Hospitals at Risk: Predicting High Rates of Hospital Acquired Infections

Jacqueline Gutman

Alex Pine

Maya Rotmensch

Table of Contents

Business Problem

Hospital-acquired infections (HAIs) are a leading cause of patient mortality and other serious adverse outcomes in hospitalized patients. The **Centers for Disease Control and Prevention** (CDC) estimates that in the United States, there are approximately 1.7 million instances of HAIs per year, and that these infections result in roughly 99,000 deaths per year (Pollack, 2010). These types of infections are considered by healthcare and public health officials to be largely preventable, with outbreaks typically resulting from healthcare professionals failing to comply with established infection control practices or providing patients with contaminated devices or medications. The CDC, in concert with the **Department of Health and Human Services** (HHS) operates a National Health Safety Network (NHSN) to track the prevalence, spread, and compliance with prevention measures in over 12,000 acute-care hospitals and other medical facilities in the United States, with an annual operating budget of over $32 million (CDC, 2014). The CDC estimates that one in 25 hospital patients will develop a HAI during their stay, and that over 200 Americans die every day of infections acquired during their hospital stay (Society for Healthcare Epidemiology of America, 2014).

To address this pervasive healthcare crisis, the CDC and its Healthcare Infection Control Practices Advisory Committee provide U.S. healthcare facilities with evidence-based guidelines and recommendations demonstrated to reduce the rates of development and transmission of HAIs. Increased adherence to these recommendations has been empirically shown to be associated with a lower incidence of **central-line associated bloodstream infections** (CLABSIs), a particularly serious type of infection affecting up to 80,000 patients annually in intensive care units with a reported mortality rate of 12 to 25 percent and an estimated cost

ranging from $16,550 to $54,000 per infection (Joint Commission, 2012). In their most recent report to Congress, the CDC proposed a long-term objective of reducing the outcome measure of CLABSIs by 50 percent over three years, as part of the HHS National Action Plan to Prevent Healthcare-Associated Infections (CDC, 2014).

Unfortunately, the continued prevalence of these types of infections underlines the fact that not all healthcare facilities are strictly compliant in consistently adhering to the CDC's recommended best practices. In a study conducted in over 100 intensive care units of Michigan hospitals, researchers found that hospitals that received an evidence-based intervention encouraging compliance with CDC infection prevention guidelines attained a 60% overall reduction in the baseline CLABSI rate, saving an estimated 2,000 lives and reducing healthcare costs by $200 million over the 36-month period during and following the intervention implementation (Provonost et al., 2006). However, implementing such an intervention nationally would require additional staffing and funding, so it is necessary to optimize our resource allocation by first identifying the set of hospitals that are likely to have the highest HAI rate in the coming year.

For our project, we have chosen to build a model that will allow the CDC to predict which hospitals are most in danger of a high prevalence of infections, to allow for the selective deployment of infection control specialists to those facilities where an intervention would have the greatest impact on preventing HAIs. This will enable the largest possible net reduction in CLABSI rates given some resource constraints, bringing the national CLABSI prevalence closer in line with the target rates recently promised to Congress in exchange for continued funding of the National Health Safety Network.

Data Understanding

Data were obtained from the Medicare Hospital Compare Downloadable Database (Centers for Medicare and Medicaid Services, 2014). These publicly available data are intended to allow health care consumers to compare the quality of care at over 4,000 Medicare-participating hospitals, and comprise the back-end to the user-friendly interface on Medicare.gov's Hospital Compare website. Archived data are available for the years 2005-2014, with data collection typically completed over the previous year.

The target variable of interest is the CLABSI standardized infection rates, which are represented in the data in two different ways. First, by a raw score denoting the risk-adjusted measure of the standardized infection ratio (SIR). This ratio is calculated from the observed number of CLABSI cases in a particular hospital divided by the expected number of cases given the national rate of infection and the number of eligible cases. The following formula illustrates how the SIR is calculated:

$$\text{SIR} = \frac{\text{observed}}{\text{expected}} = \frac{\text{number of ICU cases} + \text{number of ward cases}}{\text{ICU central line days} \times \text{national ICU rate} + \text{ward central line days} \times \text{national ward rate}}$$

For hospitals where the expected number of CLABSI cases was fewer than 1 (for example, if the national CLABSI rate was 2 per 1000 days, and a hospital had only 400 central line days), the SIR was deliberately not calculated, in order to avoid making unstable conclusions about hospitals where no infections of this type would be expected even if no preventative measures were taken. Hospitals with missing target values were dropped from our analysis. However, it should be noted that these missing targets were not randomly distributed throughout

the data, but rather were typical of smaller or more specialized hospitals that had unusually low numbers of eligible cases as measured by the number of days per patient the hospital had a patient hooked up to a central line.

An SIR of 1 means that a hospital had the same rate of CLABSIs per 1000 central line days as the national average rate. A SIR below 1 indicates fewer infections than expected given the national rate, and a SIR above 1 indicates more infections than expected. For our analysis, we are interested in predicting which hospitals have a true SIR score greater than 1. Infections that were present upon a patient's hospital admission were not considered HAIs and thus were not reported in the data.

The second way in which hospital infection rates were represented was by a CDC provided label. Each hospital could be labeled as 'Better than national average', 'Worse than national average', or 'No different than national average', as a function of these SIR scores. Because we were interested only in identifying hospitals at high risk for transmitting infections to their patients, we collapsed the hospitals labeled as 'Better' and 'No different' into a single category representing no higher than usual risk of central-line associated infection. These labels were a function of the SIR score, where 'Worse than average' implied that a confidence interval on the score fell entirely above 1, and therefore that we have a high level of certainty the rate of observed infections in that hospital was greater the national average.

For each of the three years for which we performed the analysis, data were merged from 4 separate hospital-level databases. There were many more possible features from other

databases that we could have included, but we lacked the time to investigate them all. The ones we chose were those that seemed most likely to be predictive of the HAI rate.

The first of the four databases was the Hospital Acquired Infections dataset, which contained information on the values of the target variable (including both the binary labels and the continuous measures of the target), number of eligible and observed cases of HAIs of a particular type, as well as some basic geographic information about each hospital. The Timely and Effective Care dataset included several "process of care measures", such as the proportion of times postoperative patients were correctly given an antibiotic within an hour of surgery. The Medicare Payment and Volume Measures dataset contained information for each hospital on the number of Medicare patients treated for each of several serious illnesses or surgical procedures. Finally, the Medicare Spending per Beneficiary dataset contained information on the average price-standardized Medicare payout per patient, risk-adjusted to account for differences in the age and severity of illness of the patient, and standardized to account for geographic variation in cost-of-living. The range of data types of these features included binary predictors, proportions (such as the percentage of times a hospital administered the correct standard of care in a timely manner), other continuous measures (such as the average Medicare spending per beneficiary), and the standardized incidence ratio (SIR) scores for that hospital in previous years.

HAI data were available for years 2005-2014, but the data from 2005-2011 calculated the CLABSI scores differently, and featured a set of predictors that did not correspond well with the more recent data. After some data exploration and cleaning, it was determined that the earlier years of data would require more wrangling than it was likely worth, and we subsequently retained only three years' of data, those from 2012, 2013, and 2014. Later analysis of our data

showed that adding the 2012 dataset to the 2013 data did not yield any significant information gain beyond the performance possible with 2013 alone, and therefore we felt it was unlikely that the data from 2005-2011 would contribute substantial new information to the information contained in 2012 and 2013. We believed that the marginal gains possible with this additional data would not be likely to justify the resources required to acquire and wrangle seven years' of data spread across 85 distinct data files.

Data Preparation

To build our primary dataset, we first merged the data from 2012 and 2013, including the values of all potential predictors in both years, as well as the value of the target variables (by label as well as by score) from each year. All features from the year 2014 data were dropped to avoid leakage, and only the values of the target in 2014 were retained. Initially, the 2014 HAI data included information on 4,683 hospitals. However, after dropping 1,623 hospitals without a CLABSI score and 1,055 hospitals with less than one expected CLABSI case, there were only 2,005 hospitals remaining.

The CDC assigned each hospital a label as a function of its observed and expected CLABSI cases, computing a 95% confidence interval for each hospital that assumes the underlying distribution of observed cases is Poisson distributed and labels a hospital as 'Worse than national average' if the entirety of the confidence interval lies above 1, implying that the number of observed cases is greater than the number of expected cases throughout that interval (National Cancer Institute, 2014). However, this method of assigning risk labels is extremely conservative, resulting in only 23 positives labeled as high risk out of 2,005 hospitals with data

available in 2014. Because of this, we decided to explore several alternative indicator functions that would assign each hospital a label based on the numerical value of the SIR for that year. All of these were less conservative than the original indicator function developed by the CDC.

The first indicator function mapped CLABSI SIR scores that were at least two standard deviations above the national mean to 1. All other entries are marked as 0. . This function resulted in 97 positive labels in the 2014 data. The second indicator function mapped the highest 10 percent of the scores to 1, with scores below the 90^{th} quantile labeled as 0. This function resulted in 201 positive labels in the 2014 data. The third indicator function simply mapped all SIR scores greater than 1 to 1. All other entries are marked as 0. This function resulted in 296 positive labels in the 2014 data.

Even in our least conservative set of target labels, the rate of positives in the data is no more than 1 in 8. This presented a twofold challenge. First, we had to ensure that there were sufficient positive instances in the training data so as to train our model effectively. Second, we needed to ensure that we had sufficient positives in the holdout set so we could meaningfully evaluate our model. To resolve this, we created a stratified sampling function that forced the the number of positives in the training and holdout datasets to be proportional to the size of the training and holdout set respectively. In other words, if we created a 80/20 split we forced the training dataset to include 80% of the overall positives.

The data also included the full address of each hospital, including city, state, and zip code. We decided to only include the US state in which the hospital was located, figuring that the city and street information was too granular to be predictive of HAI rates. Each state was

represented as its own dummy variable. A hospital was marked as positive for the state in which it was located, and negative for all other states.

When a datapoint was missing from a dataset, we replaced it with the mean of its column, unless it was a categorical variable, like the hospital's US state, unless we could determine that the data was systematically missing in a meaningful way. For example, in the data concerning the number of times a particular procedure was performed that year, a missing data point indicated that the hospital in question did not perform that procedure more than 10 times that year, and so missing values were imputed as 10, the maximum possible value for that hospital. To avoid doubling the number of features given our limited number of observations, we did not add indicator variables for each feature to mark where missing data values had been imputed.

Modeling and Evaluation

In order to predict which hospitals were likely to have elevated infection rates in 2014, we created 3 types of logistic regression models, each building upon the previous in order to enhance the predictive power of our model. First we started with the **Pure Autoregressive Model**, where we predict the level of risk of a hospital based solely from its risk level or SIR scores in previous years. If this type of prediction turns out to be the best we can do, then there is no need to invest resources in the data collection and modeling that would be required for more complicated models, and our business recommendation would simply be to deploy a team of infection control specialists to the hospitals with the greatest risk-adjusted rate of HAI transmission over the previous year (or set of years). Next we built upon the Autoregressive model creating a **Full-Featured Logistic Regression Model** incorporating up to 205 features.

Since the number of positives in the vector of labels originally provided in the data is quite small, we then turned to a **Proxy Logistic Regression Model** in the hopes of creating more meaningful predictions for these labels. Lastly, in an attempt to ascertain which features contribute the most to our model and in order to avoid the 'curse of dimensionality', we performed stepwise feature elimination.

Autoregressive Model

In this model, we attempt to predict target values for the current year (2014) as a linear function of available scores of the corresponding measure in previous years (2012 and 2013). Because the limitations of our data source allowed us to input only two years' worth of data in the autoregressive model, it was not possible to consider any complex path dependent relationships, such as a sensitivity to the rate of change from year to year, that may have been discoverable with more years' of data. As the autoregressive model is the least complex model we built, its performance serves as a baseline to which we compare the more complex models.

For this initial analysis, we created two types of autoregressive models that differed from one another in feature engineering. In the first model, we used binary labels indicating high infection rate (created by the aforementioned binning functions) in the previous years as features. In the second model we chose to use the continuous SIR scores of previous years as features.

The model utilizing the binary infection rate labels was trained on either the CDC-supplied category labels or our own less conservative labels, using one or both years of available data. We compared the models' ability to predict the 2014 classification of a hospital only from its classifications over the previous years. For each of the four binning indicator

functions, we applied that same indicator function to the SIR measures for each of the three years available, transforming the targets and predictors into simple binary features, and built a logistic regression model that predicted the classification of 2014 from either 2013 alone or 2013 and 2012. These models were cross-validated to determine the optimal value of the regularization hyperparameter, and then evaluated on various performance ,metrics. In most cases, there was almost no difference in the **area under the receiver operating characteristic curve** (AUC) between the model based on 2012 and 2013 data compared to the model based solely on the previous year's data (2013), and in some instances the additional year of data actually hurt performance slightly. Overall, it seems that any latent information within the 2012 data was subsumed by the more recent data from 2013. However, using the indicator functions on the historical data as well as the target variable necessarily created a very coarse-grained and unstable model that did not make full use of the information contained within the continuous 2012 and 2013 CLABSI SIR score variables. The ROC curve for this model, using both 2012 and 2013 data, is shown in Figure 1.

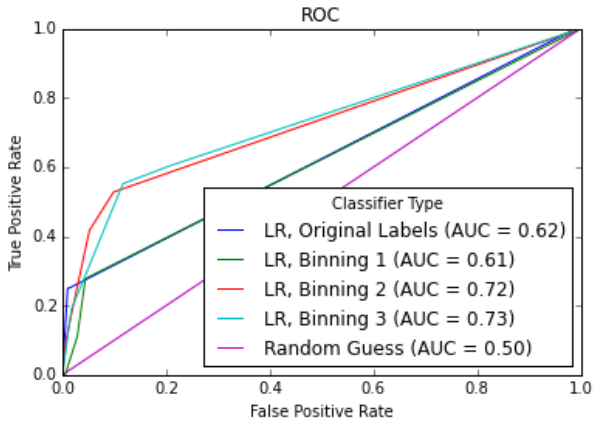

Fig. 1: Logistic Regression, trained on binned 2012 and 2013 HAI SIR scores.

To remedy the error induced by the oversimplification of this model, we applied each of the indicator functions only to the target variable and left the 2012 and 2013 predictors in their original continuous form. For all variations of the indicator functions on the target and the years of data included in the model, using the continuous predictors improved performance and reduced variance compared to using the predictors binned in the same binary form as the target. Performance of the models using the raw SIR scores for 2012 and 2013 ranged from an AUC of .70 to .78, as compared to an average AUC performance of .61 to .73 on the logistic regression using the target labels of previous years as predictors. However, using the CDC-supplied labels as our target still yielded extremely variable results from one instantiation of the model to the next, as the extreme sparsity of positives fostered a highly unstable model. In comparison, the models based on the less conservative targets tended to be much more stable. The ROC curve for this model is shown in Figure 2. Performance was nearly identical between the models built on 2013 and 2012 data compared to the model predicting 2014 classification from 2013 SIR scores alone, confirming the earlier findings that the older data did not appear to contribute any additional information not contained within the data from the most recent year.

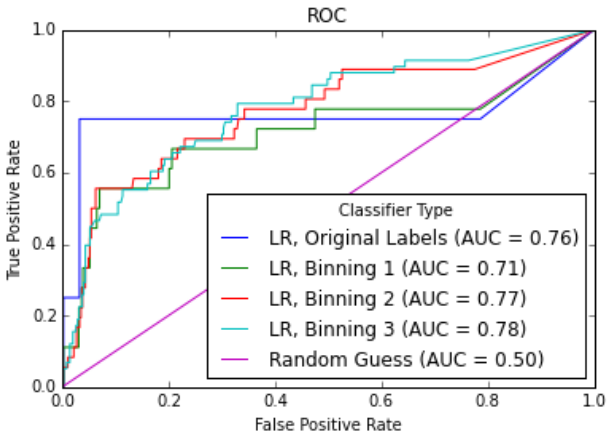

Fig. 2: ROC curve of the autoregressive model, trained on 2012 and 2013 HAI SIR scores, predicting the corresponding 2014 HAI SIR labels. Note that although the curve for the original label has a high AUC, this curve changes completely when tested on different data.

Full-Featured Logistic Regression

Building upon the autoregressive model , we want to ascertain whether hospital-level features other than the infection rate of previous years had any predictive value in estimating the current infection rate. We extracted the 203 year-indexed features introduced earlier for the years 2012 and 2013. We first reproduced the autoregressive model built in the previous step that used the raw SIR scores of 2012 and 2013 to predict which hospitals experienced a greater number of CLABSI cases than expected, given the national average rate of infection (this instantiation of the target variable results in 296 positive cases out of 2005 hospitals, or the worst 15 percent of hospitals), and then added to these two features the 203 other hospital-level features extracted from the Medicare datasets. Doing so actually slightly hurt accuracy at several possible prediction thresholds, resulting in an AUC of .73 (using L1 regularization) or .74 (using L2 regularization), down from the AUC of .78 in the corresponding autoregressive model. The ROC curve for this model, using raw 2012 and 2013 SIR scores as predictors, is shown in Figure 3.

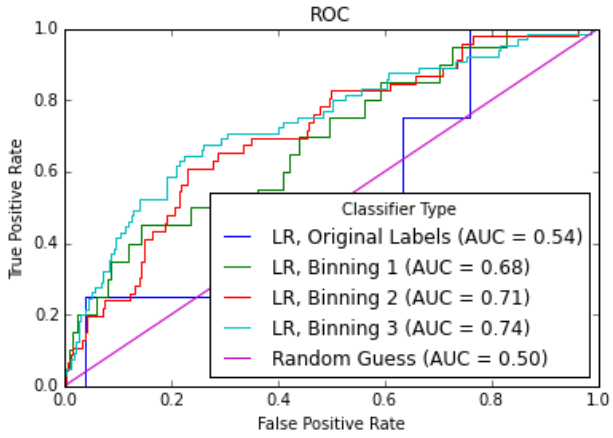

Fig. 3: ROC curve of the model with all available features, used to predict all four indicator labels. The predictors used here are the raw 2012 and 2013 HAI SIR scores. The model was tested using L2 regularization and a cross-validated regularization parameter.

Proxy Model

As we have previously shown, while our model is fairly adept in predicting the target variables we created, the models' performance for the original labels provided by the CDC is less than optimal. One difficulty in making predictions on the provided labels is that the number of positive labels in the data is quite small. In the world of online advertising, where the number of positives in a dataset can range from 1 in a thousand to 1 in a million, it has been demonstrated that proxy modeling and optimization—which trains a model on a surrogate measure whose covariates look distributionally similar to the covariates of the actual target of interest but that occurs more frequently in the data set—can make very good predictions on the value of the actual target variable of interest (Dalessandro, Hook, Perlich, & Provost, 2012). As mentioned earlier, training on the small number of positives within the set of CDC-assigned labels lends itself to a large degree of variance in the evaluation of the optimized model, so we decided to test if we could reduce the variance by training our model on a proxy variable. The target variable of interest here is whether the CDC assigned a label of 'Worse than average' in their CLABSI infection rate, but our indicator functions which assign a label of high risk as a function of a hospital's SIR score should be a good proxy to the true target, as both target and proxy are ultimately functions of the number of documented, risk-adjusted cases of hospital acquired infections of a particular type during a given year. Therefore, the proxy variable identifies a subset of hospitals which encompasses the original target subset and adds to this other hospitals which also have an alarmingly high risk of infection, but have just missed the cutoff to be labeled as 'Worse than average' by the CDC.

Before using the proxy regression to predict the CDC provided labels, we first wanted to establish how the proxy measure performed on a target variable that was sparse but that we were already predicting relatively well. To do that, we chose as our target variable the most conservative indicator function that we created (which still had twice as many positive as the original labels). The purpose for this was twofold. First, we wanted to confirm whether the proxy behaved relatively similar to the normal regression on a target variable which our model could already predict relatively well. If the proxy regression performed significantly worse, we would have cause to doubt the utility of performing this regression on the original labels. Second, we wanted to know if using a proxy target that is still relatively sparse might improve upon the basic model's predictive power. As shown in Figure 4, the results of proxy regression for the most conservative target variable (shown in green) are very similar to the results of standard logistic regression on the same target (shown in blue).

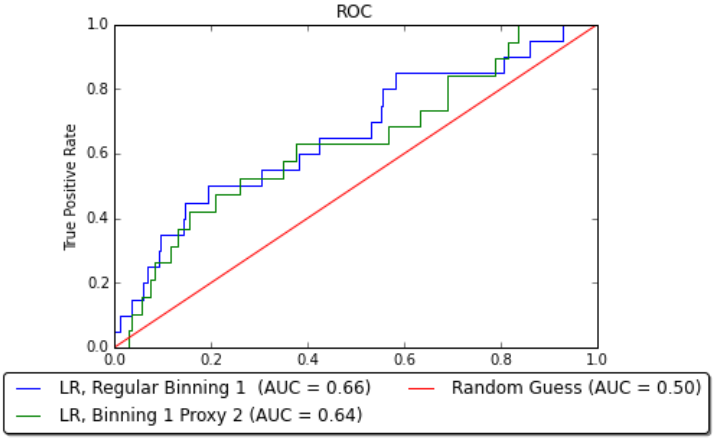

Fig. 4: ROC curve comparing proxy and regular regression. In blue is the curve created by regular logistic regression. In green is the curve created by proxy logistic regression. The target variable is the most conservative target label we created. The proxy variable is the second most conservative target created.

After establishing that proxy regression is a valid method for our data, we built a proxy logistic regression model with the original CDC provided labels as the target variable. We then

cycled through all our indicator functions and found that the second most conservative target label constituted the best proxy for our target variable (contains 201 positives compared to 23 in original). Figure 5 shows typical results for the ROC curve comparison between the regular logistic regression and proxy logistic regression on the original labels. As we we can see, the proxy regression (in green), is significantly better than the ordinary logistic regression (blue).

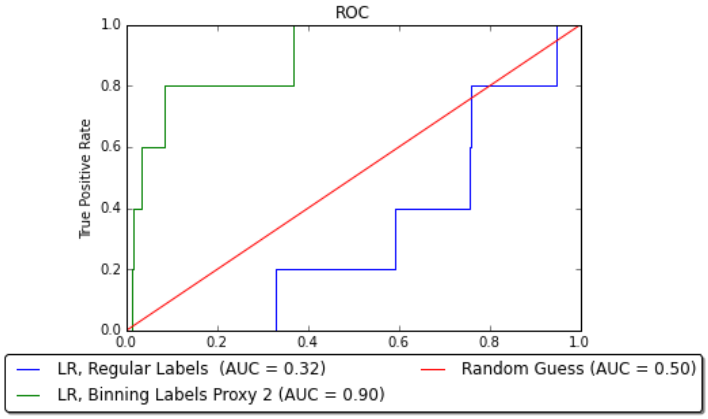

Fig. 5: ROC curve comparing proxy and regular regression. In blue is the curve created by regular logistic regression. In green is the curve created by proxy logistic regression. Target variable is the original CDC provided labels. The proxy variable is the second most conservative label created.

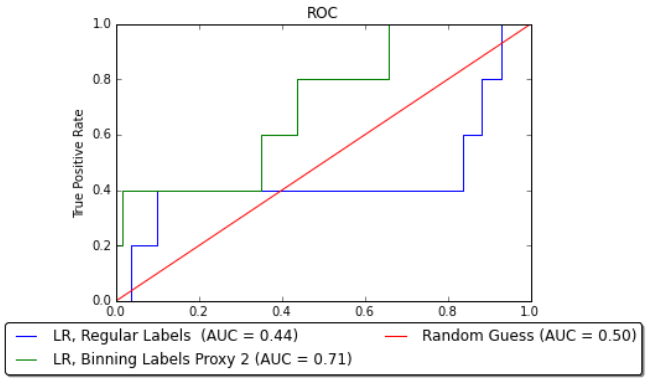

Fig. 6: ROC curve comparing proxy and regular regression, another instantiation of the model shown in Figure 5. Comparison of the two graphs demonstrates the high level of variability in results.

Comparing Figures 5 and 6 shows the variation in AUC between iterations. While the specific values of the AUC vary considerably, the proxy regression consistently outperforms the standard logistic regression. From comparing the autoregressive models in Figure 2 and 3 to

these two models, we conclude that using our less conservative target labels (which allow a greater number of hospitals to be tagged as high risk compared to the CDC-provided labels), we cannot do better than the standard logistic regression. However, when our goal is to predict the original labels, proxy regression is a very successful alternative that consistently outperforms the standard regression but produces highly variant results.

Feature Selection

Because our full-featured logistic regression actually performed slightly worse than our baseline autoregressive model, we proceeded to pare down the size of the feature set through recursive feature elimination. Performing this procedure with the most generous indicator function as the target variable (which produces our most stable model), we found that all features were eliminated except for each hospital's SIR scores from the previous years, and resulted in an AUC of .75, using the default logistic regression parameters on the remaining features (see Figure 7). Since this implies that there is little information in our non-SIR features, we decided to determine whether there was valuable signal in these additional 203 features that was simply being obscured by the predictive strength of the previous years' SIR scores. We ran the stepwise regression again, but with the 2012 and 2013 scores removed from the full feature set. As shown in Figure 8, this resulted in all but the last remaining feature being eliminated, and resulted in a prediction that was no better than random guessing (AUC of .51). These results imply that there is no signal contained within any of the features except for the SIR scores, and that there is no improved predictive ability to justify the cost of collecting the data from these disparate sources of information and incorporating them into our baseline autoregressive model.

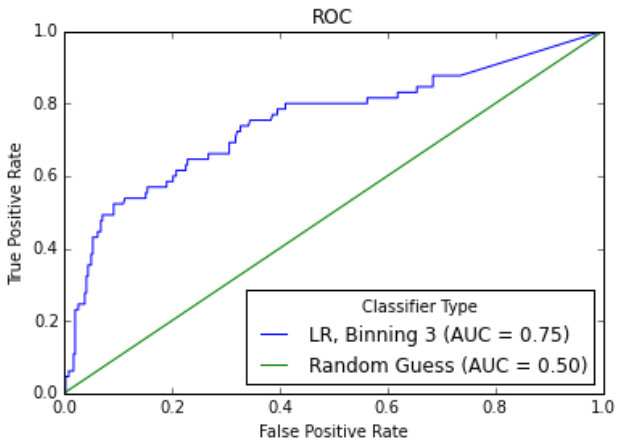

Fig. 7: ROC curve of the model after all non-HAI features have been removed through recursive feature elimination.

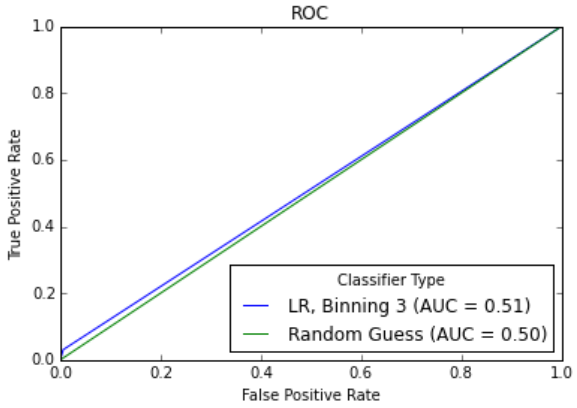

Fig. 8: ROC curve of the model was trained on all features except the HAI SIR scores.

Model Evaluation Metrics

In our analysis we used ROC curve analysis as the main method for evaluating our model. We chose to do so because ROC curves give an intuitive understanding of how the true positive rate and false positive rate change depending on a shifting probability threshold. Such intuitive and graphical results are also useful in a business presentation context, when we are trying to convince the CDC which of our models they should utilize. Nonetheless, in our business context, it is not appropriate to assume that the cost of a false negative would be comparable to the cost of a false positive. We would rather deploy a team of specialists to a hospital that has an infection rate no worse than average than fail to identify a hospital that is in

fact killing a high percentage of its patients. Optimizing for AUC assumes that the cost of each type of error is the same, therefore, while the ROC curve and AUC are still a relevant metric for analyzing our models, we should not blindly use its results for model selection since it ignores the relative costs of these decision thresholds (Provost & Fawcett, 2001). For this business problem we would rather favor recall over precision. However, we cannot simply optimize for recall because this would entail predicting all hospitals as high risk.

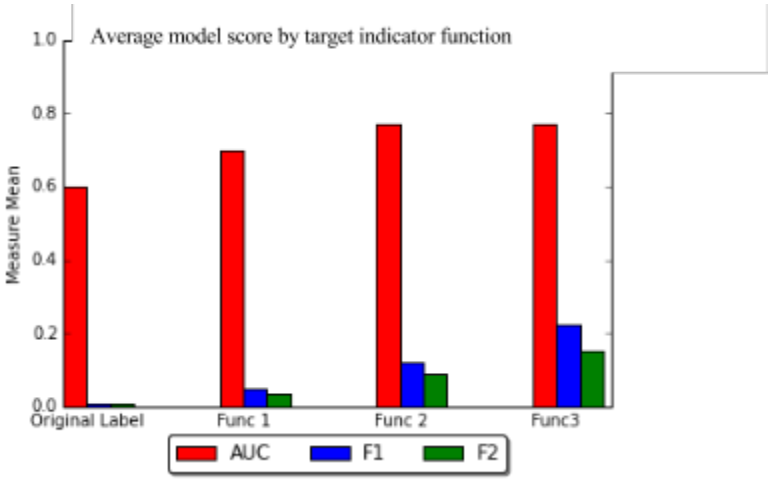

Fig. 9: A comparative view of how the autoregressive models performed for target variables ranging from most conservative (left), to least conservative (right).

Therefore, we considered two alternative metrics to optimize in evaluating the model. The first is the F1 score, which is a weighted average of the precision and recall of a model, and the second evaluation metric is the F2 score, which is similar to F1 but gives twice the weight to recall as compared to precision. After creating models that optimized for F1and F2 scores, we observed that the results of the autoregressive models varied considerably depending on which indicator function we used to define the target variable. In particular, the scores for each measure increased sharply as a function of the number of hospitals labeled as positive in the target data. This suggests that while the AUC scores do not vary much between the different indicator functions, there is a marked improvement in both F1 and F2 scores when utilizing the more

inclusive indicator functions as target variables (shown in Figure 9). Therefore, if we do want to create a balance between precision and recall, we might want to use the optimization for F2 scores and use the more generous variants the target variable rather than the original labels.

Deployment

In considering how this family of models might be deployed nationally, there are some considerations we must make note of. The ultimate goal of our model is to inform data-driven decisions about where we should send teams of infection control specialists. This is a resource allocation problem in which we are trying to optimize the number of lives saved while constraining the cost to the CDC. Hospitals selected to participate in an intervention will also benefit from a significant cost-savings in the number of infections incurred.

The CDC's budget will determine which indicator target we will use. If the CDC's budget is expansive enough to allow for sustained intervention at more than just the 1 percent or so most dangerous hospitals explicitly identified by the CDC's 'Worse than average' label, we should use one of our more comprehensive labels. In this case, the best we can do is the pure autoregressive model, using one of our more comprehensive indicator functions to set the target score, and using the raw SIR scores, for the past year or years of data as the predictors for our model. However, if it is important to our client that we predict the originally provided labels, the best model is the proxy regression, using the second least inclusive our indicator functions as the proxy target variable and retaining the CDC provided labels as the true target of our predictions.

Lastly, whichever labels we use, we still need to decide where to send teams of infection control specialists, and how many teams to send. If we were given the cost to the CDC for an

intervention, and a monetary estimate for the benefit of a successful intervention, we could use expected value analysis to figure out exactly how to set our model's positive prediction threshold to maximize the expected benefit. However, even if we had these numbers, this analysis wouldn't take into account the cost of not sending a team to hospital that needs it, because that cost does not directly affect the CDC.

To avoid having to estimate our model's positive prediction threshold, we could determine the set of hospitals to which we should send experts by ranking the hospitals from most to least likely to need an intervention using the model's probability score for each hospital. Using this method, we do not need to design a principled way of choosing a positive classification threshold for our model. We can simply rank the results of the model, and send as many teams as our budget allows. However, a downside to this method is that if the budget can fund more interventions than are actually needed, the recommendation will be to deploy specialists to hospitals predicted by the model to have a very low probability of infection.

Final Thoughts on Future Deployment

While the model we created is promising in its possible practical applications, there are several limitations that must be addressed before it is deployed nationally. The first and most notable concern is the considerable time lag between the time period during which the data is collected and the point at which it is made publicly available. The data typically represent the hospital acquired infection rates from the year ending 12 to 15 months prior to the point at which the data become available, implying that if the data are used to predict the hospital acquired infection rates of the subsequent year, they are rendered obsolete by the time the data is posted. If the data are available internally to the CDC at an earlier date, that would alleviate this concern;

otherwise, two solutions are possible. One solution is to build a model that does not make use of the preceding year, but only utilizes data from at least two years in advance. A better solution is to simply recommend a greater allocation of resources towards streamlining the data collection process to make it more dynamic and timely, eliminating paper reporting and end-of-the-year reporting of infections by allowing infections to be directly reported to a centralized database at the time of diagnosis. The funding for this initiative could be drawn from the resources previously devoted to data collection on the many features we found to be entirely uninformative for our model, as the cost of collecting these data is not justified by any improvement in our ability to predict high risk of HAIs.

A second concern is related to the generalizability of our findings. In our analysis, hospitals with missing target values for 2014 were dropped from our analysis, as is customary in building supervised learning models. However, it is important to note that we do not believe the distribution of missing target values to be random. As previously mentioned, the CDC made a conscious decision to refrain from computing either SIR scores or binary target labels for hospitals whose expected number of cases, based on the national average rate and the number of central-line days per year at that hospital, was too low. Therefore, the hospitals that have been excluded from our analysis because expected number of cases was less than 1 are not likely to be distributionally equivalent to the hospitals included in our model, and our findings may not extrapolate well to these hospitals. Out of the 3060 hospitals for which data was available, 1055 hospitals were thrown out for this reason. These hospitals may exhibit such a low number of central-line days because they are smaller than average, more rural, more specialized, or because they use non-standard methods of care. Thus it is critical to discuss with the CDC whether they

are content with these kinds of hospitals being left out of our analysis, or if it is important to make inferences about these kinds of hospitals as well. Furthermore, several specialized hospital types—most notably, veterans' administration hospitals and children's hospitals—were excluded from the analysis because no CLABSI data was available for these hospitals. It may be the case that different kinds of interventions are appropriate for these hospitals, and if the CDC considers these to be a priority it will be critical to obtain the data relevant to these types of hospitals.

Another point to consider is that while identifying the hospitals with the highest rate of infection is a reasonable predictive goal, if our business objective is to maximize the number of infections we can prevent given a limited set of resources, it may actually be preferable to identify those hospitals with the greatest number of infections, regardless of rate. For example, a hospital whose rate of infection is only slightly worse than average, but which treats a large number of central-line patients, is likely to yield a greater absolute reduction in the number of infections if it is targeted for intervention, as compared to a hospital which treats a smaller number of patients but demonstrates a high rate of infection for the patients it does treat. If reducing the number of infections nationwide as much as possible is our highest priority, then it is worth considering redefining our target variable as the number of observed cases of CLABSI infections in a particular hospital per year, and then building a linear regression on this continuous target variable instead of our logistic regression. There is no clear-cut answer to the question of which framing of the problem is a more appropriate one, and this should ultimately be determined by the CDC and its sources of funding.

Appendix

Jacqueline Gutman

- Obtained the original dataset. Researched the variables used in the dataset and the method for calculating the scores. Collaborated with Alex on parsing the data.
- Pulled the data for 2005-2014 and converted to usable format. Analyzed the distribution of missing data, distribution of target labels, and target scores.
- Wrote the indicator functions used to transform the raw scores to binary variables. Merged and analyzed the data from the Volume of medical and surgical procedures datasets.
- Wrote the first drafts of project paper and project proposal, and collaborated on the final draft.

Alex Pine

- Set up the github repository (https://github.com/pinesol/intro-ds-hai/). Wrote much of the code used to prepare the data and to train the models. Collaborated with Maya on the stratified sampling function. Imported and merged the SCIP process of care data.
- Evaluated the performance of the autoregressive model, as well as the full-featured logistic regression. Performed the feature elimination analysis.
- Prepared and gave half the oral presentation. Collaborated on writing this paper.

Maya Rotmensch

- Assisted with data cleaning and merging. Collaborated with Alex on creating the stratified sampling function. Imported and merged the Medicare spending data.
- Created the proxy regression models and evaluated their performance under different evaluation metrics and target labels.
- Created functions to evaluate models on alternative metrics to AUC (F1 and F2 scores).
- Prepared and gave half the oral presentation. Collaborated on writing this paper.

Bibliography

Centers for Disease Control and Prevention. (2014, March 7). *FY 2015 Congressional Justification of Estimates for Appropriation Committees.* Retrieved from http://www.cdc.gov/fmo/topic/Budget Information/

Centers for Medicare and Medicaid Services. (2014, October). *Official hospital compare data* [Data file]. Retrieved from https://data.medicare.gov/data/hospital-compare

Dalessandro, B., Hook, R., Perlich, C., & Provost, F. (2012). Evaluating and optimizing online advertising: Forget the click, but there are good proxies. *NYU/Stern School of Business: Center for Business Analytics,* Working Paper CBA-12-02. Retrieved from http://hdl.handle.net/2451/31637

The Joint Commission. (2012, May). *Preventing central line-associated bloodstream infections: A global challenge, a global perspective.* Oak Brook, IL: Joint Commission Resources. Retrieved from http://www.jointcommission.org/assets/1/18/clabsi_monograph.pdf

National Cancer Institute: Surveillance, Epidemiology, and End Results Program. (2014, March). *Standardized incidence ratio and confidence limits.* Retrieved from http://seer.cancer.gov/seerstat/WebHelp/seerstat.htm#Standardized_Incidence_Ratio_and_Confidence_Limits.htm

Pollack, A. (2010, February 26). Rising threats of infections unfazed by antibiotics. *The New York Times,* p. B1. Retrieved from http://www.nytimes.com/2010/02/27/business/27germ.html

Provonost, P., Needham, D., Berenholtz, S., Sinopoli, D., Chu, H., Cosgrove, S., …, Goeschel, C. (2006). An intervention to decrease catheter-related bloodstream infections in the ICU. *New England Journal of Medicine, 355*(26), 2725-2732. http://www.nejm.org/doi/pdf/10.1056/NEJMoa061115

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning, 42*(3), 203-231. Retrieved from http://people.stern.nyu.edu/fprovost/Papers/rocch-mlj.pdf

Society for Healthcare Epidemiology of America. (2014, March 28). *Testimony of SHEA/APIC to the U.S. House of Representatives Appropriations Subcommittee on FY 2015 federal funding priorities.* Retrieved from http://www.shea-online.org/View/smid/428/ArticleID/268.aspx