Machine Learning Based Prediction of Patient Readmission Using CMS Claims Data

Robert P. Yerex, Ph.D

March, 2015

UVA Medical Center Accountable Care Organization

**Objectives**. To develop a predictive model using statistical learning techniques to rank recently discharged patients on likelihood of readmission within 30 days of their discharge.

**Data Sources.** Claim and Claim Line Feed (CCLF) files provided by the Centers for Medicare & Medicaid Services (CMS).

**Statistical Learning / Predictive Model.** A Random Survival Forest (RSF) algorithm was used to predict each patient’s hazard function across the first 30 days post discharge.

**Model Performance.** The most recently trained and tested incarnation of the model had a pairwise concordance index of 0.72 averaged across the first 30 days post discharge.

# Background

The direct and indirect costs of unplanned patient readmission to hospital are staggering. By one estimate, 19.5% of all Medicare beneficiaries who were discharged from a Hospital (during 2003) were readmitted within 30 days, leading to incremental costs of $17.4 billion [1]. In addition to these direct costs, preventable readmissions compete for resources within the hospital, decreasing the quality of care and increasing costs for all patients. Costs are not equally distributed among Medicare beneficiaries. For example, in 2006 four percent of Medicaid enrollees were responsible for 48 percent of all Medicaid spending [2].

While not all, many readmissions that occur within 30 days of discharge may be preventable. The Centers for Medicare & Medicaid Services (CMS) estimates that if all hospitals attained reductions in 30 day readmission comparable to the best in class facilities, the annual reduction in costs would be $1.9 billion. It is well understood that interventions applied immediately post-discharge can have a significant positive impact in reducing unplanned, preventable readmissions [3]. In order to most effectively apply such intervention strategies on a cost effective basis, relative risk of readmission can be used to determine which patients may be most effectively targeted [4].

# Review of Published Readmission Risk Models

Readmission risk models generally fall into one of two broad categories, those used largely designed to facilitate calculation of risk-standardized readmission rates for hospital comparison purposes, and those used to identify high-risk patients. In 2011, Kansagara et. al. [4], systematically reviewed the performance of 26 predictive models targeting 30 day readmission, half of which were risk adjustment models with the other half being used to identify high risk patients. For purposes here, it is the identification of high-risk patients that is of interest. In the Kansagara study, as is also the case for all other models reviewed by this author, the outcome, or target variable was binary and can be stated as *was the patient readmitted with 30 days of discharge, or not*. The vast majority of the individual patient outcome targeted models in the Kansagara study used some form of regression adapted for binary outcomes, typically logistic regression. There are many alternative approaches which to logistic regression, most with better performance characteristics; however these alternatives are more mathematically challenging, computationally intensive, and the resulting models are less amenable to interpretation than the less sophisticated approaches using logistic regression [5].

Among the models in the Kansagara study, as with predictive modeling in general, the most critical factors impacting predictive performance were feature selection and data source. Models incorporating features related to near real time clinical measures at the time of patient discharge had performed significantly better than those that relied solely on retrospective administrative data.

# Data

The data used for development of the predictive models in this study were extracted from the Claim and Claim Line Feed (CCLF) files provided by the Centers for Medicare & Medicaid Services (CMS), to participating Accountable Care Organization (ACO) [6]. Because they are transaction oriented, considerable data manipulation was required in order to transform the CCLF records into a set of event history records each patient describing one or more spells of care that involved one or more episodes of inpatient hospital care[[1]](#footnote-1). The final data set included 5,364 patients with a total of 10,167 inpatient care episodes encompassed within 9,461 care spells. The earliest inpatient event occurred on 2013-01-01 and the last on 2014-09-17.

# Methods

As described earlier, the performance of the currently published models designed to predict 30 day readmission is poor, suggesting that there are relationships between factors

attributable to the patients and their environment, the course of their treatment, the progression of their illness, and the likelihood they will be readmitted within 30 days of discharge, which are too complex to be captured with simpler predictive model algorithms targeting categorical outcomes. In many fields where predictive modeling is used, the desire for the most accurate prediction far outweighs the need for interpretability; this is not generally true in the medical field where there is a tension between prediction and interpretation. If a model is to be designed to predict patient readmission as accurately as possible, it should not be constrained by the requirement for interpretability. Kuhn and Johnson, among others have stated that in a medical setting, it would be unethical to adopt a model that is more easily interpreted at the sacrifice of accuracy. “*As long as the model can be appropriately validated, it should not matter whether it is a black box or a simple interpretable model*” [5].

## Survival Modeling

Time to patient readmission data is amenable to the application of survival modeling, a collection of statistical techniques that take into many of the issues inherent in dealing with time to event data including censoring and competing risks [7]. While a variety of algorithms are available for estimation of the survival and related hazard functions in the presence of covariates. However most of these methods rely on restrictive assumptions such as proportional hazards, and are typically parametric in nature, requiring assumption of first, second, and third moments of the generating functions associated with the underlying selected parametric density family. Depending on the assumed density function, nonlinear effects of variables must be handled through transformations or expansion to include specialized basis functions. When there are multiple, possibly interacting covariates present, they are difficult to identify, and typically involves the researcher examining all two-way and three-way interactions, possibly relying on subjective knowledge to narrow the search.

Non-parametric methods exist for the estimation of the survival function [8], such as the Kaplan–Meier estimator [9]. The advantage of these methods is their ease of use, but they are most suitable for controlled cohort analysis.

### Statistical Learning Technique for Survival Function Estimation

Several methods within the area of statistical learning can be used for functional approximation. Cybenko’s theorem [10] proves that a type of neural network (ANN) which is equivalent to the superimposition of multiple sigmoidal functions, can be used as a near perfect approximator for arbitrary monotonic functions. Unfortunately, ANN methods do not easily take into account censoring as occurs in survival analysis. Random Forests (RF) Another well-known statistical learning method that can be thought of as universal function approximators [11]. RF is a subset of a more general class of ensemble learning[[2]](#footnote-2) based techniques that iterate over combinations of base or weak learners with the resulting learner (the ensemble of the iterated base learners) having greater predictive power . While RF is generally used for regression and classification, Ishwaran et al. [12] have extended the technique specifically for use in survival analysis and is known as Random Survival Forests (RSF). In RFS, the splitting criterion used in growing a tree explicitly invokes survival time and censoring information. The effectiveness of a particular split is measured via the difference in survival expectation for inclusion in each of the new nodes below the split.

# Model development, TRAining and testing

A Random Survival Forest (RSF) was developed using the techniques outlined by Ishwaran et al. [12]. The initial data was split into training and test sets of 7,095 and 2,366 respectively. The parameters that can be controlled when using the particular implementation of RSF [13] include; number of trees constructed (ntree), number of candidate features to try at each split (ntry), minimum number of cases in a terminal node (nodesize), the maximum depth of any tree (nodedepth), and the splitting rule (splitrule).

## Targeted Outcome / Application of Predictions

It is important to understand how the resulting predictions obtained from a model will be used in practice. Much of the literature describes models where the purpose is to predict the likelihood that a particular patient will be readmitted within 30 days of hospital discharge. In isolation, this is not of much practical use. In may be possible to use 30 day readmission predictions to support efforts to reduce the rate of such readmissions. Some recent studies have provided indications that improved post discharge care management results in reduced 30 day readmission rates [14], while others have found no measurable improvement in readmission rates [15]. Some of the most recent work in this area indicates that under some circumstances, surgical procedures in this case, readmission rates are not tied to the quality or level of post-discharge care coordination [16]. This range of findings linking post-discharge care coordination and readmission likely may point to one important issue, effective care coordination requires prioritization. Caseloads for care coordinators are likely to be quite large, and it is understood that the first few days post-discharge are the most critical [17]. In this setting, a tool that assists the care coordinator in prioritizing their cases will allow them to more effectively target their resources for best effect.

The application through which the predictions will be delivered is dubbed *Relative Readmission Risk Monitor* or R3M (R3M) for short, and is designed to assist care coordinators in prioritizing their caseload when dealing with recently discharged patients. The use case is fairly straightforward; a patient care coordinator, responsible for a subset of all the ACO patients within the UVA ACO population is provided a ranked list of patients who have been discharged over the previous 30 days. The order of ranking is on relative risk of readmission within 30 days of discharge which is calculated by integrating over the predicted hazard function  from the current offset from discharge (i.e. the number of days since discharge for a particular patient) through 30 days. This value, termed *cumulative remaining risk* or CRR can be expressed as:

,

where

,

and

,

Where  is the remaining risk for patient  at timewith the estimates for  having been generated from the RSF model[[3]](#footnote-3).  here is the survivorship function [7]. Ishwaran et. al. show that the Cumulative Hazard Function (CHF), estimate for  is in fact the Nelson-Aalen estimator [12] which can be expressed as:

,

with  the number of readmission at time  and  the total number of patients at risk for readmission at time  [8]. is calculated for each patient  having been discharged within the previous 30 days (i.e.  takes on values from 1 through 30), and then these values are ranked on an ordinal scale from highest to lowest[[4]](#footnote-4).

The ranking is dynamically updated daily. It is likely that each care coordinator will have a large number of patients assigned to them, a significant portion of which will have been discharged within the previous 30 days. Given their limited time and resources, the greatest impact in terms of reducing 30 day readmission rate will be gained through optimal application of effort to the care coordinator’s patient set. The ranked list assists the care coordinator in determining which patients to focus on. This is to be used as an adjunct to the care coordinators understanding of each patient’s situation and is not a replacement for judgment.

## Feature Selection

The process of feature selection was carried out through inspection of the *variable importance* statistics as the model was iteratively developed. Several techniques were used to avoid over-fitting; measures of performance based on bootstrapped cross-validation were used rather than apparent performance [19], the extent of any over-fitting was monitored using a modified version of Copas’ shrinkage [20]. The final versions of the model contained only 16 predictors and a training set of over 7,000 care spells reducing the likelihood of over-fitting.

Features in the final model are of two forms; simple and derived. Simple features are directly related to attributes in the data set[[5]](#footnote-5). Derived features are generated through, sometimes complex, combinations and transformations of attributes in the data set. Features fell into several broad categories;

* Demographic: Age, gender and race.
* Socioeconomic: With the patient’s address, it was possible to determine the US census block in which they resided [21]. Using this information it was possible to extract information on a variety of socioeconomic factors [22]. Included in the model were factors related to median income, and level of education.
* Access to Care: With the location information, it was also possible to create several features used as surrogates for ease of access to care.
  + Distance[[6]](#footnote-6) to PCP
  + Distance to admitting facility. For spells covering multiple claims and multiple facilities, the distance to the initial admitting facility was used.
  + Density of care providers within zip-code.
* Medical Condition:
  + ICD9 Diagnosis Codes:
    - At the individual claim level for principle and admitting diagnoses, at the claim level.
    - When a single episode and/or spell of care included several individual claims, the set of unique codes was sequentially combined into a single hashed value which then represented a category in this feature.
    - Count of total number of unique diagnoses codes.
  + Charlson Comorbidity Indicators: These were derived using the entire set of unique diagnoses removing those that were considered to be principle diagnoses.
* Care Process:
  + Admission Type: Emergency, Urgent, Elective, Trauma Center, Unknown
  + Admission Source: Physician referral, Clinic referral, HMO referral, Transfer from Other Hospital, Transfer SNF, Transfer Emergency room, Transfer ASC, Transfer Hospice.
  + Transition / Discharge Patterns. Coding from the original CCLF data files provides 33 codes for patient disposition upon discharge [6]. When there are multiple claims within a spell of care, the sequence of discharge codes describes a sequence of transitions. Nineteen distinct sequences were observed with 67% limited to a single transition, 18% included two transitions, 9% involved three transitions, and 6% involved four or more transitions. The largest number of transitions for a single patient spell was 14. It should be noted that many of these transitions were not care related and could be attributed to billing cycles and other administrative procedural issues. In addition to coding the patterns of transition, a simple count was also included in the model.
  + Length of Spell (LOSp): This was calculated as the total time elapsed (in days) from the beginning of the patients care spell to the end. In addition, the length of stay for each episode was tracked and the ratio of LOSp to average episode duration was calculated and used as a feature. When a spell includes only one episode, this value is 1. When there were multiple episodes/claims within a single spell and that spell is long in duration, this value is greater than one. For the patient with 14 episodes within a single spell, the overall LOSp was 87 days and the average length of within spell episode was 4.3 resulting in a ratio of 20.2.

## Variable Importance

Due to its complexity, the type of algorithm used to build this predictive model is not amenable to analytical interpretation of the sort available when using less sophisticated (and significantly less accurate) algorithms such as logistic regression. It is possible to determine the relative importance of the variables included in the model. Typically referred to as *Variable Importance* or *VIMP*. Table 1 presents the relative variable importance of each of the features.

|  |  |
| --- | --- |
| **Feature** | **RVIMP** |
| LOSp | 1.000 |
| Charlson | 0.961 |
| prncplDiagCDSequence | 0.898 |
| admitDiagCDSequence | 0.805 |
| patientAge | 0.794 |
| admitTypeSequence | 0.617 |
| dschrgCDSequence | 0.566 |
| medianIncome | 0.390 |
| eduHS | 0.159 |
| patientRace | 0.031 |
| routeIRF | 0.007 |
| routeSNF | 0.004 |
| routeICF | 0.000 |

Table : Relative Variable Importance

* LOSp: Length of Spell in days
* Charlson: Charlson comorbidity score. The range for this data set was from 0 to 11.
* prncplDiagCDSequence: Concatenated sequence of primary diagnosis codes (ICD9).
* admitlDiagCDSequence: Concatenated sequence of admission diagnosis codes (ICD9).
* patientAge: Patient’s age in years at time of discharge
* admitTypeSuence: Concatenated sequence of admission type codes.
* dischrgCDSequence: Concatenated sequence of discharge codes.
* medianIncome: The median income (in dollars) reported for the census block in which the patient resides.
* eduHS: The probability that an adult living in the patient’s census block completed High School.
* patientRace: The patien’ts reported race.
* routeIRF, routeSNF, routeICF: A set of three features used to indicate if the patient’s discharge pathway included an SNF (Skilled Nursing Facility), IRF (Inpatient Rehabilitation Facility), or ICF (Intermediate Care Facility respectively).

The process for calculating the VIMP values is essentially one in which each of the features in the final model is removed, one at a time, and the model is re-run and cross-validated. The reduction in predictive power then represents the importance of each feature. When interaction terms between features are included in the model, the process becomes more complex as groups of features must be tested. In the case of *Random Survival Forest* models, interaction terms are not directly included in the model formulation, rather the impact of interactions is implicitly accounted for in the branching process used to grow the large number of trees (in this case 500).



Figure : Impact of Number of Trees on Error Rate

## Error Measurement, Model Tuning, and Performance

For statistical learning models where survival time and hazard rates are the predicted outcome or target, the underlying distributions of  and  are highly skewed and the data itself is right censored, error measures that compare predicted survival times to actual survival times are far too restrictive. The use of biased estimators compounds this issue. An alternative is to cast the task as one of ranking survival times rather than estimating those times outright. Individual pairs of patients can then be ranked as to which has a shorter estimated time to event (readmission in this case) and then testing this against the known outcomes. This approach is a form of Concordance Index (CI) [23], which itself is related to the Mann-Whitney Parameter [24], adjusted for censored data. When applied on the test data set, a pair of patients is called concordant if the risk of the event predicted by a model is lower for the patient who experiences the event at a later time point. The concordance probability (C-index) is the frequency of concordant pairs among all pairs of subjects. It can be used to measure and compare the discriminative power of a risk prediction models. In this setting, the concordance probabilities are weighted by the inverse of the probability of censoring in order to adjust for right censoring. Cross-validation based on bootstrap resampling or bootstrap subsampling can be applied to assess the discriminative power of various modelling strategies on the same set of data. While useful as an overall model performance measure, the CI is less useful for model tuning in which features are added and removed as small changes in predictive power are difficult to detect.

For model comparison purposes it is important to understand how the ranking is to be used in practice. In use, the R3M application presents an overall ranked list of recently discharged ACO patients (those within 30 days of discharge). The exact ordering of this ranking is less important than the groupings. Recast as a top *p* percent problem, with the user focusing on the patients in the top grouping, performance may best be measured by understanding what a *practical* error is. The goal is reduce the number of patients not identified within the top *p* percent that are indeed readmitted within 30 days of their discharge. Secondarily it is most efficient from the care coordinators viewpoint, to reduce the number of patients identified to be within the top *p* percent who are at lower risk of being readmitted within thirty days.

Because the level *p* determines the relative size of the identified at risk patients, setting *p* to 1, thus placing all patients above the cutoff would reduce the error rate to zero. However this would be impractical as the purpose is to provide a tool that allows the care coordinator to focus on a smaller group. By setting a (user controllable) cutoff point, the application essentially identifies two groups. For a given level *p* (sensitivity), the correct determination of membership in the two groups can be used as a way to compare individual models. This is essentially a definition of the *Net Reclassification Improvement* index often used in genomic modeling where the number of markers and complexity of models is very large [25].

### Current Model Performance

The current iteration of the model was the set of 9,461 care spells between 2013-01-01 and 2014-09-17 described earlier. The training set included 7,095 cases, with the remaining 2,366 allocated to the test set. Figure 1 presents a graph of the C-index between days 1 and 60 for the current version of the predictive algorithm. The reference model is the non-parametric Kaplan-Meier estimator [9]. The RSF model outperforms the reference model in all but a small range of days. Performance of both models is difficult to ascertain within the first five days due to the small number of cases where patients were readmitted within that time period.

Figure : C-index Statistic

# References

[1] M. T. Kassin, R. M. Owen, S. Perez, I. Leeds, J. C. Cox, K. Schnier, V. Sadiraj, and J. F. Sweeney, “Risk Factors for 30-Day Hospital Readmission among General Surgery Patients,” *J. Am. Coll. Surg.*, vol. 215, no. 3, pp. 322–330, Sep. 2012.

[2] J. Billings and T. Mijanovich, “Improving The Management Of Care For High-Cost Medicaid Patients,” *Health Aff. (Millwood)*, vol. 26, no. 6, pp. 1643–1654, Nov. 2007.

[3] B. W. Jack, V. K. Chetty, D. Anthony, J. L. Greenwald, G. M. Sanchez, A. E. Johnson, S. R. Forsythe, J. K. O’Donnell, M. K. Paasche-Orlow, C. Manasseh, S. Martin, and L. Culpepper, “A reengineered hospital discharge program to decrease rehospitalization: a randomized trial,” *Ann. Intern. Med.*, vol. 150, no. 3, pp. 178–187, Feb. 2009.

[4] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani, *Risk Prediction Models for Hospital Readmission: A Systematic Review*. Washington (DC): Department of Veterans Affairs (US), 2011.

[5] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 2013 edition. New York: Springer, 2013.

[6] Northrup Grumman Corporation Information Systems, “Accountable Care Organization – Operational System (ACO-OS) Claim and Claim Line Feed (CCLF) Information Packet (IP) – June Release,” Centers for Medicare & Medicaid Services, NGC.ICDA.0301.04.0.0713, Jul. 2013.

[7] E. T. Lee, *Statistical Methods for Survival Data Analysis*, 3 edition. New York: Wiley-Interscience, 2003.

[8] M. G. Akritas, “Nonparametric Survival Analysis,” *Stat. Sci.*, vol. 19, no. 4, pp. 615–623, Nov. 2004.

[9] E. L. Kaplan and P. Meier, “Nonparametric Estimation from Incomplete Observations,” *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, Jun. 1958.

[10] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Math. Control Signals Syst.*, vol. 2, no. 4, pp. 303–314, Dec. 1989.

[11] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[12] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random Survival Forests,” *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, Sep. 2008.

[13] U. B. Mogensen, H. Ishwaran, and T. A. Gerds, “Evaluating Random Forests for Survival Analysis Using Prediction Error Curves,” *http://www.jstatsoft.org/v50/i11/paper*, Sep. 2012.

[14] A. F. Hernandez, M.A. Greiner, G.C. Fonarow, B. G. Hammill, P. A. Heidenreich, C. W. Yancy, E. D. Peterson, and L. H. Curtis, “Relationship between early physician follow-up and 30-day readmission among medicare beneficiaries hospitalized for heart failure,” *JAMA*, vol. 303, no. 17, pp. 1716–1722, 2010.

[15] D. Peikes, A. Chen, J. Schore, and R. Brown, “Effects of care coordination on hospitalization, quality of care, and health care expenditures among medicare beneficiaries: 15 randomized trials,” *JAMA*, vol. 301, no. 6, pp. 603–618, 2009.

[16] R. P. Merkow, M. H. Ju, J. W. Chung, and et. al., “Underlying reasons associated with hospital readmission following surgery in the united states,” *JAMA*, vol. 313, no. 5, pp. 483–495, 2015.

[17] S. Kripalani, A. T. Jackson, J. L. Schnipper, and E. A. Coleman, “Promoting effective transitions of care at hospital discharge: A review of key issues for hospitalists,” *J. Hosp. Med.*, vol. 2, no. 5, pp. 314–323, Sep. 2007.

[18] A. O. Hero, J. A. Fessler, and M. Usman, “Exploring estimator bias-variance tradeoffs using the uniform CR bound,” *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2026–2041, Aug. 1996.

[19] B. Efron and R. Tibshirani, “Improvements on Cross-Validation: The .632+ Bootstrap Method,” *J. Am. Stat. Assoc.*, vol. 92, no. 438, pp. 548–560, Jun. 1997.

[20] J. B. Copas, “Regression, Prediction and Shrinkage,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 45, no. 3, pp. 311–354, Jan. 1983.

[21] *Census Block Conversions API*. Federal Communications Commision, 2009.

[22] A. C. S. Office, “American Community Survey: User Notes.” [Online]. Available: http://www.census.gov/acs/www/data\_documentation/user\_notes/. [Accessed: 15-Mar-2015].

[23] V. C. Raykar, H. Steck, B. Krishnapuram, and C. Dehing-oberije, *On Ranking in Survival Analysis: Bounds on the Concordance Index*. .

[24] J. A. Koziol and Z. Jia, “The Concordance Index C and the Mann–Whitney Parameter Pr(X>Y) with Randomly Censored Data,” *Biom. J.*, vol. 51, no. 3, pp. 467–474, Jul. 2009.

[25] S. Kundu, Y. S. Aulchenko, C. M. van Duijn, and A. C. J. W. Janssens, “PredictABEL: an R package for the assessment of risk prediction models,” *Eur. J. Epidemiol.*, vol. 26, no. 4, pp. 261–264, Apr. 2011.

1. As defined here, an episode of inpatient care can be described as a contiguous period of time when a patient was treated in a specific location for a specified condition. A spell of care is a chain of related episodes in chronological order. [↑](#footnote-ref-1)
2. Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one [↑](#footnote-ref-2)
3. A detailed derivation of the survival function estimation procedure used in RSF can be found in *Ishwaran et. al*. [12]. [↑](#footnote-ref-3)
4. is not an absolute measure of remaining risk as scaling and centering operations performed during training and scoring impact the mean and variance as well as the empirical distribution of the derived survival point values. As such the values are only useful for ranking the set of patients that were scored at the same time. In the trade-off between bias and variance, ensemble methods produce more stable predictions with less variance at the cost of an increase in bias [18] [↑](#footnote-ref-4)
5. Attributes may be transformed, centered or rescaled [↑](#footnote-ref-5)
6. Straight line distance as well as driving distance (when available) was used [↑](#footnote-ref-6)