# Predictive Modelling of High Cost Healthcare Users in Ontario

Health Analytics Branch, MOHLTC

SAS Health User Group Forum
April 12, 2013

Ontario

## Introduction

- High cost healthcare users (HCUs) are patients who incur the highest costs to healthcare
- The top 5% of users account for 68% of healthcare costs (FY 2010/11)
- Studying high cost healthcare users is important for:
  - Improving health outcomes
  - Effectively managing HCUs
  - Providing appropriate care
  - Allocating resources appropriately
  - Easing fiscal pressures on healthcare

## Introduction

- A predictive model is a statistical model that uses information on characteristics of units to predict a future outcome (for those units)
- We can use predictive modelling to predict who would become an HCU in the future, using various demographic, SES, clinical, and utilization information
- A Model that predicts who will become an HCU in the future can help:
  - Forecast expenditures and manage budgets appropriately
  - Implement proactive healthcare to prevent patients from becoming HCUs
  - Reduce resource use/impact/cost of HCUs

## Methodology

- Purpose of our predictive model:
  - To predict who will / will not become an HCU in the immediate future year, given various patient-level characteristics in the current year and two previous years
- Study period:
  - Model will estimate HCU status among patients from FY 10/11 using patient characteristics from FY 07/08 - FY 09/10
  - Model is validated by applying it to patient characteristics from FY 06/07 - FY 08/09 to predict HCU status in FY 09/10 (out of sample prediction power)
- Statistical technique:
  - Logistic regression model

# Methodology

- Population scope
  - All Ontario residents that are serviced by the health care system in Ontario during FY 09/10 in one of the following care types (database in brackets):
    - Physician services – OHIP (CHDB)
    - Acute care – AIP (DAD)
    - Day surgery – DS (NACRS)
    - Emergency – ER (NACRS)
    - Complex continuing care – CCC (CCRS)
    - Rehabilitation – Rehab (NRS)
    - Inpatient mental health – MH (OMHRS)
    - Long-term care – LTC (CCRS)
    - Home care – HC (HCD)
    - Dialysis – (NACRS)
    - Oncology – (NACRS)
    - Outpatient clinic – (NACRS)

## Methodology

- Population scope (Cont.)
  - Exclusions:
    - Patients who die during the FY 09/10
    - Patients who are under 5 years of age in FY 09/10
    - WSIB claims
    - Telemedicine claims

## Methodology

- Identify HCUs
  - High cost healthcare users: the top 5% cost incurring users in FY 10/11
  - Procedure:

1. Sum costs across all in-scope care types for each user
   a) Patient cost for AIP, ER, DS, Rehab, CCC, MH, and HC are derived from unit cost X weighted volume of services
   b) Cost for OHIP claims are represented by fees approved
   c) Patient cost for LTC are estimated using average cost per patient per day X patient length of stay
   d) Oncology, dialysis, and outpatient clinic costs were not included
2. Sort users in descending order of total expenditures, and classify the top 5% of users as HCUs
3. Create and add a binary variable to the data to identify patients as either HCU or not

# Methodology

- Data preparation
  - Identify potentially relevant variables for predicting HCUs from corpus of healthcare databases
  - Filter and extract data from various databases (e.g., remove duplicates, select only most updated record for an assessment)
  - Create rules for resolving discrepancies (e.g., conflicting postal codes, records with overlapping assessment periods)
  - Merge all data (care-specific databases, RPDB, PCCF+, etc)
  - Derive predictor variables (e.g., visit count variables, clinical group variables)
  - Transform continuous variables (e.g., on a log scale, or to a categorical variable) if necessary
  - Reduce levels for a categorical variable through clustering
  - Impute missing values using multiple imputation
  - Reduce number of variables (e.g., identify co-linearity, cluster related variables, screen out redundant and irrelevant variables)

## Methodology

- More details on imputation of missing variables
  - Imputation of missing variables was performed using PROC MI (multiple imputation) procedure in SAS
  - Monotone regression method was a method of choice
  - In the regression method, a regression model is fitted for a variable with the covariates constructed from a set of effects. Based on the fitted regression model, a new regression model is simulated from the posterior predictive distribution of the parameters and is used to impute the missing values for each variable

## Methodology

- More details on variable reduction through variable clustering
  - The PROC VARCLUS procedure in SAS divides a set of numeric variables into disjoint or hierarchical clusters
  - PROC VARCLUS was used as a variable-reduction method. A large set of variables was replaced by the set of cluster components with little loss of information.
    - A given number of cluster components does not generally explain as much variance as the same number of principal components on the full set of variables, but the cluster components are usually easier to interpret than the principal components.
  - MAXEIGEN value was chosen 0.7
    - This option specifies that when choosing a cluster to split, VARCLUS should choose the cluster with the largest second eigenvalue, provided that its second eigenvalue is greater than the MAXEIGEN= value.

## Methodology

- Creating a predictive model
  - Using data, create a statistical regression model that estimates the outcome (HCU or not) using factors (covariates) that may be influencing the outcome, such as:
    - Demographic variables (e.g., age, sex, RIO score)
    - Clinical variables (e.g., ICD-10 based chapters, with additional splits for diabetes, CHF, COPD)
    - SES variables (e.g., deprivation index (material and social deprivation))
    - Utilization variables for all care types from current year and previous two years, to account for disease progression (e.g., Number of visits, length of stay)
  - Execute model in SAS using PROC LOGISTIC

## Methodology

- Creating a predictive model (Cont.)
  - Measure the performance of the model using:
    - Model goodness-of-fit (e.g., Akaike Information Criterion (AIC))
    - Predictive ability of model for current year (e.g., c statistic)
    - Significance and impact of parameter estimates (e.g., p-value, standardized estimates)

## Methodology

- Evaluating the predictive power of the model
  - Apply model to FY 06/07–FY 08/09 data to predict HCU status of each patient in FY 09/10, given knowledge of model covariates, as if HCU status is unknown (out of sample prediction power)
  - Measure model performance using:
    - Specificity and sensitivity, positive and negative predictive value, accuracy
      - Select the top 1%, 5%, 10%, 15%, etc. of patients with the highest risk of becoming HCU
    - Receiver operating characteristic (ROC) curve
    - Calibration (goodness-of-fit) curve

## Results

- Population of patients: 10,300,856
- Number of HCUs: 520,492 (5% of population)
- Number of variables in initial model: 97
- Variables that were transformed: 64
- Number of variables reduced due to clustering: 28
- Number of variables in final model: 69

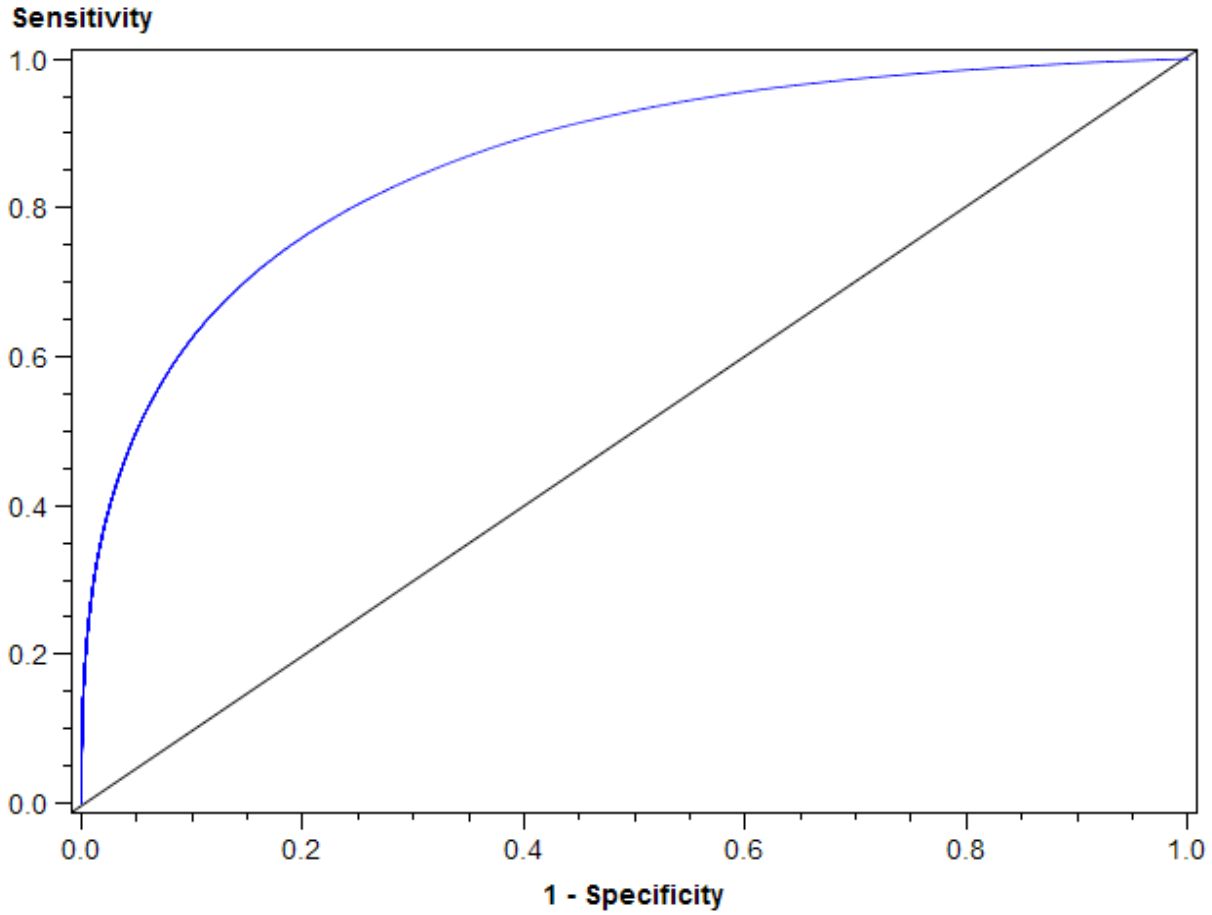
# Results

- Performance of model:
  - **C statistic: 0.865**
  - Percent Concordant: 86.1
  - Percent Discordant: 13.0
  - Percent tied: 0.9
- Predictive (out-of-sample) performance of model:

| Metric | Selection of patients based on predicted probabilities – the top: | | | | Formula | Notes |
|---|---|---|---|---|---|---|
| | 1% | 5% | 10% | 15% | | |
| Sensitivity | 15.8% | 42.2% | 57.1% | 66.4% | TP/(TP+FN) | picks up % of all high users |
| Specificity | 99.8% | 97.0% | 92.5% | 87.7% | TN/(FP+TN) | correctly identifies % of those who are not high users |
| Positive Predictive Value | 79.9% | 42.6% | 28.8% | 22.4% | TP/(TP+FP) | good at confirming high users |
| Negative Predictive Value | 95.7% | 96.9% | 97.6% | 98.0% | TN/(FN+TN) | reassuring that a patient will not become a high user |
| Accuracy | 95.5% | 94.2% | 90.7% | 86.7% | (TP+TN)/(P+N) | % of true positive and true negative out of all patients |

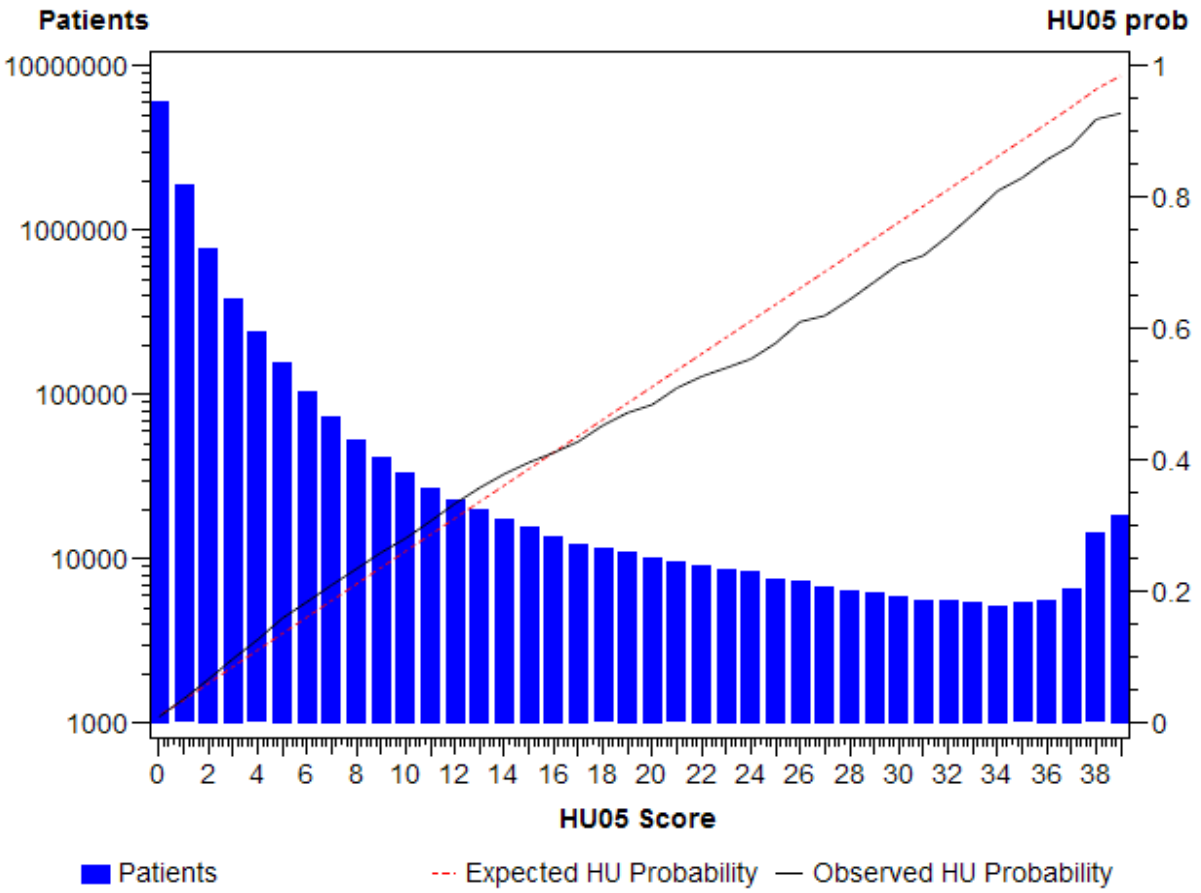
# Results

**Receiver Operating Characteristic (ROC) plot of model performance on scored 2008 data**



# Results

## Goodness of fit (calibration) curve on scored 2008 data



## Discussion

- Highlights
  - Very strong (in sample) performance (c = 0.865)
  - Graphs show very strong out-of-sample performance
  - Sensitivity/specificity analysis shows good predictive power
- Limitations
  - No population based case-mix groupers are accessible
    - Used ICD-9 and ICD-10 codes to group patients into ICD-10 chapters as a proxy, where available
  - Large number of predictor variables complicates the application of the model

# Thank You

**Project members:**
Yuriy Chechulin, Senior Methodologist, Health Analytics Branch
Amir Nazerian, Methodologist, Health Analytics Branch
Saad Rais, Methodologist, Health Analytics Branch
Kamil Malikov, Manager, Health Analytics Branch

**The Health Analytics Branch (HAB)** of the Ministry of Health and Long-Term Care, provides high quality information, analyses, and methodological support to enhance evidence-based decision making in the health system. As part of the Health System Information Management and Investment (HSIMI) Division, HAB manages health analytics requests, identifies methods, and creates reports and tools to meet ministry, LHIN and other client needs for accurate, timely, and useful information.

***Health Analytics Branch: Evidence you can count on.***

Ontario