# Intern Project Plan

June 21, 2015

## Overview

The primary project of focus will be the productization of the *Relative Readmission Risk Monitor* (R3M) tool. The goal is to have a beta test version of the application available by Q4 2015. There are several tasks that need to be undertaken:

**Model testing and optimization:** While the current incarnation of the model shows significant promise in terms of correctly ranking recently discharged ACO patients in relative risk order, before putting the tool into production additional tuning and testing is needed:

**Topic 1: Measuring Accuracy: Is there an adaptation of the CI that is more applicable to the case of assigning patients to one of two groups (high risk / low risk), that is selecting the top of patients?**While there are a number of ways to measure the accuracy of a predictive model, none of the typical approaches appear to be appropriate. Models that seek to assign a predicted category of outcome can be measured for accuracy using one of the true/false positive quantification techniques. When the predicted outcome is a hazard rate (as in the case of RM3), a measure known as the *Concordance Index*, is typically

Some references for Concordance Index:

* Koziol, James A., and Zhenyu Jia. “The Concordance Index C and the Mann–Whitney Parameter Pr(X>Y) with Randomly Censored Data.” Biometrical Journal 51, no. 3 (July 1, 2009): 467–74. doi:10.1002/bimj.200800228.
* Gerds, Thomas A., Michael W. Kattan, Martin Schumacher, and Changhong Yu. “Estimating a Time-Dependent Concordance Index for Survival Prediction Models with Covariate Dependent Censoring.” Statistics in Medicine 32, no. 13 (2013): 2173–84. doi:10.1002/sim.5681.
* Koziol, James A., and Zhenyu Jia. “The Concordance Index C and the Mann–Whitney Parameter Pr(X>Y) with Randomly Censored Data.” Biometrical Journal 51, no. 3 (July 1, 2009): 467–74. doi:10.1002/bimj.200800228.
* Raykar, Vikas C., Harald Steck, Balaji Krishnapuram, and Cary Dehing-oberije. On Ranking in Survival Analysis: Bounds on the Concordance Index, n.d.
* Wolbers, Marcel, Paul Blanche, Michael T. Koller, Jacqueline C. M. Witteman, and Thomas A. Gerds. “Concordance for Prognostic Models with Competing Risks.” Biostatistics 15, no. 3 (July 1, 2014): 526–39. doi:10.1093/biostatistics/kxt059.

**Topic 2: Model optimization: Pursuing refined feature selection and parameter optimization, how accurate can the model be?**

There are a number of areas in which feature selection could be improved:

* **Groupings of diagnoses**: Patients typically are given multiple diagnoses during their spell of care. Currently these are combined using the Charlson Index. **Are there other ways to combine multiple diagnoses?**
* **Accessibility of care:** Patients who reside in rural areas where there are fewer care providers, and the distance travelled to a PCP increase are at a significant disadvantage. **Can distances (both direct and driving) between a patient’s address and that of their PCP and the admitting hospital be used as an engineered feature that would be a surrogate for accessibility of care?**
* **Prevalence of care: Similarly, can the density of care providers in a patient’s local be used to develop a useful engineered feature?**

**System integration:** In order to be useful, the R3M application will need to be integrated with one or more of the EMR type systems used at UVA.

**Topic 3: What is the best way to integrate R3M with the current UVA systems?**

This integration will include:

* Determining how to extract from the system (Epic?) a list of patient’s discharged within the last 30 days.
* Automating the extraction of ACO claims data for these patients from our PostgreSQL DB of CMS claims records.
* Extracting from Epic, relevant information of the patient’s most recent stay at UVA (from which they were just discharged). As currently formulated, the model uses information on the current inpatient spell such as length of stay and diagnoses extracted from the CMS records. In practice, the CMS records are delayed by 45 days. As such data from previous hospital stays taken from the CMS records will need to be combined with current records from Epic.
* When scoring is completed on a set of patients, the relative risk ranking will be made available to the its primary target, the Patient Care Coordinator. This can be done through a dashboard or report in Epic, or through an auxiliary application. Pros and cons of each approach should be studied.

## Getting Started:

Before starting, you will need to get setup with the appropriate development platform. Using your own laptop will make for a much quicker process. A few things that I would recommend:

1. Setup a development environment focused on R. I use Eclipse with the StatEt plugin but you are welcome to use whatever you are comfortable with such as RStudio. The set of libraries that I currently use in building the model, and/or transforming the data include:

* DBI
* RPostgreSQL
* timereg
* xtable
* plyr
* reshape
* prodlim
* data.table
* Hmisc
* kimisc
* aod
* stringr
* hash
* bit64
* jsonlite
* pec
* survival
* rms
* randomForestSRC
* party
* ggplot2
* caret
* data.table

1. Install PostgreSQL: See <http://www.postgresql.org/> We are using verion 9.4
2. Background research:   
   There are several papers from the Journal of Statistical Software (<http://www.jstatsoft.org/>) that provide excellent information on using some of the more complex R packages including the following (by title):
   * + Building Predictive Models in R Using the caret Package
     + Evaluating Random Forests for Survival Analysis Using Prediction Error Curves
     + Analyzing Competing Risk Data Using the R timereg Package