Hospital Readmission Prediction : Preliminary Results and Remarks

Maximillien Burq Sebastien Boyer Joan LaRovere

December 12, 2015

1 Problem formulation and contribution

When discharged from hospital after an inpatient stay - be that for asthma, pneumonia, a heart attack, whatever the condition - there is an 8 to 18% chance of readmission to hospital non-electively within 30-days of discharge for the same condition. As figure 9 indicates, there is a high variation by geography. Although some readmissions are clinically necessary or even planned as part of optimal patient care, these variations also suggest a higher likelihood of readmission by institution. This is important on many levels. Firstly, readmission is clinically risky. Receiving the necessary hospitalization on the first admission and having clear discharge instructions are the ideal in clinical care. Secondly, it is expensive. Readmission within 30 days of discharge accounts for a $41.3 billion dollar market and accounts for 11% of total hospital costs.

In 2012, the Hospital Readmissions Reduction Program (HRRP) was established under the Affordable Care Act in which the Centers for Medicare and Medicaid Services (CMS) are required to reduce payments to Inpatient Prospective Payment System (IPPS) hospitals with excess readmissions, as part of a focus to tie payment to quality and value over the coming years. Readmission is defined as admission within 30-days of discharge. CMS uses an all-cause definition for readmission but does allow for planned hospitalization as an exception since 2014. A hospital's excess readmission ratio is a measure of a hospital's readmission performance compared to the national average for the hospital's set of patients with that applicable condition. The risk adjustment methodology endorsed by the National Quality Forum (NQF) is used to calculate the excess readmission ratios, which includes adjustment for factors that are clinically relevant including certain patient demographic characteristics, comorbidities and patient frailty. An excess readmission ratio is established for each applicable condition. Readmission measures were adopted for Acute Myocardial Infarction (AMI), Heart Failure (HF) and Pneumonia (PN). The current focus is on readmissions occurring after initial hospitalization for selected conditions. Overall readmission rates regardless of initial diagnoses are collected for a national average and hospital specific, but these overall rates are not currently used in the HRRP to calculate readmission penalties. However reduction in payments occurs across all Medicare admissions. The greater the rate of excess readmission, the higher the penalty incurred with a maximum of 3% reduction in reimbursement of the hospital's base inpatient claims. CMS also adjusts for certain demographic characteristics of both the patients being readmitted and each hospital's patient population (such as age and illness severity) before comparing a hospital's readmission rate to the national average.

Since that time, measures were expanded to include acute exacerbation of chronic obstructive pulmonary disease (COPD), elective total hip arthroplasty (THA) and total knee arthroplasty (TKA), with coronary artery bypass graft (CABG) surgery and additional pneumonia diagnoses (aspiration pneumonia and sepsis patients coded with pneumonia present on admission but excluding severe sepsis) soon to be added to the calculation of a hospital's readmission payment adjustment factor. Three years of discharge data with a minimum of 25 cases are used to calculate a hospital's excess readmission ratio for each applicable condition.

Those most likely to incur these penalties are major teaching hospitals and those serving larger portions of low-income patients. In 2015 penalties and the numbers of hospitals receiving them increased mainly due to the number of conditions being measured. Hospitals would do well to anticipate additions to the readmission rate calculation and put processes in place to reduce these anticipated categories.

Although many hospitals have been addressing this problem given the clinical risk to patients, now further incentivized by the recent financial penalties imposed through the Affordable Care Act for high 30-day readmission rates, it is perhaps insurance companies and CMS that benefit most from reduce readmission rates as reducing the current $41.3 billion dollar bill is ultimately in the payer's interest.

Clinicians strive to avoid readmission for their patients and do their best to be certain that they are clinically ready for discharge but obviously this system is imperfect. Clinicians may miss clues that predict readmission and social and psychological factors may also impact hospital readmission. Building a predictive model of 30-day readmissions to assist hospitals and clinicians would aid this process, with Dell well positioned to solve this problem. Originally a data hardware company, through recent acquisitions, Dell has positioned itself as an IT services company. It currently operates the data centers for many hospitals and is strategically positioned to make the leap to leverage this data with analytics capability and gain insights to feedback to these hospitals. Predictive analytics applies to many of areas of health, not just 30-day hospital readmission, and is undoubtedly a game changer for the future. This is not an if but rather a when and who question and it is strategic of Dell to attempt to capitalize on its position and try to capture this market.

2 Dataset Description

Dell provided a data set collected from one anonymous US hospital with all personal information masked for data privacy. The dataset contains information about 1500 patient-admissions in the hospital as well as the corresponding outcome for readmission. The information about each patient is composed of 26 variables which fall into three categories as shown in table 1.

As figure 1 shows, this data set contains information about patients ranging from 60 yrs of age to 100 yrs of age among which 48% are females. Figure 3 shows that 7% of patients were readmitted within 30 days after their discharge and that one third of overall patients in our dataset were readmitted at some point.

We dealt with missing values using default values and customized those based on the significance of each variable. The outcome of interest is a binary variable defined as readmittance to hospital or non-readmittance to hospital within the first 30 days after being discharged. This threshold is justified by the current legal context of readmission in the US but we also benchmarked our algorithm on other thresholds (as shown at the end).

3 Predictive analytics

3.1 Performance Metrics and Baselines

Prediction Objective

Choosing the information to predict is a key component of our problem. We started with the simple 0-1 prediction of whether a patient is going to be readmitted within 30 days. However, the goal is to create a tool that is going to help doctors make the right decision. Therefore it seems important to provide more information, such as the confidence of our prediction. Equivalently, we focused on computing a risk of a patient, measured as an estimated probability that he will be readmitted within 30 days.

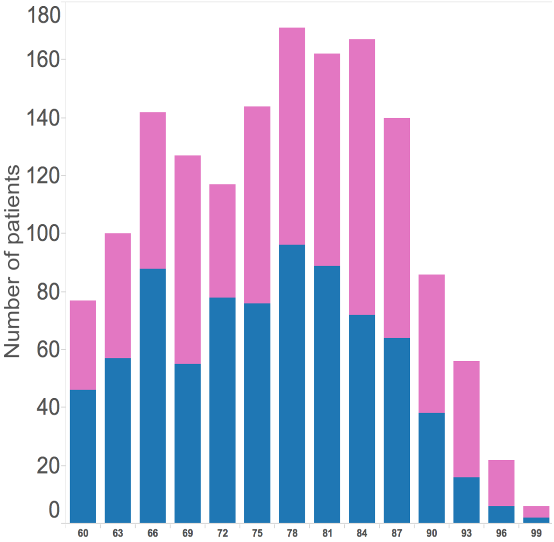


Figure 1: Distribution of Age of patients (pink=female, blue=male)

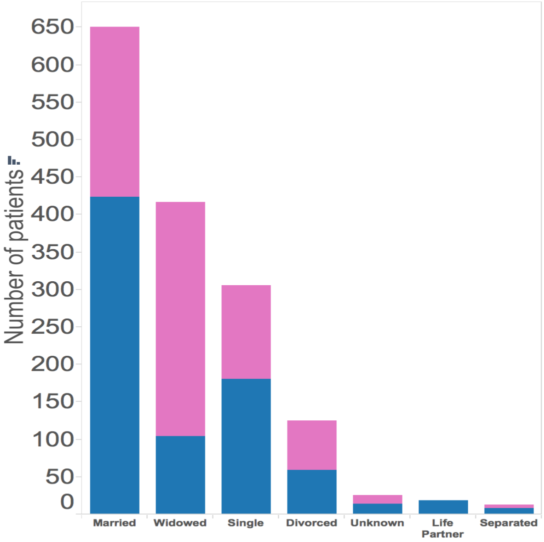


Figure 2: Marital status of patients

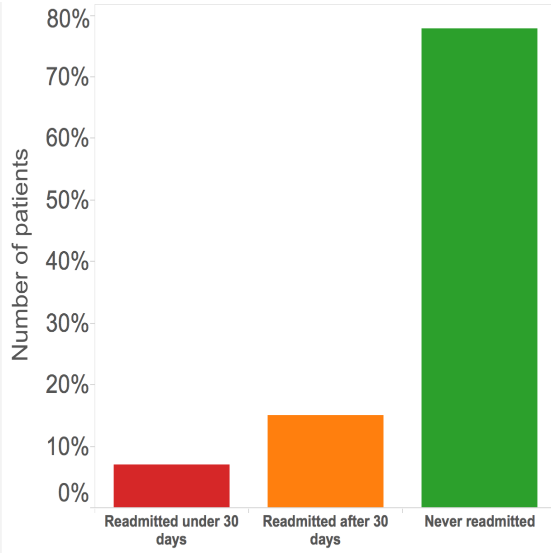


Figure 3: Categories of readmission in our dataset

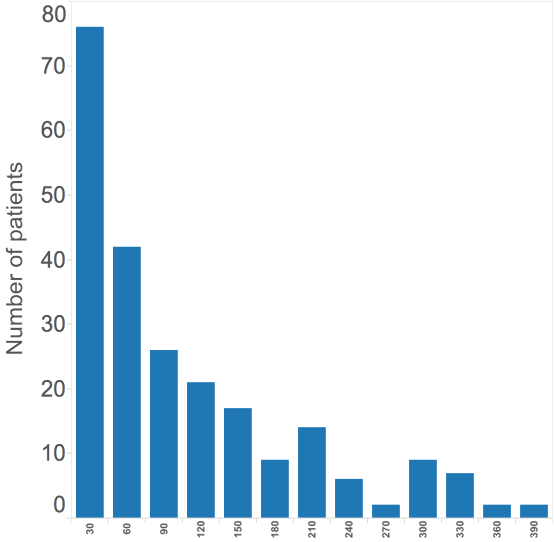


Figure 4: Distribution of number of days before readmission (for readmitted patients only)

Performance Metrics

We focused until now on the Area under the ROC curve (called AUC) as our main performance metric for our different algorithms. This metrics has a simple interpretation :

$$\begin{aligned} AUC = x \iff & \text{ If one future-readmited patients and one non-future-readmited patient} \\ & \text{ are presented to the algorithm it will know which one is which x \% of the time} \end{aligned}$$

We believe that this metric is very useful for a variety of reasons:

- It captures completely the tradeoff between sensitivity and specificity.
- It allows us to compare all our algorithms together without any bias.
- A very simple baseline is available (predicting randomly for every patient) that we can compare our algorithms against.

However, there are also some drawbacks to the AUC. For instance, it doesnt necessarily capture that the costs of false positives and false negatives are very different. In this respect, the value of the ROC at some points might be more insightful. Secondly, the AUC doesnt take into account any notion of confidence of the prediction. In this respect, a log loss function would be better but we didn't explore this yet. Also taking into account the relative cost of false positive and false negative could lead to a real optimization of cost based on these predictions. We haven't explored this area of study yet.

Baselines

As a baseline we used the score of a random guess (AUC= 0.5). Notice that there is an inherent bias in the dataset, in that we can assume that the doctors tried to avoid readmissions as much as possible for the patients that we are analyzing. Therefore, the goal of the algorithm would not be to try to predict better than a doctor, but rather to predict after the doctor. The question of decision making under capacity constraints that automatically arises is a very interesting and complicated one, and is not the purpose of the current study.

3.2 Feature Engineering

The first part of our analytics work on the dataset focused on extracting as much information from the data available. The idea here was to change the format of the data we had to maximize the available information given as inputs into a predictive algorithm. This process is called features engineering.

3.2.1 Dummy categorical columns

When dealing with categorical columns we encountered the following issue. A lot of the values were of the form :

$$\begin{aligned}\text{Value of patient } i &= sub\_value\_1, sub\_value\_2, ...\\ \text{E.g. , Value of patient } i &= \text{Alone, Pets}\end{aligned}$$

This presents a challenge since the same set of subvalues is very unlikely to appear more than a very few times (particularly in small datasets like this one). A predictive algorithm could therefore not learn how each of the subvalues individually influenced the output. To tackle this, we transformed each of the categorical column into several dummy columns or subcolumns corresponding to each of the subvalues. Those subcolumns contained only 1s and 0s (1 only if the subvalue appeared in the sequence). For example ,

$$\begin{aligned}\text{Value of patient } i \text{ for subvalue "Alone"} &= 1\\ \text{Value of patient } i \text{ for subvalue "Pets"} &= 1\\ \text{Value of patient } i \text{ for subvalue "Children"} &= 0\end{aligned}$$

We applied this technique to 17 columns resulting in a dataset of 154 subcolumns.

Table 1 presents the results we were able to achieve using each of the common machine learning algorithms and the variables just created. Each result is computed using 5-fold cross-validation. That is, for each algorithm, we randomly split the data set into a train set (80%) and a test set (20%) and we measure the AUC, then we repeat this process 5 times with different random split. The AUC displayed below is the average over the 5 splits.

Learning Algorithm	AUC
Random guess	0.500
CART	0.602
Logistic Regression	0.629
SVM-RBF	0.633
Gradient boosting classifier	0.642
Random Forests	0.648

Table 1: Results achieved after the first step of cleaning and choosing appropriate data format

3.3 Text information extraction

The second part of our analytics work was to extract information from hand typed text. Two of the columns available to us,Patient Reason for Admission and Chief Complaint/Injury, presented information in text format. For example :

$$\text{'Chief Complaint - Injury' value of patient } i = \text{"pre surg check"}$$

Again it is impossible for a predictive algorithm to use those columns as they are because if two strings don't match exactly they would be considered as two distinct categories even though they might share some semantic similarities. For instance, we would like our predictive algorithm to understand that "unable to breathe is similar to shortness of breath whereas abdominal pain is very different from chest pain

3.3.1 Manual Feature Engineering

First we tried to create a manual classification of the handwritten text. To do this, we extracted two conditions (respiratory and heart) that made the most clinical sense.

We then created two sets of keywords that related to each of these conditions, first without and then with the help of a medical doctor. We then labeled 1 in the corresponding columns whenever there were substrings of the hand-typed text in the keyword dictionary.

The first keyword lists were: "shortness of breath", "pneumonia", "copd, "resp" and "cough" for respiratory conditions "heart", "cardiac", "chest pain" for heart conditions. These increased the AUC by 1.1%.

We then increased the number of keywords, especially in relation to specific conditions that are associated with heart failure: "heart", "card", "chf", "chest pain", "congestive heart failure exacerbation", "elevated tropinin", "NSTMI", "sob","fib","bradycardia","chest tightness", "icd placement","stent","arrest","cad", "swelling","syncope", "pacemaker","angi". This increased the AUC by 1.7%.

Note that in order to avoid overfitting, we made sure to ignore the readmission variable when we chose the keywords.

3.3.2 Automated Topic Modeling

We then implemented an automated Topic model algorithm. The goal of this algorithm is to automatically group all the sentences into a few buckets (typically five or ten) such that each bucket shared similar semantic meaning. Doing so, instead of having to deal with 1500 completely different sentences (which will be useless for a predictive algorithm) we can use the buckets into which the sentences fall as features. Those buckets will appear several time thus enabling the predictive algorithm to learn how the buckets correlate with the output. We emphasize the fact that the algorithm automatically comes up with the buckets that best describe the diversity of the data, avoiding need for any human intervention.

$$\begin{aligned}
\text{Text of patient } i &= \text{"Shortness of breath with some cough''} \\
\text{Bucket of patient } i &= \text{Breath} \\
\text{Text of patient } j &= \text{"resp. distress''} \\
\text{Bucket of patient } j &= \text{Breath}
\end{aligned}$$

Doing so we obviously lost some granularity in the information obtained but the good news is that the number of buckets can be tricked by a human. In other words, as we get more data we can start increasing the number of buckets and thus the granularity of the information we capture. The topics generated by the algorithms are described in table 6 of the Appendix. This increased the AUC by 2.7%.

3.3.3 Results and Further Developments

By extracting text information using a Topic model method we were able significantly outperform the existing algorithms.In particular, we gain almost 3 points of AUC which is very encouraging. Again, we believe that as more data will be gathered, more overlap between diseases will emerge and the predictive power of such technique will increase a lot.

Learning Algorithm	AUC
Random guess	0.500
Random Forests without Text	0.648
Random Forests without Non-expert topics	0.659
Random Forests without Expert topics	0.665
Random Forests with automatic Topic models	0.675

Table 2: Best results achieved so far

Further development of this algorithm would involve combining human knowledge of medical conditions to the flexibility of an automated topic modeling algorithm. One way to go would be to manually add constraints. For instance, instead of letting the algorithm decides freely what categories best describe the set of sentences, a doctor could help the algorithm defining those buckets by manually entering some word-bucket correlations (e.g., Biventricular ICD is correlated with the Bucket Heart disease).

3.4 Variable importance and selection

Measuring the relevance of variables fo the particular task at stake here (readmission prediction) is crucial both to improve the performance of our predictive algorithm and to better understand what really affects readmission rate. We dedicate this to the exploratory analysis we have done on how useful each of the variable are and how we used that to improve our algorithm.

We used the *Randomized Logistic regression* algorithm to compute the variable importance for all variable available after the feature engineering process (154). Figure 5 shows the results in terms of score (a decimal number between 0 and 1) for each variable.

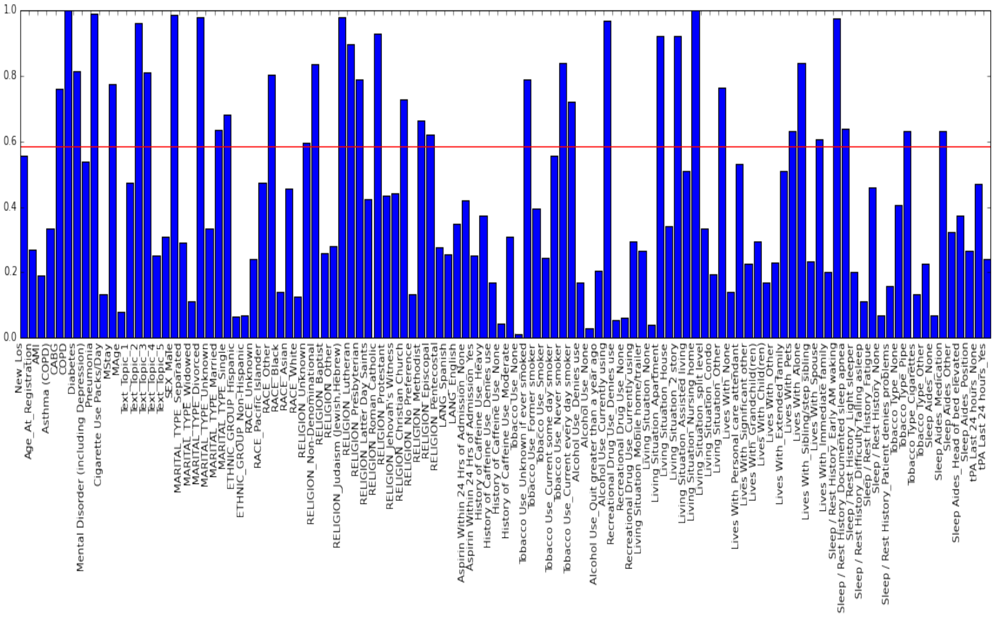


Figure 5: Importance of each of the available variable. The importance is measured using the *Randomized Logistic regression* algorithm. The red bar is the threshold used to select the final variables (only the variable with an importance above the threshold are used in the final model)..

We then benchmark different threshold level and apply our best algorithm only on the variables above the threshold. The results in terms of AUC are displayed in figure 6. The best choice for the threshold is 0.59 which gives 35 variables used in the best algorithm. Using this feature selection process presented above, we were able to boost the performance of the algorithm by almost 3 AUC points.

Learning Algorithm	AUC
Random guess	0.500
Random Forests with Topic model	0.675
Logistic regression with Topic model and Feature selection	0.702

Table 3: Best results achieved so far

4 Design final algorithm

The final and best predictive algorithm contains the 35 variables displayed in table 7 of the Appendix.It uses a randomized logistic regression and performs according to the results shown in table 4.

Our code is currently written in python and uses the following library : *numpy*, *pandas*, *sklearn*.

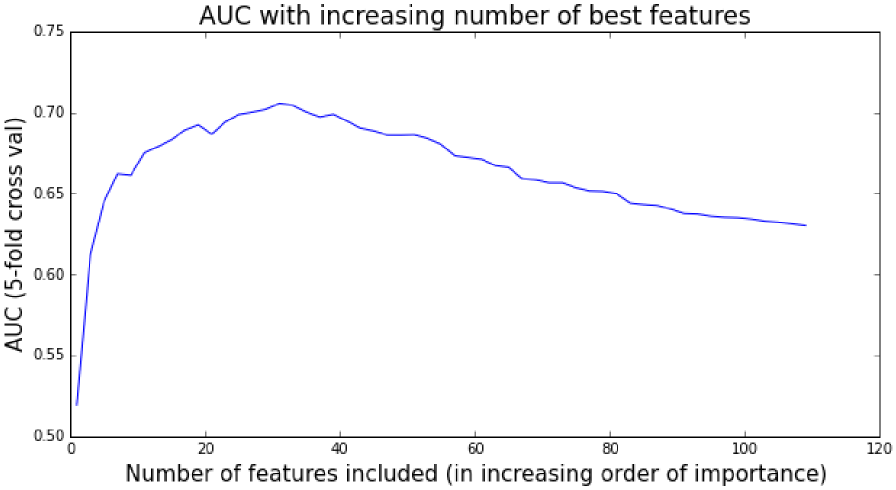


Figure 6: AUC of logistic regression for different number of variables included (in orer of decreasing relevance)

Readmission category	AUC
15 days	0.721
30 days	**0.702**
60 days	0.695
180 days	0.679

Table 4: Performance of the final predictive algorithm on different readmission horizons

5 Insights

5.1 Incremental impact of the factors on readmission risk

This section summarizes one of the main insights gained from developing the predictive algorithm discussed above. Using this predictive algorithm we design a set of counterfactual experiments to get of sense of the incremental risk that each of the factor account for. Particularly we proceeded as follows for each categorical variables :

1. Switched the categories (from 0 to 1 or from 1 to 0) for all the patients.

2. Predicted the new risk for all patients (counterfactual prediction)

3. Computed the difference between the 1-prediction and the 0-prediction for all patients

4. Averaged all the differences

The figure7 displays the results for all categorical variables remaining in the best 35 variables. Blue bars show a positive impact on the patient (reducing risk) and red bars show a negative impact. The figure reads as follows : Other things being equal, living in a Nursing Home decreases the risk of readmission by 0.1%, thus preventing 1 readmission every 1000 patients according to the dataset and our best predictive algorithm.

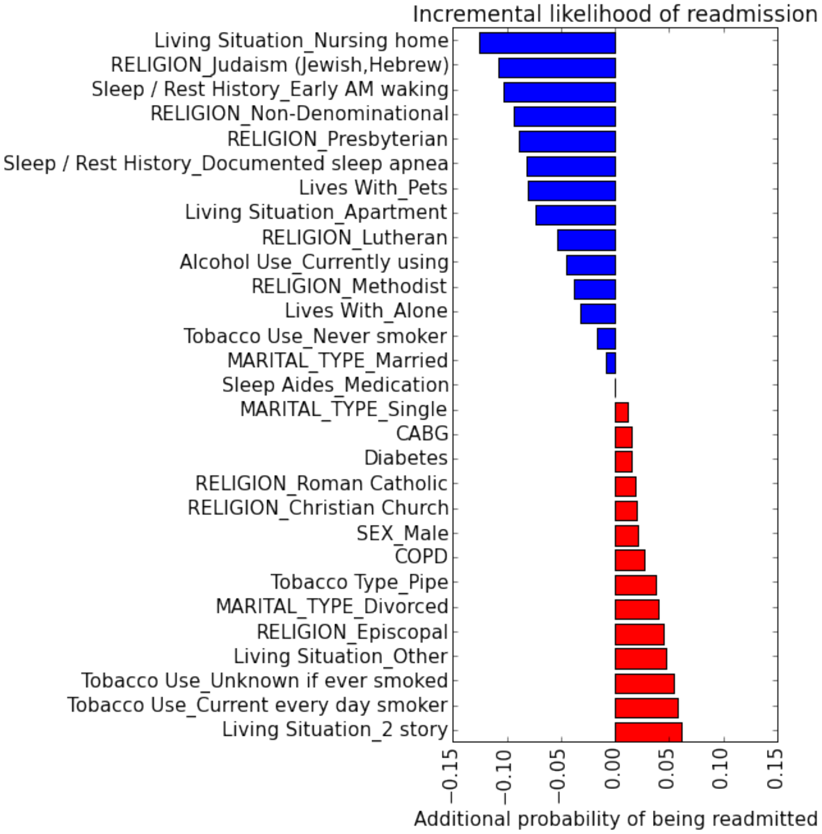


Figure 7: Incremental risk of the factors included in the most significative 35 variables.

5.2 Clustering

We performed a clustering analysis on the dataset in order to determine those groups of patients who present with common risk factors regarding likelihood of readmission in order to create greater customer segmentation and a more targeted tool.

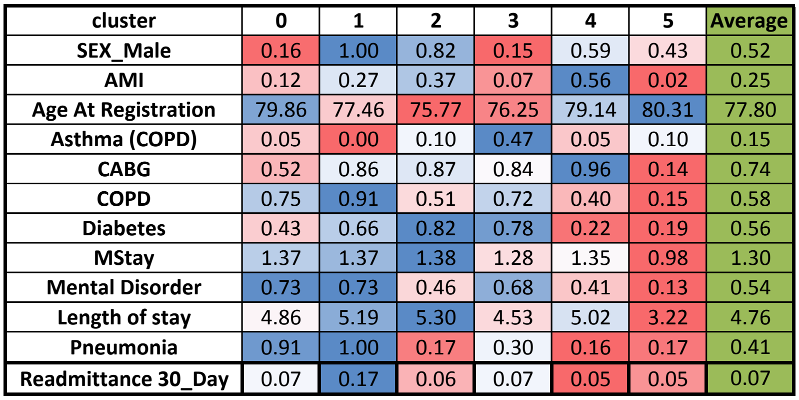
6 Product and Business Case

6.1 Getting the right data

Given the limitations of the present dataset, there are many aspects of the analysis that we were not able to perform. In the next section we provide interesting directions for future work, as well as important information to collect.

Multiplicity of patients

It would be more informative to know which of the admissions correspond to the same individual. This is possible for some of the patients because of similar demographic information, and allows us to leverage the historical information to enhance the predictions. It seems that many of the entries in the dataset correspond to identical patients. Accessing this information would provide valuable insights such as:



cluster	0	1	2	3	4	5	Average
SEX_Male	0.16	1.00	0.82	0.15	0.59	0.43	0.52
AMI	0.12	0.27	0.37	0.07	0.56	0.02	0.25
Age At Registration	79.86	77.46	75.77	76.25	79.14	80.31	77.80
Asthma (COPD)	0.05	0.00	0.10	0.47	0.05	0.10	0.15
CABG	0.52	0.86	0.87	0.84	0.96	0.14	0.74
COPD	0.75	0.91	0.51	0.72	0.40	0.15	0.58
Diabetes	0.43	0.66	0.82	0.78	0.22	0.19	0.56
MStay	1.37	1.37	1.38	1.28	1.35	0.98	1.30
Mental Disorder	0.73	0.73	0.46	0.68	0.41	0.13	0.54
Length of stay	4.86	5.19	5.30	4.53	5.02	3.22	4.76
Pneumonia	0.91	1.00	0.17	0.30	0.16	0.17	0.41
Readmittance 30_Day	0.07	0.17	0.06	0.07	0.05	0.05	0.07

Figure 8: Results of the clustering analysis. For each cluster, we computed the percent of people in these clusters who had specific comorbidities (to be compared with the average percent overall). Blue (respectively red) bars indicates that the percent is higher (respectively lower) than average.

- Multiplicity of conditions for a single patient (not all conditions are revealed during a single stay at the hospital)
- Complications due to previous illness- Ability to cross-reference information and use data linked with the previous stay.

Clinical Data

One of the main difficulties with the present dataset was the lack of clinical data. We saw in the Variable Importance section that the comorbidities are very useful in predicting readmission. The clinical data chosen is understandable in that these match clinical factors focused on by CMS. However, a more optimal model might include more clinical data.

Reason for Readmission

It is very likely that some of the readmissions were planned by the doctors as follow-ups. Because we do not have information on which admissions were planned, it is hard to determine if our algorithms are learning hidden clinical information, or simply the hospital's policy in terms of follow-up examinations.

6.2 Quantifying algorithm quality

The objective is to develop a module aimed at predicting the likelihood of readmission as part of a broader discharge software suite for hospitals. Therefore, rather than accuracy benchmarks, the quality of our algorithms should be assessed in terms of the gain (for hospitals and patients) compared to the status quo.

It is notoriously hard to try to get a precise estimate of the "welfare gain" from the patient perspective, especially prior to any tests on a real scale. However, it is possible to measure the monetary gain from the hospital's perspective.

To do that for this study, we would need to balance the gain from preventing a fraction of the readmissions (the true positives that our algorithm would accurately predict) compared to the cost of keeping some patients longer than necessary in the hospital (the false positives that our algorithm would classify as at risk but would have not been readmitted).

Cost of readmission

We are aware that the numbers here are subject to large variations, and we only ask for intervals and orders of magnitude to help us think the problem through.

- What is the average cost of an admission ? Who incurs the cost (hospital, insurance company, etc)
- Does that vary for a readmission ?
- How does that vary by patient condition ?
- Are there some (known) factors that impact this cost (insurance provider, demographic factors, etc)
- What are the main actions available to hospitals for patients with high readmission risks ? (additional medical exams, keeping them longer, etc)
- How does that vary by patient condition ?
- What are the typical costs of such actions ?
- By how much do they reduce the likelihood of readmission ?

7 Conclusions and Remarks

Many strategies are effective in reducing readmission including clarity around discharge instructions, integration with post-acute care providers and primary care physicians, as well as efforts to reduce complications and morbidity surrounding the hospital admission. Ideally strategies should be created to target the customer segments as built out through clustering algorithms on a more comprehensive dataset.

If we put a name and a face to the clustering algorithms shown in figure 8 it is easier to understand how we can leverage cluster analysis to come up with simple rules to predict the probability of readmission for new patients and segment them accordingly. 'Perky Peg' is an elderly woman with no major chronic cardiorespiratory illness who comes into hospital for a single issue, such as a knee replacement. Her risk of 30-day readmission is 2%. Let's compare her with 'Chronic Carl.' He is an elderly man with an acute admission for coronary artery bypass grafting with pneumonia and a history of chronic obstructive pulmonary disease. He has a 15% chance of 30-day readmission. With this customer segmentation, we can introduce new techniques or approaches to preventing readmission that align with the target market. For example, 'Perky Peg' would benefit from robotic discharge training to be sure she understands her aftercare to prevent readmission. 'Chronic Carl' would be best served in discharge to a nursing home before returning home. He would also benefit from wearables/Internet of Things technology to notify caregivers of worsening clinical status to provide early intervention before readmission is necessitated.

Going forward Dell should continue to develop this methodology and test and improve the algorithm using larger datasets to bring a robust solution to the health care market. In addition to more clinical data, social media and customer relationship management data should be explored. Internet of Things/Wearables data should also be obtained, at least in a pilot fashion, and interfaced with the algorithm to see its cost/benefit and impact on reducing 30-day readmission. Emerging technologies such as robotics, artificial intelligence and telemedicine should be explored as a means to improve the inpatient and outpatient customer experience and reduce 30-day readmission.

8 Appendix

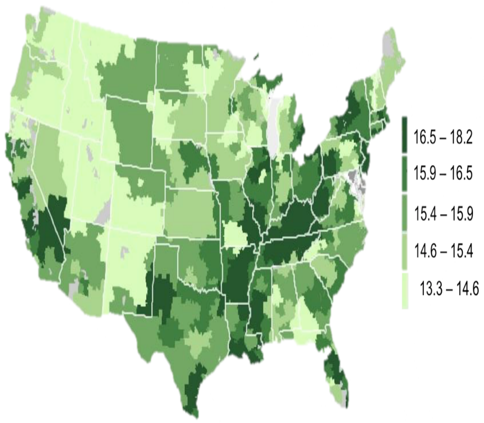


Figure 9: Distribution of Age of patients (pink=female, blue=male)

Category	Variable name	Variable Type	Variable id
Demographics	Sex	categorical	1
	Marital status	categorical	2
	Age	integers	3
	Race	categorical	4
	Ethnic group	categorical	5
	Language	categorical	6
	Religion	categorical	7
Behavioral	Tobacco use	categorical	8
	Cigarette per day	integers	9
	Recreational drug use	categorical	10
	Caffeine use	categorical	11
	Alcohol use	categorical	12
	Sleeping habits	categorical	13
	Sleep aides use	categorical	14
	Aspirin use	categorical	15
	Living situation (appart,...)	categorical	16
	Living with (daughter,...)	categorical	17
Clinical	Diabetes	categorical	18
	Asthma	categorical	19
	CAPG	categorical	20
	COPD	categorical	21
	Pneumonia	categorical	22
	Mental disorder	categorical	23
	AMI	categorical	24
	Patient reason for admission	text	25
	Chief complaint	text	26

Table 5: Description of variables as given in the dataset

Text category	Words associated and weights
Topic1	chf(10.7%), exacerbation(0.0755), stated(0.0610), swollen(0.0232) renal(0.0174), sobcopd(0.0174), chronic(0.0145), legs(0.0116) pleural(0.0116), pneumoniapneumonia(0.0116), abd(0.0116), lower(0.0116) pt(0.0108) swelling(0.0094), overload(0.0087), fluid(0.0087) syncope(0.0087), week(0.0087), insufficiency(0.0087), feet(0.0087)
Topic2	sob(0.1986), shortness(0.1235), x(0.0606), pneumonia(0.0363), failure(0.0217), days(0.0203), heart(0.0194), breathchf(0.0170), mi(0.0170), breathshortness(0.0170), fever(0.0145), increased(0.0121) vomiting(0.0121), transfer(0.0097), exertion(0.0097), nausea(0.0073) pt(0.0073), exertionsob(0.0073), told(0.0073), breathcopd(0.0073)
Topic3	breathing(0.0530), sobsob(0.0530), difficulty(0.0446), cough(0.0418), cath(0.0307), cardiac(0.0307), blood(0.0303), trouble(0.0223), stent(0.0223), left(0.0206), weakness(0.0195), distress(0.0195), days(0.0185), foot(0.0168), low(0.0167), chfsob(0.0139), resp(0.0139), weak(0.0138), high(0.0112), onset(0.0112)
Topic4	breath(0.1954), short(0.0736), sobchf(0.0736), copd(0.0634), cp(0.0152), abn(0.0152),cath(0.0152), pleural(0.0152), fell(0.0150), feeling(0.0109), sobexacerbation(0.0101), effusion(0.0101), stresscardiac(0.0101), respiratory(0.0101), chfshortness(0.0101),severe(0.0101), passed(0.0101), home(0.0099), feel(0.0076), right(0.0076)
Topic5	chest(0.2080), pain(0.1891), painchest(0.0496), back(0.0213), sobpneumonia(0.0165),pressure(0.0142), coughing(0.0142), swelling(0.0136), acs(0.0123), arm(0.0118), uti(0.0095), pneumoniashortness(0.0095), left(0.0085), pulmonary(0.0071), leg(0.0071), last(0.0071), weaknesschf(0.0071),hard(0.0071), exacerbationsob(0.0071), edema(0.0071)

Table 6: The five topics automatically generated by the *Topic model*algorithm. Topics are described by the words most related to that topic.

Category	Variable name
Demographics	MARITAL_TYPE_Single
	SEX_Male
	RACE_Other
	RELIGION_Judaism (Jewish, Hebrew)
	RELIGION_Presbyterian
	RELIGION_Lutheran
	RELIGION_Non-Denominational
	RELIGION_Roman Catholic
	RELIGION_Christian Church
	RELIGION_Episcopal
	RELIGION_Unknown
Behavioral	Living Situation_Nursing home
	Sleep / Rest History_Early AM waking
	MARITAL_TYPE_Divorced
	Alcohol Use_Currently using
	Lives With_Alone
	Living Situation_Apartment
	Living Situation_2 story
	Tobacco Use_Unknown if ever smoked
	Lives With_Pets
	Living Situation_Other
	Tobacco Use_Never smoker
	Tobacco Use_Current every day smoker
Clinical	COPD
	Pneumonia
	Diabetes
	CABG
	Mstay
Text	Text_Topic_2
	Text_Topic_5

Table 7: Description of variables retained in the best performing algorithm