, department of the discharge, year of the discharge, as well as the total amount in Canadian dollars billed by physicians during the hospitalization. (See Table 1 for variable description.)

# 3 Study Design

## 3.1 Cross-validation of the prediction

We predict readmission with a supervised learning approach. The supervised learning approach allows us to compare different algorithms objectively with cross-validation. The dataset for a study is portioned into a training dataset and a testing dataset; the model is trained with the training dataset and tested by the testing dataset. This allows researchers to compare different algorithms objectively, since the observations in the testing dataset are not used in the training/construction of the model, and choose the most predictive model according to the performance (Tan et al. 2006). To avoid the results being influenced by the partitioning of the original dataset, a multi-fold cross-validation is necessary. Ten-fold cross-validation is performed in our case to avoid bias associated with the partitioning of training and testing datasets. To test sensitivity of the predictions, we compare different versions of the models, in particular Logistic regression and Deep Learning with different sample size and predictors.

## 3.2 Algorithms

Throughout the paper we report and compare the results using the area under the ROC curve (AUC). We compare the performance of five different algorithms: (1) simple logistic regression, (2) decision trees, (3) naïve Bayes, (4) random forest and (5) deep learning. Logistic regression is one of the most applied methods in many disciplines when the dependent variable is binary; however, it is weak when the feature space is large and sparse. In our case the Logistic regression fails when there are many sparse features such as the diagnostic codes with many different categories; so we remove main diagnostic code, region code, DRG, destination after the stay, specialty of the treating physician at admission and discharge, as well as department of admission and discharge from the Logistic regression. The other algorithms are machine learning algorithms. Generally speaking (see, e.g., Hastie et al. 2011, James et al. 2013, for more detailed discussion), decision

tree is fast and easily interpretable, but the prediction power is often moderate. Naïve Bayes is even faster than decision tree and is one of the fastest algorithms in machine learning. The features are assumed to be conditionally independent from each other and this "naïve" assumption largely increases computation speed. While the gain in computation speed is sometimes at the expense of less reliable prediction, the algorithm is nonetheless widely applied because running time is a key determinant in some practical cases. Decision Trees and naïve Bayes are the two algorithms applied by Hosseinzadeh et al. (2013) in the study of readmission in Quebec with machine learning approach. Random Forest is often the winner in prediction accuracy; it is a tree-based algorithm and has some similarity with classical decision tree algorithm; nonetheless, each node of each tree is generated randomly, making a "random forest" of a huge number of trees. Deep Learning is based on layered architectures of artificial neural networks. The idea is comparable to the way human brain works; more informative features are extracted and figured out at each deeper layer. Deep Learning is also strong in prediction, but the computation is often heavy due to complex structures of the layers. Fortunately, recent development in capacity of analyzing big data has facilitated implementation of these highly predictive algorithms. Here we apply the Random Forest and Deep Learning algorithms provided by H2O, which is now jointly offered with Spark under the name "Sparkling Water". The algorithms divide data into subsets and then analyze each subset simultaneously. These processes are combined in the estimation of parameters with a parallel stochastic gradient method (Recht et al. 2011). Table 2 shows the running time on Intel Core Duo required for one replication using 90% of the observations as training dataset and 10% as testing dataset.

# 4 Results

## 4.1 Key results

Table 3 shows the AUC obtained from different algorithms. Among the different algorithms, Deep Learning is the most predictive, and Random Forest is only slightly lower in AUC than the Deep Learning algorithm. Figure 1 shows the ROC curves of each algorithm at admission and discharge. The ROC curves for Deep Learning and Random Forest are more on the upper-left side compared to the ROC curves for the other algorithms, indicating better performance of the two state-of-the-art algorithms. The other algorithms predicted moderately but the Decision Trees algorithm failed in the prediction at hospital admission. Figure 2 shows the variation in AUC by MDC categories and discharge years, using the same color scheme as in Figure 1. The AUC varies across different MDC categories. The pattern of the variation is similar for different algorithms;

that is to say, when the AUC is relatively higher or lower for a specific MDC category we see this clearly for all algorithms. For example, the AUC is relatively lower for the newborns (MDC=15). This is probably because the vast majority of newborns remains in hospital but are predominantly healthy. At hospital discharge, the AUC of the two algorithms Deep Learning and Random Forest is always above the acceptable level 0.7; in fact, the AUC is always above 0.75 except in one case (MDC=20) and in some cases the AUC goes above 0.9 (MDC=8, 14, 22 and 25). Clearly, the prediction at hospital discharge is feasible for the general population in Quebec. As for the prediction at hospital admission, the AUC of the two best algorithms is above 0.7 in many MDC categories, but is below 0.7 in ten out of twenty-six categories (MDC=4, 5, 10, 11, 19, 20, 21, 23, 24, and 25). Nonetheless, in these cases the AUC is not too far from the acceptable level of 0.7 and we recall that the prediction here only used the variables that are available at the very beginning of each stay. Hence, a good thing in practice is to add more information (such as the DRG) quickly once available into the prediction during early stages of a stay. In contrast, the prediction at hospital admission is already highly reliable in two cases (MDC=8 and MDC=14); the former has an AUC above 0.8 and the latter above 0.9. Although Deep Learning performs the best in general, the difference in AUC is often minimal with the Random Forest algorithm. In certain cases the Random Forest algorithm does lead the Deep Learning algorithm; therefore, we recommend the use of both algorithms in practice for reasons of comparison.

Figure 3 and Figure 4 show the importance of variables at hospital admission and discharge, respectively, for the algorithms Deep Learning and Random Forest. Variable importance is determined by calculating the relative influence of each feature (explanatory variable) on the response variable, with the method of Gedeon (1997). At hospital admission, specialty of the treating physician, department of admission and the year of admission are very important for both algorithms. Total amount billed by physicians before the hospitalization (during one year and during two years) and age of patient are very important for the Random Forest algorithm but is relatively less important for Deep Learning, which gives relatively more importance to region code of the patients and region code of the hospitals as well as type of healthcare. As for the prediction at hospital discharge, DRG and diagnostic of major disease/condition are the two most predictive variables with the Deep Learning algorithm, accounting for nearly 75% of variable importance. These two variables are also important for the Random Forest algorithm, altogether accounting for more than 30% of variable importance; but destination after the end of hospital stay is the most predictive variable with the importance of 32.62%.

From Figure 4, it is clear that the medical codes DRG and diagshort contribute a lot to the prediction of readmission, among other predictive categorical variables such as specentree, specsortie, serventree and servsortie. It is not straightforward to include such variables in a simple Logit regression because this kind of categorical variable corresponds to a large number of binary/dummy variables that are sparse, i.e. with many zeros and very few ones, and this leads to a convergence problem in Logit regression (data not shown). In contrast, machine learning algorithms handle such variables with ease and this is particularly true for the Deep Learning algorithm in our case. As is well-known, the number of observations must be greater than the number of variables included in an analysis for reason of identification (see, e.g., Wansbeek and Meijer 2000). The categorical variables largely increased the number of variables in our analysis and for this reason we varied the sample size to compare the gain from Deep Learning in terms of AUC with respect to Logistic regression. The results are shown in Table 4 where the comparison was made at hospital admission and hospital discharge.

Deep Learning always performs better than Logistic regression; additionally, while the AUC increases slightly with sample size for Logistic regression, the gain from increasing sample size is relatively more remarkable for Deep Learning. At hospital admission, Deep Learning gains less than four percentage points in AUC with respect to Logistic regression using 1‰ of the sample, but the gain increases to more than seven percentage points when using the whole sample. Similarly, Deep Learning gains less than four percentage points in AUC compared to Logistic regression at hospital discharge when using 1‰ of the sample; the gain nonetheless rose to almost nine percentage points with the whole sample. In other words, categorical variables, in particular medical codes, bring new information to increase the AUC, and Deep Learning also benefits from big data ("big data" here refers to large sample size). Of course, with a fixed number of variables included in the analysis, the increase in AUC slows down when the sample size reaches a certain level (taking 50% of the sample or the whole sample give very similar results), which is well in line with the discussions in Varian (2014). One step further, Table 5 shows the AUC obtained from different sets of variables using the Deep Learning algorithm. It is clear that the medical codes DRG and diagshort increase the AUC significantly and the other variables regionetab, destination, specentree, serventree, specsortie, servsortie, which are not included in the Logistic regression, further raise the AUC. However, these latter variables are reported individually to be less predictive compared with the DRG and diagshort codes.

## 4.2 Discussion

From the above results, the most predictive variables are categorical variables with many different categories, though the continuous variables are also informative. For example, the amount billed by physicians contributes a lot to a better prediction especially when the medical codes are not available at early stages of hospital stay. Classical statistical models are not typically conceived for extracting information from variables with many different categories such as the diagnostic codes. Machine learning, in contrast, is a fast-developing research area where the algorithms are particularly designed to handle complicated real data, which offers an interesting option for hospitals. While our dataset is relatively big, machine learning algorithms work also well at smaller scales.

Comparing with previous results in the literature, Hosseinzadeh et al. (2013) also used data from the RAMQ to study 30-day readmission in Quebec. They took a machine learning approach with the Decision Trees and naïve Bayes algorithms. They did not particularly distinguish between the prediction at hospital admission and at hospital discharge. The AUC was around 0.67 in their paper and was increased to around 0.84 when they removed some "outliers" from the sample. Because we do not have the same dataset, it is not easy to know the exact reasons for the difference in AUC associated with the Decision Trees and naïve Bayes algorithms. The variables included in the analyses may have played a role. They started with a much larger number of features (more than 20,000) and pre-selected the variables with two feature selection algorithms. As they admitted, however, there was no consensus on how to distinguish among broad range of feature reduction methods. The variables in our analyses are selected based on prior knowledge, which leads nonetheless to more than 2,000 features. This is not necessarily better than applying a feature selection algorithm, but it seems from the AUC that we did not lose information compared to their approach. Because they did not give a detailed list of variables included in their study, it is not easy to know whether some specific features/variables played a key role in our prediction compared to theirs. Second, the most important reason is perhaps the choice of algorithm. The two state-of-the-art H2O algorithms (Random Forest and Deep Learning) became available only very recently, and these algorithms largely increased the AUC requiring only a reasonable amount of additional running time. To the best of our knowledge, our results are better than any existing result in the literature on hospital readmission, in a general sense.

It is an important debate as to whether predictions generated by AI will change medical practice and how. In the first place, more precise predictions can modify the burden of proof: while it

is more common to make safer decisions in the absence of information, more information and better predictions will shift decisions to riskier ones. Safer medical practice generally requires more tests and examinations given a patient medical condition. Furthermore, as shown by Agrawal et al. (2016), better information about the costs of risky actions will cause the decision-maker to revert to the safe action. Hence, a consequence of more precise predictions is that hidden costs related to risky decisions will have to be better evaluated. This may concern physiological, psychological or even economic consequences of decisions. For instance, program rationale for the judicious use of health resources suggests that many medical tests and examinations can be detrimental to the patient and also costly. However, this type of awareness campaign with the medical profession is not intended to reverse the burden of proof. Rather, it puts forward that there are hidden opportunities to choose riskier solutions (e.g., reducing blood samples in medical practice); and also that apparently safe decisions can be detrimental to the patient and thus risky. More effective campaigns that could reduce the need for unnecessary medical procedures should educate physicians about the use of precise machine predictions.

Better predictions also increase benefit-cost ratio of interventions. This is notably the case with interventions that reduce the costs associated with hospital readmissions. Indeed, the decision to intervene may be guided by the cost of the intervention and the benefit related to the likelihood that the intervention will succeed. In case of success, the patient would not be readmitted to the hospital for complications, which will save costs of hospitalization. What is more, the benefit of the intervention increases as the probability of being readmitted is getting high (see, e.g., Bayati et al. 2014). Thus, the precision with which the probability of readmission is measured has a direct effect on the benefit-cost ratio as it improves benefits through a possible better targeting of patients. Indeed, in the presence of cost constraints, a limited number of patients can be treated. In the absence of reliable prediction, the net benefit of the intervention is reduced because targeting is inefficient. On the other hand, if the prediction is sufficiently precise, the impact can be maximized by targeting patients with higher readmission probability. Gaining precision between two competing models (such as Logit and Deep Learning) could therefore help increasing the benefit of the intervention. To illustrate this, let's consider heart failure (HF) clinics in Quebec (Campagna et al. 2017). In Quebec, 4.43% of patients suffering from HF have active follow-up in HF clinic in 2015-2016. This represents 7760 patients with suspected HF over 175 109 patients suffering from HF without distinction of etiology or form. The total cost of HF clinics in Quebec is estimated at $11.8M which is $1515 per patient. The cost of readmission to the hospital averages $10348 per patient in the absence of a clinic, and $6209 per patient with a clinic, given the 40% decrease in

length of hospital stay due to the intervention. In addition, the intervention also has the effect of reducing the readmission rate by 40%. Overall, the benefit-cost ratio of HF clinics in Quebec is estimated at 1.23 if we consider readmission rate for heart failure in Canada of 23.6%. Nevertheless, it is possible to increase this ratio by better targeting patients through readmission prediction. Ultimately, if we select the most at-risk patients (whose 30-day risk of rehospitalization is close to 1), then the estimated ratio would be of around 5 (i.e. benefits are 5 times higher than costs). More realistically, if we target patients with an error representing approximately 30% of patients who would not have been readmitted even without clinical follow-up, the ratio would be of about 3.6, whereas it would be 4.1 with a model that would predict with 80% efficiency. The difference of 0.5 in the benefit-cost ratio represents approximately $790 per patient, or more than $6.1 million overall, which does not seem negligible. By increasing to 90% efficiency as the deep learning algorithm seems to allow, we would reach more than $13.8M gain, a slightly higher amount than the current cost of HF clinics in Quebec.

# 5 Conclusion

Hospital readmission can be costly to the health care system in any country. The Canadian, New Zealand and Australian governments, for example, have used the 30-day readmission rate as a quality indicator of hospital services (Goldfield 2010). In this paper, different methods have been used to estimate the probability of patient readmission within 30 days after a hospital stay. In particular, the traditional logistic regression method was found to give acceptable result, but is still difficult to implement using all needed characteristics. More innovative machine learning methods have been found to give better predictions while using more information. In particular, Deep Learning and Random Forest, the two most complex prediction methods, have found to be giving superior prediction results, while not being prohibitively hard to estimate. As expected, using information available at discharge is superior to using only information at admission, but information at admission still gives a decent prediction. As a final remark, deep learning is indeed strong in extracting information from variables with many categories such as the diagnostic codes, which may contain valuable information on patients but are difficult to be handled by classical statistical models.

Our study must be considered in light of several limitations. First, the current dataset does not include death outside of hospital stays, and therefore some observations that may not be considered remain in the dataset. Ideally, if out-of-hospital deaths could be identified, models would

consider death as a competing risk for readmission. Second, all data are from Quebec and the results need to be validated externally with other datasets in future studies. Third, we did not include hospital-specific information, which may further increase the AUC in practice for certain hospitals. Notwithstanding the limitations, we believe that the results are meaningful for practical use in hospitals. Our dataset is representative at least for Quebec, and most of the information we use here (e.g., medical codes) is easily available to hospital, although some coding might be necessary after data being extracted. The ease of estimations, as well as the relatively high reliability of predictions by our models brings new possibilities for decision makers in the health care system. Indeed, it would be easy to setup a system that will at the end of everyday predict the readmission probabilities for recent patients and provides a statistical report informing on the incoming readmission in the system. Such information could then be used to prepare for the incoming cohorts. On the individual level, prediction of the probability of readmission of a given patient allows physicians themselves to plan which patients will be most likely readmitted, and do a better follow-up on the patients' condition, as well as making care outside of the stay easier. Such a system is inexpensive to deploy, but can reduce medical costs effectively with the information it provides. What is more, better information for all actors should result in better health outcomes across the board.

# 6 References


Agrawal A, Gans JS, Goldfarb A. 2016. Exploring the impact of artificial intelligence: Prediction versus judgment. University of Totonto.

Au A, McAlister FA, Bakal JA, Ezekowitz J, Kaul P, van Walraven C. 2012. Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. American Heart Journal, 164(3): 365-72.

Bayati M, Braverman M, Gillam M, Mack KM, Ruiz G, Smith MS, Horvitz E. 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. PLoS ONE, 9(10): e109264.

Campagna C, Bourgon Labelle J, Echevin D, Farand P. 2017. Economic evaluation of heart failure management by specialized clinics in Quebec. University of Sherbrooke.

Coleman EA, Min SJ, Chomiak A, et al. 2004. Posthospital care transitions: patterns, complications, and risk identification. Health Services Research, 39(5): 1449-1465.

Conway D, White J. 2012. Machine learning for hackers. O'Reilly Media, Inc.

Cotter P, Bhalla V, Wallis S, Biram V. 2012. Predicting readmissions: poor performance of the lace index in an older UK population. Age and Ageing, 41(6): 784-789.

Gedeon TD. 1997. Data mining of inputs: analysing magnitude and functional measures. International Journal of Neural Systems, 8(02): 209-218.

Goldfield N. 2010. Strategies to decrease the rate of preventable readmission to hospital. Canadian Medical Association Journal, 182(6): 538-539.

Gruneir A, Dhalla I, Walraven C, Fischerand H, Rochon P. 2011. Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. Open Medicine 5(2): 31.

Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1): 29-36.

Hastie TJ, Tibshirani RJ, Friedman JH. 2011. The elements of statistical learning: data mining, inference, and prediction. Springer.

Hosseinzadeh A, Izadi MT, Verma A, Precup D, Buckeridge DL. 2013. Assessing the predictability of hospital readmission using machine learning. Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference.

James G, Witten D, Hastie T, Tibshirani R. 2013. An introduction to statistical learning (Vol. 6). New York: Springer.

Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S. 2011. Risk prediction models for hospital readmission: A systematic review. Journal of the American Medical Association, 306: 1688-1698.

Kind A. 2007. Bouncing back: Patterns and predictors of complicated transitions thirty days after hospitalizations for acute ischemic stroke. Journal of the American Geriatrics Society, 55(3): 365-373.

Krumholz H, Normand S, Keenan P. 2008a. Hospital 30-day acute myocardial infarction readmission measure: Methodology. Report prepared for Centers for Medicare and Medicaid Services.

Krumholz H, Normand S, Keenan P. 2008b. Hospital 30-day heart failure readmission measure: Methodology. Report prepared for Centers for Medicare and Medicaid Services.

Krumholz H, Normand S, Keenan P. 2008c. Hospital 30-day pneumonia readmission risk measure: Methodology. Report prepared for Centers for Medicare and Medicaid Services.

Ohman EM, Granger CB, Harrington RA, Lee KL. 2000. Risk stratification and therapeutic decision making in acute coronary syndromes. Journal of the American Medical Association, 284(7): 876-878.

Pham J. 2007. Care patterns in medicare and their implications for pay for performance. New England Journal of Medicine, 356(11): 1130-1139.

Recht B, Re C, Wright S, Feng N. 2011. Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent in J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira & K.Q. Weinberger, eds. Advances in Neural Information Processing Systems. Curran Associates, Inc. 24: 693-701.

Schneeweiss S, Seeger JD, Maclure M, Wang PS, Avorn J, Glynn RJ. 2001. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. American Journal of Epidemiology. 154(9): 854-864.

Stone J, Hoffman G. 2010. Medicare hospital readmissions issues, policy options and ppaca. Congressional Research Service Report for Congress.

Tan PN, Steinbach M, Kumar V. 2006. Introduction to data mining. Pearson Education India.

Varian HR. 2014. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2): 3-28.

Walraven C, Dhalla I, Bell C, Etchells E, Zarnke K, Austin P, Forster A. 2010. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Canadian Medical Association Journal, 6(182): 551-557.

Wang H, Robinson RD, Johnson C, Zenarosa NR, Jayswal RD, Keithley J, Delaney KA. 2014. Using the LACE index to predict hospital readmissions in congestive heart failure patients. BMC Cardiovascular Disorders, 14: 97.

Wansbeek T, Meijer E. 2000. Measurement error and latent variables in econometrics, Amsterdam: North-Holland.

Wolff J. 2002. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. Archives of Internal Medicine, 162(20): 2269-76.

Yu S, Farooq F, van Esbroeck A, Fung G, Anand V, Krishnapuram B. 2015. Predicting readmission risk with institution-specific prediction models. Artificial Intelligence in Medicine, 65(2): 89-96.

## TABLES & FIGURES

### Table 1: Variables used in predictions

| Variable | Definition | Type | N cat. | Mean | Median | SD | % missing |
|---|---|---|---|---|---|---|---|
| **At admission** | | | | | | | |
| serventree | Department at admission | Categorical | 88 | | | | 0.00% |
| specentree | Specialty of the physician at admission | Categorical | 54 | | | | 0.00% |
| region | Home region of the patient | Categorical | 19 | | | | 0.00% |
| regionetab | Region of the healthcare facility | Categorical | 19 | | | | 0.00% |
| typesoins | Type of care | Categorical | 6 | | | | 0.00% |
| anneeent | Admission year | Categorical | 37 | | | | 0.00% |
| female | Sex of the patient | Categorical | 2 | | | | 0.00% |
| montant1an | Amount in CAD billed by physicians during 1 year before admission | Continuous | | 262.24 | 46.13 | 766.85 | 0.00% |
| montant2ans | Amount in CAD billed by physicians during 2 years before admission | Continuous | | 419.57 | 86.02 | 1099.01 | 0.00% |
| age | Age of the patient | Continuous | | 46.50 | 49.00 | 26.59 | 0.00% |
| **At discharge** | | | | | | | |
| diagshort | Main diagnostic (simplified) | Categorical | 966 | | | | 0.00% |
| destination | Destination at discharge | Categorical | 27 | | | | 0.00% |
| DRG | Diagnostic-related group | Categorical | 618 | | | | 7.00% |
| MDC | Major diagnostic category | Categorical | 25 | | | | 7.00% |
| servsortie | Department at discharge | Categorical | 88 | | | | 0.00% |
| specsortie | Physician specialty at discharge | Categorical | 54 | | | | 0.00% |
| anneesort | Discharge year | Categorical | 19 | | | | 0.00% |
| nsejour | Stay number | Continuous | | 3.77 | 2.00 | 4.78 | 0.00% |
| dureesejour | Length of stay | Continuous | | 6.36 | 2.00 | 28.42 | 0.00% |
| montant | Amount in CAD billed by physicians during hospitalization | Continuous | | 190.20 | 0.00 | 565.57 | 0.00% |
| mortalite | Mortality risk | Continuous | | 1.27 | 1.00 | 0.61 | 7.01% |
| grav | Gravity level | Continuous | | 1.51 | 1.00 | 0.75 | 7.00% |

Note: The variables available at admission are, of course, also available at discharge. See supporting information for more descriptions of the variables. All costs are converted to 2012 Canadian dollars using the consumer price index from Statistics Canada.

**Table 2: Time required for one replication**

| **Algorithm** | Decision Trees | Naïve Bayes | Random Forest | Deep Learning |
|---|---|---|---|---|
| **Time** | 2.08 min | 1.91 min | 4.98 min | 9.00 min |

Note: This is the approximate running time on a computer (Intel Core Duo) required for one replication to train (using 90% of the observations) and test the model (using 10% of the observations) with all variables available at hospital discharge. In practice, if the model is already trained, it only takes a split second to calculate the probability of readmission for a certain patient.

**Table 3: AUC by algorithm and model**

| **Algorithm** | Logistic regression | Decision Tree | Naïve Bayes | Random Forest | Deep Learning |
|---|---|---|---|---|---|
| **Admission** | 0.7121 | 0.5017 | 0.7463 | 0.7814 | 0.7877 |
| **Discharge** | 0.7889 | 0.7264 | 0.8155 | 0.8706 | 0.8776 |

Note: More information is available at hospital discharge and so the AUC is always higher with respect to the AUC at hospital admission. Decision Tree fails at hospital admission.

**Table 4: Change of AUC with sample size**

| | 1‰ sample | 5‰ sample | 1% sample | 5% sample | 10% sample | 50% sample | Whole Sample |
|---|---|---|---|---|---|---|---|
| **Admission** | | | | | | | |
| Deep Learning | 0.7460 | 0.7499 | 0.7681 | 0.7698 | 0.7789 | 0.7865 | 0.7877 |
| Logistic regression | 0.7078 | 0.7160 | 0.7096 | 0.7102 | 0.7118 | 0.7121 | 0.7121 |
| **Discharge** | | | | | | | |
| Deep Learning | 0.8101 | 0.8140 | 0.8246 | 0.8501 | 0.8667 | 0.8773 | 0.8776 |
| Logistic regression | 0.7709 | 0.7830 | 0.7852 | 0.7866 | 0.7886 | 0.7886 | 0.7889 |

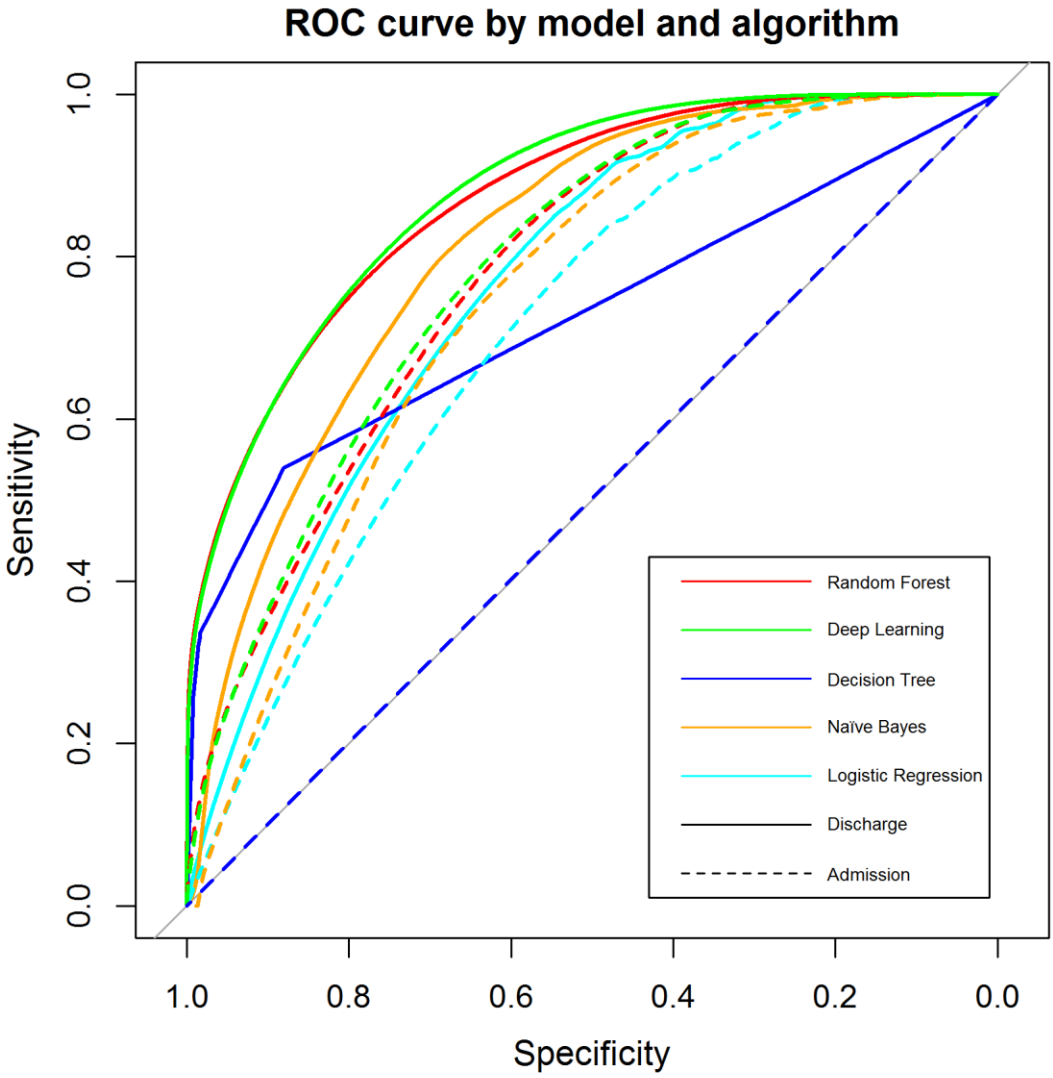
Note: Deep Learning adds certain categorical variables not used in Logit regression as explained in Section 3.2. Whole sample represents 1,633,099 hospital stays.

**Table 5: Change of AUC when adding more variables**

| | Variables in Logit | Add DRG only | Add diagshort only | Add DRG and diagshort | Add all variables |
|---|---|---|---|---|---|
| Deep Learning | 0.8120 | 0.8390 | 0.8274 | 0.8420 | 0.8776 |

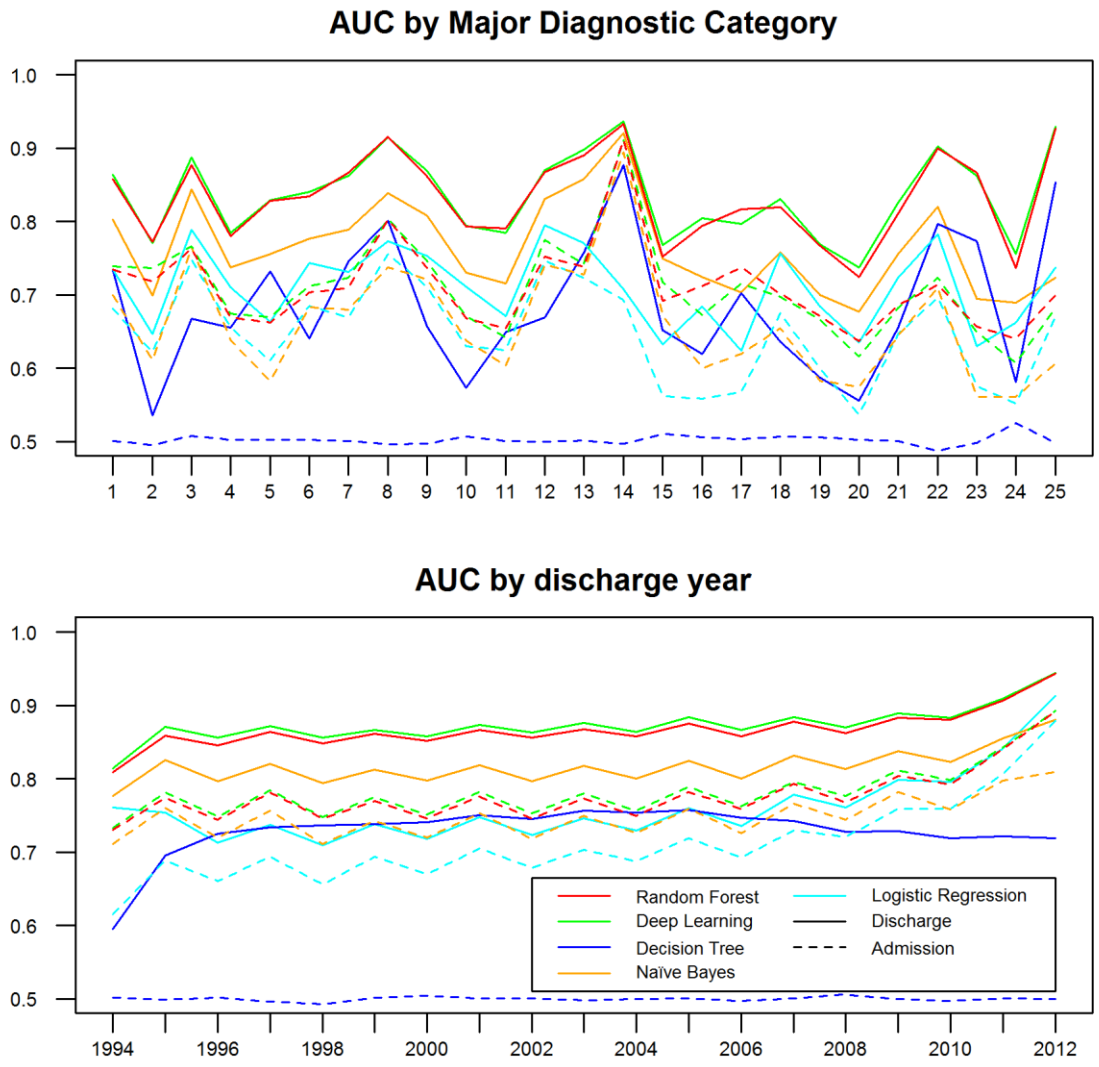
Note: Here we use the whole sample in all cases; the Deep Learning uses the same variables as in Table 4.

**Figure 1: ROC curves**



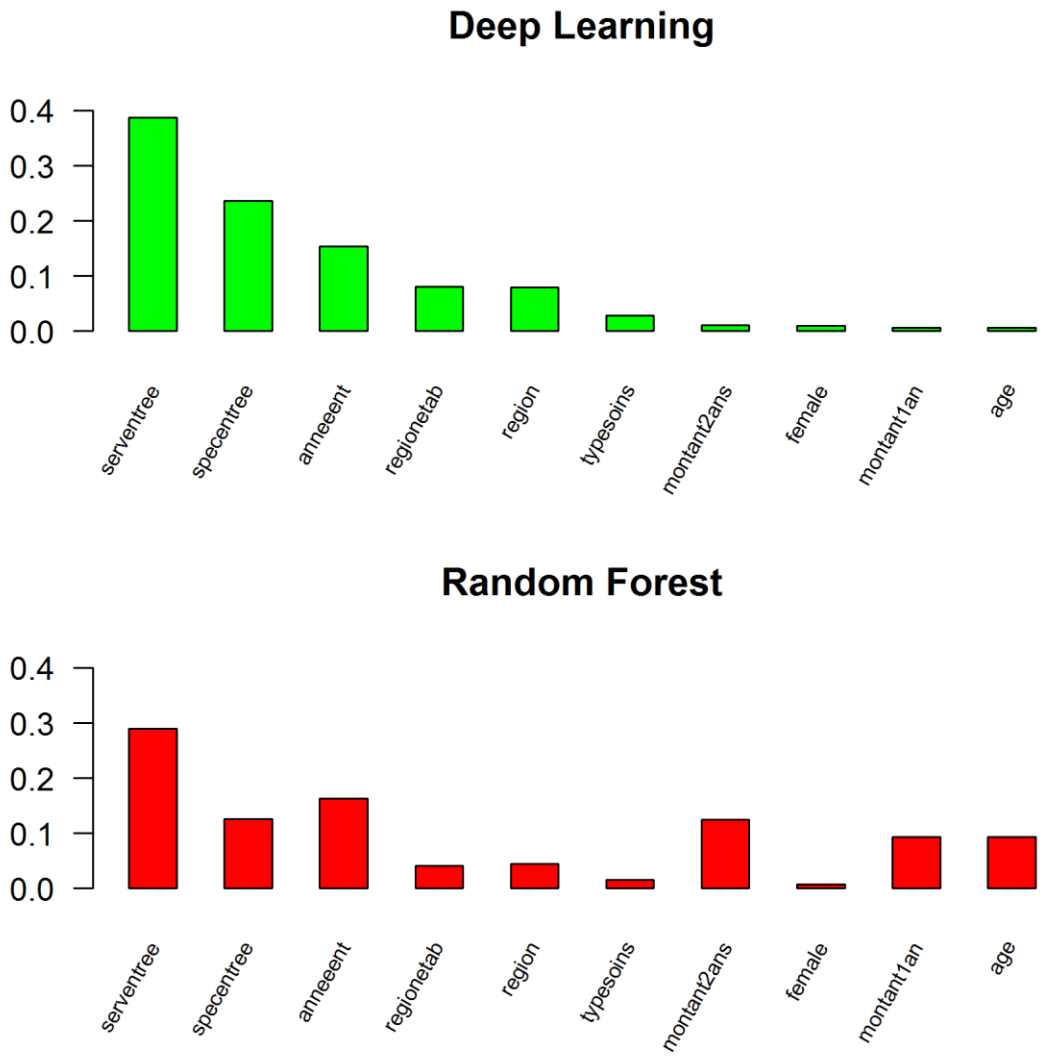

Note: The AUC is the area under ROC curve, which is between 0.5 and 1. The more an ROC curve is to the upper-left side of the figure, the bigger is the AUC. The solid lines correspond to predictions at discharge and the dashed lines correspond to predictions at admission. Each color corresponds to a specific algorithm as indicated in the window. Clearly, Deep Learning and Random Forest are the two best algorithms whose ROC curves are closer to the upper-left side.

**Figure 2: AUC by MDC and discharge year**



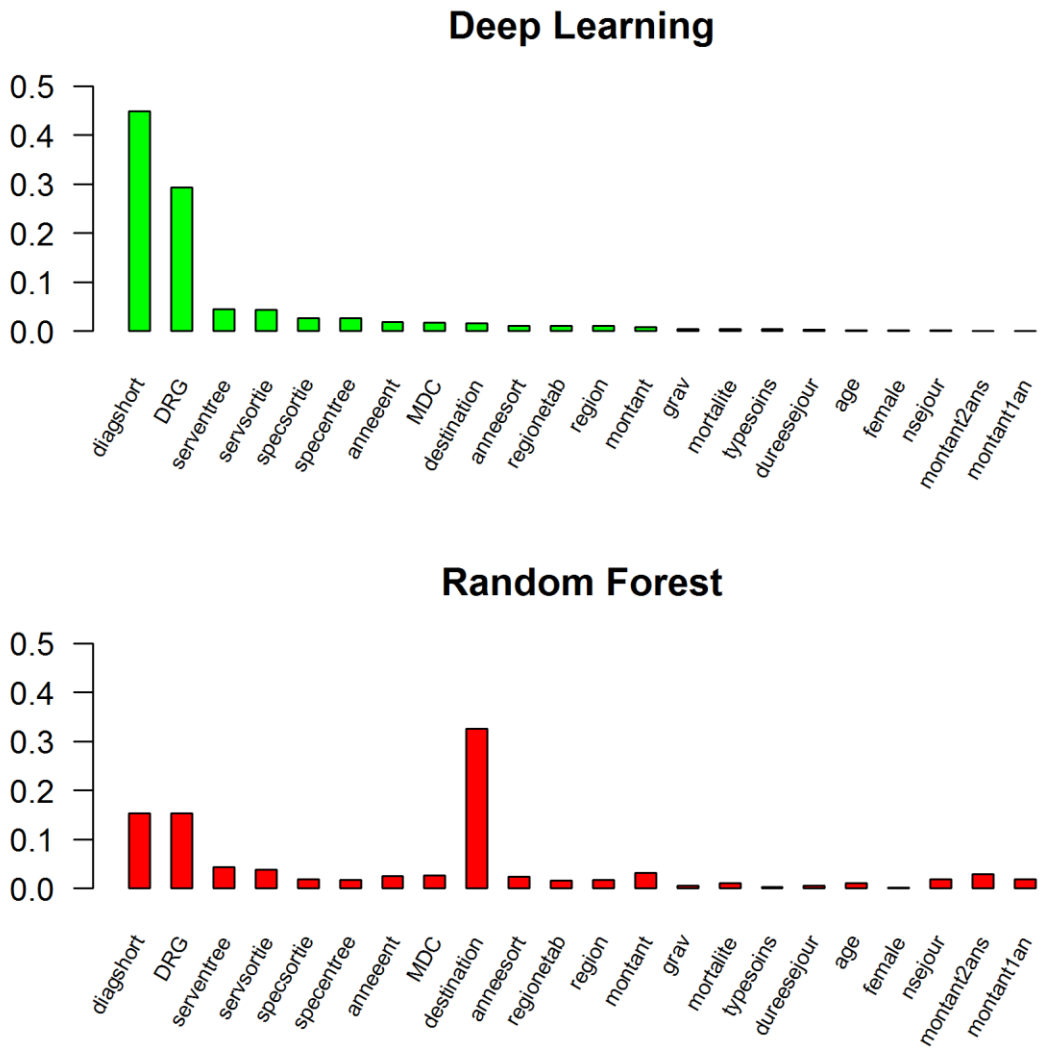

Note: The vertical axis shows the AUC. The solid lines correspond to predictions at discharge and the dashed lines correspond to predictions at admission. Each color corresponds to a specific algorithm as indicated in the window. MDC description: https://en.wikipedia.org/wiki/Major_Diagnostic_Category.

**Figure 3: Variable importance for the prediction at hospital admission**




Note: Variable importance is the relative influence of each explanatory variable on the response variable (readmission). The variables on the horizontal axis are the variables available at hospital admission. The variables are sorted by their importance in the Deep Learning algorithm; hence specentree is relatively the most important and age is relatively the least important for prediction with the Deep Learning algorithm.

**Figure 4: Variable importance for the prediction at hospital discharge**




Note: The figure has the same explanation as Figure 3, except that the variables on the horizontal axis are the variables available at hospital discharge (all explanatory variables). The variables are sorted by their importance in the Deep Learning algorithm; hence diagshort is relatively the most important and montant1an is relatively the least important for prediction with the Deep Learning algorithm.