Content Specification and

Methodology Documentation

Population Classification

Version 1: 2017

# Document History

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | Initials | Product Ver. | Doc. Draft | Changes: Section/Description of Change |
| 01/25/2018 | GS | 1.0 | 1.0 | Initial Draft |
| 02/13/2018 | GS | 1.0 | 1.1 | Change the description for Neonate logic. |
| 06/13/2018 | GS | 1.0 | 1.2 | Adding SVCSCAT ending in '24' when we define the counts for office visits. |

# Document Approval and Review

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Title | Product Ver. | Date of Approval |
| George Sirbu | Lead Data Scientist | 1.0 | 2018.01.25 |
| Christopher Gagen | Senior Methodologist | 1.0 | NA |
| Janet Young | Lead Clinical Scientist | 1.0 | NA |
| Matthew Alter | Methods Analyst | 1.0 | NA |
| Victor Yuditsky | Director, Emerging Analytics | 1.0 | NA |
| Denise Culkowski | Senior Methodologist | 1.0 | NA |

# Introduction

The Population Classification model is designed to classify patients based on interventions for patients diagnoses for use in a healthcare management system. While existing systems may classify patients based on similar clinical conditions and/or anticipated resource use, these methods are not optimized to consider the type of interventions for patients that would be most appropriate based on current needs. Individuals need different levels of care management across time, based upon various factors like characteristics of current disease condition, recent significant changes in health status, short term acute situations, lack of engagement with primary care, high patient complexity or patient utilization behaviors. The model that is presented in the document below is addressing these issues. It allows for planning and optimization of care management resources as well as providing an overall understanding of needs for a population. For example, the patients classified in “Engagement” will need to be engaged with the goal of developing a PCP relationship. On the other hand, somebody in “Coordination” will need significant care coordination from the provider. The model is more than just a statistical identification of high risk or high cost patients. The main advantage of our approach is that it provides a framework for more specific targeted analytics through actionable categories that provide insights on the entire population.

# Objective

This document specifies the content and methodology needed to create the Population Classification. The methodology is based on research done by Emerging Analytics using a sample from MarketScan data. The model is applicable only for situation where complete claims data is available, including inpatient, outpatient, drugs and enrollment. Some of the data elements are proprietary and available only in MarketScan data but not usually part of the electronic claim records, for example Stage and Diagnosis Category as identified by Disease Staging software. Whenever needed we will specify in the document the software or the methodology that was used to construct the features required by the Population Classification Model.

# Process Overview

1. Population categories

Based on the research done by clinician experts, we identified ten (10) categories for classifying patients. The categories are defined by the anticipated holistic care management resource needs of the individuals. They are

* Engagement
* Prevention
* Support
* Treatment Navigation
* Coordination
* Monitoring
* Recovery guidance
* Rebalancing
* Surveillance
* Crisis management

A description of typical patient in each of the ten categories with potential interventions, patients’ goals and examples are given below.

**Engagement -** Individuals not currently seeking care, or seeking care only via facilities such as ED, urgent care, etc.

Potential Intervention: Engage patient

Patient Goal: Primary care relationship

Example: 35 year old female without office visits, 1 visit to ER with sprained ankle

**Prevention -** Individuals who are basically healthy and whose primary focus is prevention of disease/maintaining wellness.

Potential Intervention: Wellness care tips/preventive reminders

Patient Goal: Staying well

Example: 25 year old male without significant chronic conditions with 1 office visit in past year

**Support -** Individuals with well-managed high-risk conditions whose primary focus is prevention of flare ups/disease progression.

Potential Intervention: Disease specific advice and information

Patient Goal: Living with illness

Example: 51 year old female with well controlled type 2 diabetes mellitus

**Treatment** **Navigation -** Patients newly diagnosed with conditions having multiple potential treatment pathways.

Potential Intervention: Decision guidance and support

Patient Goal: Making the best decision

Example: 45 year old female with new diagnosis of lower back pain

**Coordination -** Individuals with multiple conditions, medications, physicians, etc., which require a high level of coordination.

Potential Intervention: Significant care coordination

Patient Goal: Coordination care delivery

Example: 54 year old male with COPD, Crohn’s disease, arrhythmia, major depression on 7 maintenance drugs with visits to multiple specialists

**Monitoring -** Individuals with concerning resource use patterns such as frequent ER use or high days-supply of opioids.

Potential Intervention: Disease management

Patient Goal: Managing condition

Example: 45 year old male with knee pain and 90 day supply of opioid drugs in last 90 days

**Recovery Guidance -** People who have had a major acute event and are recovering, expected to return to their original state

Potential Intervention: Temporary assistance in recovery

Patient Goal: Getting healthy

Example: 61 year old male 18 days post anterior fusion of cervical spine

**Rebalancing -** Individuals who have had a significant change in their health status and need help finding their “new normal”.

Potential Intervention: Temporary intense care management

Patient Goal: Finding new normal

Example: 42 year old male with newly diagnosed type 2 diabetes mellitus

**Surveillance -** Patients undergoing focused intense care for a significant condition of which the care management team should be aware.

Potential Intervention: Observe, intervene rarely as needed

Patient Goal: Completing Treatment

Example: 52 year old female currently receiving radiation therapy following mastectomy for breast cancer

**Crisis Management -** Significantly ill Individuals who require a high degree of active care management on an ongoing basis.

Potential Intervention: Disaster control, end of life care

Patient Goal: Coping with Significant Disease

Example: 59 year old female with chronic pulmonary disease, CAD, diabetes, cirrhosis with recent admission for sepsis

1. Data Input

The training data used for Population Classification was a sample of 345 patients and their correspondent claims from 2013 MarketScan commercial data. Although the model is based on training from commercial MarketScan data we consider that the rules are general enough to be deployed to other type of populations, including Medicare or Medicaid. The details for the data input requirements for scoring new data are given in Section 3.6.

The patients were first selected based on a random sample and then further filtered based on expert opinion in order to balance the number of patients expected to be in the more clinically severe categories. For example, in a *typical* population (i.e. similar to MarketScan commercial population in terms of patients’ clinical characteristics) it’s expected that over 50% of the members are in Prevention but only less than 1% would be identified in Crisis Management. Our clinical experts reviewed the data and identified the category for the 345 patients. The labels were used further in the process as the ground truth for the classification algorithm. The distribution of the ten categories in the training data is given below.

Table 1 – Training data distribution by population classification categories.

|  |  |
| --- | --- |
| **Population Classification Category** | **Percent (of Total)** |
| Engagement | 5.2% |
| Prevention | 41.4% |
| Support | 20.0% |
| Treatment Navigation | 2.0% |
| Coordination | 6.7% |
| Monitoring | 7.2% |
| Recovery guidance | 1.4% |
| Rebalancing | 6.1% |
| Surveillance | 6.1% |
| Crisis Management | 3.8% |

1. Machine Learning Rules and Logic Based Rules

For developing the classification system, we used a decision tree model approach. This approach creates rules and has the advantage of making transparent the reasoning process behind the model when browsing the tree. This contrasts with other modeling techniques in which the internal logic can be difficult to work out. Another advantage of the decision tree model is that the process automatically includes in its rules only the features that really matter in classification, the rest being ignored. In our case this led to rules with a more comprehensible form. These matched well with two requirements of the project that the users can access and interpret the logic for the generated results.

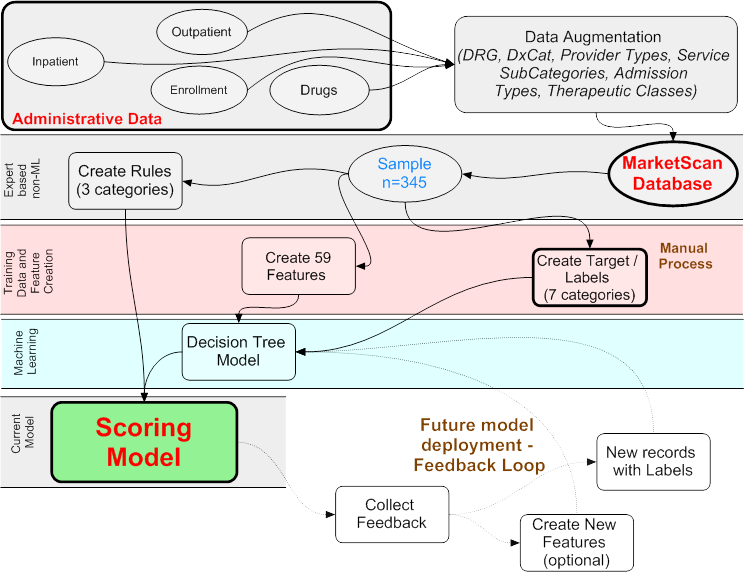
However, some of the ten categories are specific to certain diseases and/or they are based on a complex precondition exclusion logic. But the intent for the features considered in the model is to be more general and related to patient’s use of medical resources (e.g. count of inpatient admissions, count of days supply for maintenance drugs, etc.).

For these reasons, we used a hybrid approach where we identify three (3) categories based on specific (deterministic) logic while for the other seven (7) we use the above mentioned decision tree model. The three categories where we used expert defined logic are Treatment Navigation, Rebalancing and Surveillance. We will call them the logic based rules in the document below.

The details for the decision tree model are given in Section 3.5a below. The methodology for combining the two approaches is described in Section 3.5c below.

At a high level, the diagram for the project workflow is given in the diagram below.

Figure 1. Population Classification model development workflow.



1. Features Construction

For the decision tree model building we considered fifty-nine (59) features in the training data. In the end, the algorithm selected only ten (10) features that were most important for predicting the seven (7) categories: Engagement, Prevention, Support, Coordination, Monitoring, Recovery guidance and Crisis management. Recall that for other three (3) categories we are using logic based rules. The ten features in the model that are mandatory for scoring are highlighted in bold red in the tables in sections below.

In the table below we listed all features and the description for how they were constructed. In some of the descriptions we are using only the short names of the variables that are available in MarketScan database. For the full name and descriptions of these variables please see Appendix 3, Table A3.1.

* For all the constructed features in the training model we used a cutoff date as 31 December 2013 for defining 3, 6 or 12 months look back periods. For scoring, the cutoff date will be defined by the user. The longest look back period is one year.
  1. Inpatient data

For these features, we used CCAEI2013 MarketScan dataset. The record is the summary of the inpatient admission claim. It is augmented further with the diagnosis type (Acute vs. Chronic) mapping as given in Appendix 1, Table A1.1.This mapping is based on the principal diagnosis category variable, PDXCAT, from MarketScan dataset. The rest of the variables needed from CCAEI are listed in Appendix 3, Table A3.1. We did not have control of the version of Disease Staging that was used to create the PDXCAT variable in MarketScan.

* For all the look back periods we look at the difference between the cutoff data and the admission date.
* If there is no admission that meets the criteria than we set the count to zero (0) and winsorize the days since admission to 365.
* For calculating the number of days since an admission, use the cutoff date and the admission date.

Table 2 – Features based on inpatient data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Claim Type** | **Variable Name** | **Description** | **Details** | **Lookback Time Period** |
| Inpatient | CountAdm3mChronic | Number Chronic Admission, last 3 months | Based on PDXCAT, identify Chronic as given in the map in Appendix 1, Table A1.1. | 90 days |
| Inpatient | CountAdm6mChronic | Number Chronic Admission, last 6 months | Same as CountAdm3mChronic except for the look back time period. | 180 days |
| Inpatient | CountAdm3mAll | Number Any Admission, last 3 months |  | 90 days |
| Inpatient | CountAdm6mAll | Number Any Admission, last 6 months |  | 180 days |
| Inpatient | SumLOS3mAll | Total Inpatient Days in the last 3 months | Sum of all LOS for all admissions in the past 3 months | 90 days |
| Inpatient | DaysAnyAdm | Days Since Any Admission |  | 366 days |
| Inpatient | DaysChronicAdm | Days Since Chronic Admission | Based on PDXCAT, identify Chronic as given in the map in Appendix 1, Table A1.1. Count days | 366 days |
| Inpatient | DaysSurgAdm | Days Since Surgical Admission | Based on ADMTYP = 1 | 366 days |
| Inpatient | DaysMedicalAdm | Days Since Medical Admission | Based on ADMTYP = 2 | 366 days |
| Inpatient | DaysMtrntyAdm | Days since Maternity Admission | Based on ADMTYP = 3 | 366 days |
| Inpatient | DaysMHSAAdm | Days since MHSA Admission | Based on ADMTYP = 4 | 366 days |

* 1. Drug data

For these features, we used CCAED2013 MarketScan dataset. It is augmented with the flag for maintenance drugs using the map in Appendix 1, Table A1.2. It is based on NDC code variable, NDCNUM. The rest of the variables needed from CCAED are listed in Appendix 3, Table A3.1.

* For all the look back periods we look at the difference between the cutoff data and the service date on the claim.
* If there are no drugs identified by the criteria below then the count is set to zero (0).

Table 3 – Features based on drug data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Claim Type** | **Variable Name** | **Description** | **Details** | **Lookback Time Period** |
| Drug | FlagChronicRx | Flag for Maintenance Drugs at any time during the past year | Based on NDCNUM and maintenance Rx map. | 366 days |
| Drug | FlagNewChronicRx | Flag for **new** Maintenance Rx in the past 3 months | Based on NDCNUM and maintenance Rx map for identifying chronic Rx. Logic of new Rx is based on excluding cases where a drug with same THERCLS exists in the same year before 90 days period. | 90 days for new drug, 366 days for history of similar drugs |
| Drug | CountTherClass3m | Number Therapeutic Classes in the past 3 months | Count of distinct values of THERCLS. | 90 days |
| Drug | CountTherClass6m | Number Therapeutic Classes in the past 6 months | Same as CountTherClass3m except for the look back time period. | 180 days |
| Drug | CountTherClassChronic3m | Number Therapeutic Classes for Maintenance drugs in the past 3 months | Count of distinct values of THERCLS but only for Rx in the maintenance Rx map. | 90 days |
| Drug | CountTherClassChronic6m | Number Therapeutic Classes for Maintenance drugs in the past 6 months | Same as CountTherClassChronic6m except for the look back time period. | 180 days |
| Drug | CountDaysSupp3m | Number Days Supply in the past 3 months | Sum of DAYSUPP for all Rx. | 90 days |
| Drug | CountDaysSupp6m | Number Days Supply in the past 6 months | Same as CountDaysSupp3m except for the look back time period. | 180 days |
| Drug | CountDaysSuppChronic3m | Number Days Supply for Maintenance drugs in the past 3 months | Sum of DAYSUPP only for Rx in the maintenance Rx map. | 90 days |
| Drug | CountDaysSuppChronic6m | Number Days Supply for Maintenance drugs in the past 6 months | Same as CountDaysSuppChronic6m except for the look back time period.. | 180 days |
| Drug | CountDaysSuppOpiates3m | Number Days Supply for Opiates in the past 3 months | Sum of DAYSUPP only for Rx with either THERCLS or GENERID in the correspondent lists given in Appendix 1, Table A1.3. | 90 days |
| Drug | CountDaysSuppOpiates6m | Number Days Supply for Opiates in the past 6 months | Same as CountDaysSuppOpiates3m except for the look back time period. | 180 days |

* 1. Outpatient data

For these features, we used CCAEO2013 MarketScan dataset, the outpatient claims dataset. Few variables come from CCAEF2013. The dataset contains linked information to facility headers claims. The full list of variables needed from CCAEO and CCAEF are given in Appendix 3, Table A3.1. The outpatient data is augmented with the following extra information.

* DXCAT and the correspondent DxStage.

By default, the CCAEO MarketScan data has a single DXCAT variable available even when multiple DX codes are present. However, the DxStage is not available at all, so to obtain that information we need to run Disease Staging (we used version 6.33) for all patients’ outpatient claims with claim line SEQNUM as the unique record identifier. The DxStage correspondent to the DXCAT was then merged into the data that we used for feature construction. *One caveat is that the outpatient claims typically do not have a principal diagnosis and Disease Staging will possibly output multiple DXCAT variables when multiple DX codes are in the input data*. In these cases, for other data sources where we need to run Disease Staging, we recommend using the DXCAT from the first DX available on the claim. The full list of variables needed from CCAEO for running Disease Staging is given in Appendix 3, Table A3.2.

* CASEID, SVCDATE and TSVCDATE from the facility header tables CCAEF.

A non-blank value in the linked CASEID variable is used to identify the outpatient ER claims that are followed by an inpatient admission. The link is done based on combination of ENROLID and FACHDID variables. The SVCDATE and TSCVDATE variables have the date when service incurs and respectively when it ends. It is used to find the number of days in the hospital for long term care.

* The diagnosis type (Acute vs. Chronic)

It is given in the mapping from Appendix 1, Table A1.1. It is based on the single DXCAT variable available in CCAEO MarketScan datasets. As stated above, *one caveat is that for claims with multiple DX codes there is no proper way to define a principal diagnosis but instead we recommend choosing the DXCAT correspondent to the first DX available on the claim*.

* ATG

It is a procedure grouper used by OPEG model and based on CPT codes. We used the version v2015\_2 of the OPEG lookup table, it is referenced in Appendix 1, Table A1.4

* Significant conditions based on DXCAT.

There are some conditions that are identified as critical but they may not necessarily be captured in higher stage by Disease Staging. The mapping is referenced in Appendix 1, Table A1.5. As stated above, *one caveat is that for claims with multiple DX codes there is no proper way to define a principal diagnosis but instead we recommend choosing the DXCAT correspondent to the first DX available on the claim*.

As in the other cases, for the look back periods we look at the difference between the cutoff data and the service date on the claim. And when there are no claims identified by the criteria below then the count is set to zero and we winsorize the days since event to 365.

Table 4 – Features based on outpatient data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Claim Type** | **Variable Name** | **Description** | **Details** | **Lookback Time Period** |
| Outpatient | Count3mOV | Number Office Visits in the past 3 months | Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. Identify office visits by records where the value in SVCSCAT ends in ‘25’ or '24'. Count unique values of claim IDs (MSCLMID). | 90 days |
| Outpatient | Count12mOV | Office Visit in the past year | Same as Count3mOV except for the look back time period. | 366 days |
| Outpatient | Count3mOVChronic | Number Office Visits Chronic in the past 3 months | Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. Identify office visits by records where the value in SVCSCAT ends in ‘25’ or '24' and the DXCAT is Chronic as given in the map in Appendix 1, Table A1.1. Count unique values of claim IDs (MSCLMID). | 90 days |
| Outpatient | Count6mOVChronic | Number Office Visits Chronic in the past 6 months | Same as Count3mOVChronic except for the look back time period. | 180 days |
| Outpatient | Count3mOVChronicSpecialist | Number of Office Visits with a Specialist, Chronic only, in the past 3 months | Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. Identify office visits by records where the value in SVCSCAT ends in ‘25’ or '24' and the DXCAT is Chronic as given in the map in Appendix 1, Table A1.1. Select only the records were the visit was to a specialist by filtering values in STDPROV to the range 100-799 but excluding the ones corresponding to the list given in Appendix 1, Table 1.7. Count unique values of claim IDs (MSCLMID). | 90 days |
| Outpatient | Count6mOVChronicSpecialist | Number of Office Visits with a Specialist, Chronic only, in the past 6 months | Same as Count3mOVChronicSpecialist except for the look back time period. | 180 days |
| Outpatient | Count3mOVSpecialties | Number of Specialists Seen in an Office Visit, in the past 3 months | Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. Identify office visits by records where the value in SVCSCAT ends in ‘25’ or '24'. Select only the records were the visit was to a specialist by filtering values in STDPROV to the range 100-799 but excluding the ones corresponding to the list given in Appendix 1, Table 1.7. Count unique values of specialties (STDPROV). | 90 days |
| Outpatient | Count3mOVChronicSpecialties | Number of Specialists Seen in an Office Visit, Chronic only, in the past 3 months | Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. Identify office visits by records where the value in SVCSCAT ends in ‘25’ or '24' and the DXCAT is Chronic as given in the map in Appendix 1, Table A1.1. Select only the records were the visit was to a specialist by filtering values in STDPROV to the range 100-799 but excluding the ones corresponding to the list given in Appendix 1, Table 1.7. Count unique values of specialties (STDPROV). | 90 days |
| Outpatient | Count3mER | Number ER Visits in the past 3 months | Based only on ER facility claims (SVSCAT ends in ‘20’ and FACPROF is ‘F’). Also, there is no associated inpatient admission for the ER event (linked CASEID is blank). Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. Count unique values of claim IDs (MSCLMID). | 90 days |
| Outpatient | Count12mER | Number ER Visits in the past year | Same as Count3mER except for the look back time period. | 366 days |
| Outpatient | Count3mERChronic | Number ER Visits, Chronic only, in the past 3 months | Based only on ER facility claims (SVSCAT ends in ‘20’ and FACPROF is ‘F’). DXCAT is Chronic as given in the map in Appendix 1, Table A1.1. Also, there is no associated inpatient admission for the ER event (linked CASEID is blank). Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. Count unique values of claim IDs (MSCLMID). | 90 days |
| Outpatient | Count6mERChronic | Number ER Visits, Chronic only, in the past 6 months | Same as Count3mERChronic except for the look back time period. | 180 days |
| Outpatient | DaysER | Days Since Any ER in the past year | Number of days (based on SVCDATE) to the most recent ER event. The ER event is identified by facility claims with SVCSCAT ending in ‘20’ and FACPROF is ‘F’. Also, there is no associated inpatient admission for the ER event (linked CASEID is blank). Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. | 366 days |
| Outpatient | DaysMajorER | Days Since Major ER in the past year | Number of days (based on SVCDATE) to the most recent major ER event. The ER event is identified by facility claims with SVCSCAT ending in ‘20’ and FACPROF is ‘F’. Also, there is no associated inpatient admission for the ER event (linked CASEID is blank). A major event is defined by a significant DXCAT as given in the map referenced in Appendix 1, Table A1.5 or by a DxStage 3 or higher. Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. | 366 days |
| Outpatient | DaysHH | Days Since Home Health in the past year | Number of days (based on SVCDATE) to the most recent major Home Health event. as identified by SVCSCAT ending in ‘33’. Also, there is no associated inpatient admission for the Home Health event (linked CASEID is blank). Exclude records where PROC1 is in the list in Appendix 1, Table 1.6. | 366 days |
| Outpatient | DaysMajorSurg | Days Since a major outpatient procedure in the past year | Number of days (based on SVCDATE) to the most recent major procedure as identified by ATG value of “SURG MAJOR”. Also, there is no associated inpatient admission for the procedure event (linked CASEID is blank). | 366 days |
| Outpatient | DaysOxygen | Days Since Oxygen treatment in the past year | Number of days (based on SVCDATE) to the most recent Oxygen treatment as identified by PROC1 in the list in Appendix 1, Table 1.8. Also, there is no associated inpatient admission for the treatment (linked CASEID is blank). | 366 days |
| Outpatient | Days3mLTC | Days in Long Term Care (LTC) in the past 3 months | Sum of all days spent in LTC as identified by SVCSCAT value beginning with ‘102’. The length of a LTC stay is taken from SVCDATE (start) to TSVCDAT (end) from the linked facility header table (CCAEF) for each LTC identified in the outpatient data. | 90 days |
| Outpatient | Flag3mLTC | Long Term Care (LTC) flag in the past 3 months | Flag an LTC event as identified by SVCSCAT value beginning with ‘102’. | 90 days |

* 1. Outpatient and Inpatient data

The main difference for the next set of features is that they consider **all the diagnosis codes from both inpatient and outpatient claims,** CCAEO2013 and CCAES2013 and MarketScan dataset. The diagnosis codes are not used directly but rather they are ran through Disease Staging, version 6.33 to get the correspondent DXCAT and DXSTAGE. The full list of variables needed from MarketScan datasets CCAEO and CCAES to run through Disease Staging to obtain DXCAT and DXSTAGE are given in Appendix 3, Table A3.2. For all the features listed in this section, **all** the DXCAT and DXSTAGE are considered.

As in the other cases, for the look back periods we look at the difference between the cutoff data and the service date on the claim. And when there are no claims identified by the criteria below then the count is set to zero and we winsorize the days since event to 365.

Table 5 – Features based on both outpatient and inpatient data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Claim Type** | **Variable Name** | **Description** | **Details** | **Lookback Time Period** |
| Outpatient and Inpatient | Flag12mMHSA | Flag for MHSA diagnosis in the past year | Based on any DXCAT in the list in Appendix 1, Table 1.9 or ADMTYP=4 | 366 days |
| Outpatient and Inpatient | Flag3mGYN | Flag for a pregnancy diagnosis in the past 3 months | Based on any DXCAT in the list in Appendix 1, Table 1.10. | 90 days |
| Outpatient and Inpatient | Count3mAcute | Number Total Acute Conditions in the past 3 months | Exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Identify Acute conditions by any DXCAT, using the map in Appendix 1, Table A1.1. Count unique DXCAT values. | 90 days |
| Outpatient and Inpatient | Count6mAcute | Number Total Acute Conditions in the past 6 months | Same as Count3mAcute except for the look back time period. | 180 days |
| Outpatient and Inpatient | Count3mAcuteStage3 | Number Acute Conditions Stage 3 or higher in the past 3 months | Exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Identify Acute conditions by any DXCAT, using the map in Appendix 1, Table A1.1. and the corresponding DXSTAGE is 3 or higher. Count unique DXCAT values. | 90 days |
| Outpatient and Inpatient | Count3mAcuteSign | Number Significant Acute Conditions in the past 3 months | Exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Identify Acute Significant conditions by any DXCAT, using the map in Appendix 1, Table A1.1. (for Acute) and the map in Appendix 1, Table A1.5. (for Significant). Count unique DXCAT values. | 90 days |
| Outpatient and Inpatient | Count3mChronic | Number Total Chronic Conditions in the past 3 months | Exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Identify Chronic conditions by any DXCAT, using the map in Appendix 1, Table A1.1. Count unique DXCAT values. | 90 days |
| Outpatient and Inpatient | Count12mChronic | Number Total Chronic Conditions in the past year | Same as Count3mChronic except for the look back time period. | 366 days |
| Outpatient and Inpatient | Count3mChronicStage3 | Number Chronic Conditions Stage 3 or higher in the past 3 months | Exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Identify Chronic conditions by any DXCAT, using the map in Appendix 1, Table A1.1. and the corresponding DXSTAGE is 3 or higher. Count unique DXCAT values. | 90 days |
| Outpatient and Inpatient | Count6mChronicStage3 | Number Chronic Conditions Stage 3 or higher in the past 6 months | Same as Count3mChronicStage3 except for the look back time period. | 180 days |
| Outpatient and Inpatient | Count3mChronicSign | Number Significant Chronic Conditions in the past 3 months | Exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Identify Chronic Significant conditions by any DXCAT, using the map in Appendix 1, Table A1.1. (for Chronic) and the map in Appendix 1, Table A1.5. (for Significant). Count unique DXCAT values. | 90 days |
| Outpatient and Inpatient | Count6mChronicSign | Number Significant Chronic Conditions in the past 6 months | Same as Count3mChronicSign except for the look back time period. | 180 days |
| Outpatient and Inpatient | Count12mChronicSign | Number Significant Chronic Conditions in the past year | Same as Count3mChronicSign except for the look back time period. | 366 days |
| Outpatient and Inpatient | Flag3mNewChronicSign | New Significant Chronic Condition in the past 3 months | Exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Identify Chronic Significant conditions by any DXCAT, using the map in Appendix 1, Table A1.1. (for Chronic) and the map in Appendix 1, Table A1.5. (for Significant). Exclude cases where same DXCAT exists in the year before 90 days period. | 90 days for new condition, 366 days for history of same condition |
| Outpatient and Inpatient | Days3mStage3 | Number of days with Conditions Stage 3 or higher in the past 3 months. | Exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Identify cases where any DXSTAGE value is 3 or higher. Count unique SVCDATE | 90 days |
| Outpatient and Inpatient | Count3mBody | Unique Count of Body Systems past 90 days | Crosswalk all DXCAT to the body system based on the map given in Appendix 1, Table 1.12. Count unique body systems. | 90 days |
| Outpatient and Inpatient | Count3mBodySign | Unique Count of Body Systems with Significant Conditions past 90 days | Identify Significant conditions by any DXCAT, using the map in Appendix 1, Table A1.5. (for Significant). Crosswalk Significant DXCAT to the body system based on the map given in Appendix 1, Table 1.12. Count unique body systems. | 90 days |

1. Building the model
   1. Decision Tree model

As detailed in Section 3.3, we used a decision tree algorithm to build a model for classifying the patients into one of 7 categories: Engagement, Prevention, Support, Coordination, Monitoring, Recovery guidance or Crisis Management. The details about building the decision tree model are given in Appendix 2.

The Table 6 below describes the rules from the actual decision tree using the naming conventions for the features that we introduced in Section 3.4. Recall that the model retained only 10 features found to be important for the classification and not the entire list of 59. The 10 features are Count6mChronicStage3, CountDaysSuppOpiates3m, Count12mChronicSign, Count12mOV, CountDaysSupp3m, DaysAnyAdm, DaysMajorER, CountTherClassChronic6m, CountAdm3mAll and Count6mERChronic. However, in the next updates and recalibrations of the model it is possible that the list of the features or the thresholds used by the decision trees will be changed. Also notice that there are 13 mutually exclusive rules that are created by the decision tree model and the correspondent index is also given in the table.

* There is one exception to the decision tree rules below. For the recent neonates without complications we want to classify them by default in “Prevention” if the correspondent information from claims will put the child in "Engagement". If the model is assigning a category higher than "Prevention" then leave it unchanged. More exactly, if a patient has an inpatient admission in the last 30 days from the cutoff date with the DRG in the list given in the table Appendix 1, Table A1.28 and is assigned to "Engagement " than re-assign the patient to “Prevention”, otherwise leave unchanged.

Table 6 – Decision Tree Model

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Rule Index** | **Decision Tree Logic** | | | | | |  | |  | |  | | **Category** | |
|  | Count6mChronicStage3 <= 0 | | | | | | | | | |  | |  | |
|  |  | CountDaysSuppOpiates3m <= 50 | | | | | | | | |  | |  | |
|  |  |  | Count12mChronicSign <= 0 | | | | | | | |  | |  | |
|  |  |  |  | Count12mOV <= 0 | | | | | | |  | |  | |
| **1** |  |  |  |  | CountDaysSupp3m <= 37 | | | | | → | | **Engagement** | |
| **2** |  |  |  |  | CountDaysSupp3m > 37 | | | | | → | | **Prevention** | |
|  |  |  |  | Count12mOV > 0 | | | | | | |  | |  | |
| **3** |  |  |  |  | DaysAnyAdm <= 32 | | | | | → | | **Recovery Guidance** | |
| **4** |  |  |  |  | DaysAnyAdm > 32 | | | | | → | | **Prevention** | |
|  |  |  | Count12mChronicSign > 0 | | | | | | | |  | |  | |
| **5** |  |  |  | DaysAnyAdm <= 38 | | | | | | | → | | **Recovery Guidance** | |
|  |  |  |  | DaysAnyAdm > 38 | | | | | | |  | |  | |
|  |  |  |  |  | Count12mChronicSign <= 4 | | | | |  | |  | |
| **6** |  |  |  |  |  | DaysMajorER <= 255 | | | | → | | **Monitoring** | |
|  |  |  |  |  |  | DaysMajorER > 255 | | | |  | |  | |
| **7** |  |  |  |  |  |  | | CountTherClassChronic6m <= 6 | | → | | **Support** | |
| **8** |  |  |  |  |  |  | | CountTherClassChronic6m > 6 | | → | | **Coordination** | |
| **9** |  |  |  |  | Count12mChronicSign > 4 | | | | | → | | **Coordination** | |
| **10** |  | CountDaysSuppOpiates3m > 50 | | | | | | | | | → | | **Monitoring** | |
|  | Count6mChronicStage3 > 0 | | | | | | | | | |  | |  | |
| **11** |  | CountAdm3mAll <= 0 | | | | | | | | | → | | **Monitoring** | |
|  |  | CountAdm3mAll > 0 | | | | | | | | |  | |  | |
| **12** |  |  | Count6mERChronic <= 0 | | | | | | | | → | | **Crisis Management** | |
| **13** |  |  | Count6mERChronic > 0 | | | | | | | | → | | **Recovery Guidance** | |

* 1. Logic based rules

As mentioned in introduction of the overall process in Section 3.3, for three categories we use logic based rules instead of decision tree model rules. This is due to the complexity of the patients labeled in these categories that cannot be captured in the features we considered and described in Section 3.4.

* For all the logic related to DXCAT and DXSTAGE that is described below, we are using the same data as in Section 3.4d. That is, we combined outpatient and inpatient claims and ran Disease Staging version 6.33 with identifiers at the claim line level.

**Treatment Navigation**

The overall logic for this category is that the patient is newly diagnosed with a specific condition that has multiple potential treatment pathways and hence it requires professional guidance and support from medical providers. The logic is described by the steps given below.

* Use all the diagnosis codes from both inpatient and outpatient claims ran through Disease Staging as described in Section 3.4d. The full list of variables needed from MarketScan datasets CCAEO and CCAES to run through Disease Staging to obtain DXCAT and DXSTAGE are given in Appendix 3, Table A3.2.
* Identification of the patients in this category is based on DXCAT and DXSTAGE combination as given in the list in Appendix 1, Table A1.13.
* As before, exclude outpatient records where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11.
* Limit the search only for the past 30 days from the cutoff date.
* The condition must be new, exclude the patients who met the same criteria for that condition prior to 30 days and within 366 days from the cutoff date. The match is done by Condition column as given in the list from Appendix 1, Table A1.13.
* Exclude the cases where the patient has a higher stage for the same DXCAT at any time during the past year.

**Rebalance**

The overall logic for this category is that the patient has a significant change in their health status and it is not expected that they will become condition free again. Hence the patients need help from medical care management to learn how to cope with the new conditions.

Below are the details for four criteria that can qualify a patient for Rebalance category. They are not mutually exclusive and any one of them is enough to flag the patient for the category.

*Criteria 1 – If condition is present in the past 3 months, regardless of prior history*

* Use all the diagnosis codes from both inpatient and outpatient claims ran through Disease Staging as described in Section 3.4d. The full list of variables needed from MarketScan datasets CCAEO and CCAES to run through Disease Staging to obtain DXCAT and DXSTAGE are given in Appendix 3, Table A3.2.
* Filter the claims from the past 90 days. As before, exclude outpatient diagnosis where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11. Then identify the patients based on
  + - Any DXCAT from all outpatient and inpatient claims that is in the list given in Appendix 1, Table A1.14 with DXSTAGE 3 or higher

or

* + - Any DXCAT from any facility ER claim that is in the list given in Appendix 1, Table A1.15. The ER facility claims is identified by SVCSCAT ending in ‘20’ and FACPROF is ‘F’

or

* + - PROCGRP in list given in Appendix 1, Table A1.16. The PROCGRP is based on the OPEG lookup table referenced in Appendix 1, Table A1.4 by linking the procedure (PROC1) to the correspondent column from the lookup table.

or

* + - MHSA inpatient claims based on (ADMTYP=4) or DRG as listed in Appendix 1, Table A1.17. This is based only on inpatient claims (CCAEI in MarketScan)

*Criteria 2 – If new condition is present in the past 3 months with no prior history of similar condition based on similar disease category rollup or drugs*

* Use all the diagnosis codes from both inpatient and outpatient claims ran through Disease Staging as described in Section 3.4d. The full list of variables needed from MarketScan datasets CCAEO and CCAES to run through Disease Staging to obtain DXCAT and DXSTAGE are given in Appendix 3, Table A3.2.
* Filter the claims from the past 90 days. As before, exclude outpatient diagnosis where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11.
* Identify patients based on any DXCAT and DXSTAGE combination that is in the list given in Appendix 1, Table A1.18. DXSTAGE on the claim must be higher than the one listed in the table.
* Exclude the patients who had a previous history of the same rollup DXCAT, that is the first three letters of the DXCAT value. The previous history is considered past 90 days and up to 366 from the cutoff date.
* Exclude the patients who had a previous history of associated drugs as identified by THERCLS given by the list in Appendix 1, Table A1.19. The match must be for the same associated DXCAT listed in the table. The service date of the drug claim (SVCDATE) must be 90 days prior to most recent date (but within 366 days from the cutoff date) from the claim that qualified the patient as identified by the DXCAT in the step above.

*Criteria 3 – If new condition based on DRG is present in the past 3 months with no prior history of same DRG*

* Filter the inpatient claims from the past 90 days.
* Identify patients based on DRG in the list in Appendix 1, Table A1.20. Exclude the patients who had the same DRG prior to 90 days but within 366 days from the cutoff point.

*Criteria 4 – If new condition based on ICD code is present in the past 3 months with no prior history of same ICD code.*

* This is the only time where we use ICD diagnosis codes.
* Use all the diagnosis codes from both inpatient and outpatient claims, same data that is used as the input for Disease Staging as described in Section 3.4d. However, for this criterion there is no need to run the data through the software.
* Filter the claims from the past 90 days. As before, exclude outpatient diagnosis where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11.
* Identify patients based on diagnosis codes. ICD-9 codes must match only the first four characters given in the list while the ICD-10 codes must match the whole value as given in the same list in Appendix 1, Table A1.21. Exclude the patients who had the same correspondent Description of the diagnosis prior to 90 days but within 366 days from the cutoff point. The Description is given in the same table linked previously.

**Surveillance**

The overall logic for this category is to identify patients who recently (3 months) needed focused intense care for a significant condition that can be potentially life threatening. The condition is focusing mainly on identifying active cancer and advanced renal failure.

Below are the details for five criteria that can qualify a patient for Surveillance category. They are not mutually exclusive and any one of them is enough to flag the patient for the category.

*Criteria 1 – If active condition is present based on general PROCGRP and DXCAT codes*

* Use all the diagnosis codes from both inpatient and outpatient claims ran through Disease Staging as described in Section 3.4d. The full list of variables needed from MarketScan datasets CCAEO and CCAES to run through Disease Staging to obtain DXCAT and DXSTAGE are given in Appendix 3, Table A3.2.
* Filter the claims from the past 90 days. As before, exclude outpatient diagnosis where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11.
* Append the PROCGRP code based on PROC1 variable and OPEG lookup table v2015\_2 as referenced in Appendix 1, Table A1.4
* Identify the patients who have evidence of active cancer treatment based on both diagnosis (DXCAT) and procedure (PROCGRP), but they don’t necessarily need to be on the same claim:
  + - Any DXCAT in the General Active Cancer Diagnostics list as given in Appendix 1, Table A1.22 where type is “Dx”

and

* + - Any PROCGRP in the General Active Cancer Procedures list as given in Appendix 1, Table A1.23 or the DXCAT is “OTH22” (this value is also listed Appendix 1, Table A1.22 where type is “Px”)

*Criteria 2 – If active condition is present based on specific combinations of PROCGRP and DXCAT codes*

* Use all the diagnosis codes from both inpatient and outpatient claims ran through Disease Staging as described in Section 3.4d. The full list of variables needed from MarketScan datasets CCAEO and CCAES to run through Disease Staging to obtain DXCAT and DXSTAGE are given in Appendix 3, Table A3.2.
* Filter the claims from the past 90 days. As before, exclude outpatient diagnosis where the PROC1 is in the list in Appendix 1, Table 1.6 or the SVCSCAT is in the list in Appendix 1, Table 1.11.
* Append the PROCGRP code based on PROC1 variable and OPEG lookup table v2015\_2 as referenced in Appendix 1, Table A1.4
* Identify the patients who have evidence of active cancer treatment based on specific combinations of diagnosis (DXCAT) and procedures (PROCGRP). Match any of the DXCAT and PROCGRP combinations in the Specific Active Cancer list as given in Appendix 1, Table A1.24. The DXSTAGE correspondent to the DXCAT must be higher than the one listed in the table for the same DXCAT. Also, as in Criteria1, the DXCAT and PROCGRP don’t necessarily need to be on the same claim.

*Criteria 3 – If various conditions based on DRG are present in the past 3 months*

* Filter the inpatient claims from the past 90 days.
* Identify patients based on DRG in the list in Appendix 1, Table A1.25.

*Criteria 4 – If active Chemotherapy treatment is present in the past 3 months*

* Use all the drug claims from the past 90 days as described in Section 3.4b.
* Identify patients based on NDC in the list in Appendix 1, Table A1.26.

*Criteria 5 – If new chronic chemotherapy treatment is present in the past 3 months with no prior history*

* Use all the drug claims from the past 90 days as described in Section 3.4b.
* Identify patients based on NDC in the list in Appendix 1, Table A1.27.
* Exclude the patients who had prior history of any chemotherapy based on NDC and the same list given in Appendix 1, Table A1.27. The history is checked prior to 90 days but within 366 days from the cutoff point.
  1. Final category

The final category for a patient is based on merging the decision tree model from Section 3.5a and the logic based rules from Section 3.5b. The rules from the decision tree model are mutually exclusive, i.e. a patient can be in one and only one of the 7 categories. However, one patient can also be identified at the same time by none or any of the three logic based rules. The merging is done using the ranking below that will pick one and only one final classification out of the 10 categories based on the best ranking for the patient.

Table 7 – Categories Ranking

|  |  |
| --- | --- |
| Ranking | Population Classification |
| 10 | Crisis Management |
| 9 | Surveillance |
| 8 | Rebalancing |
| 7 | Recovery Guidance |
| 6 | Monitoring |
| 5 | Coordination |
| 4 | Treatment Navigation |
| 3 | Support |
| 2 | Prevention |
| 1 | Engagement |

1. Scoring new population

The logic for the decision tree and the rules are clinically valid regardless of the input data type. However, it is recommended that the scoring is done for patients with full enrolment and claims history otherwise there is possible for data gaps to produce results that are underestimating the severity of the patient’s condition. The model will run nonetheless, even when no claims at all are found for a patient and in that case the model will assign the “Engagement” category.

* The only parameter that is required from the user is the cutoff date.
* The input data required is the enrollment file containing the IDs for the patients that are required to be classified plus their correspondent medical claim files from the past year going back from the cutoff date. The medical files need to contain all the fields listed in tables in Appendix 3, Table A3.1 and Table A3.2

Also recall that for scoring new data we will only need the values from 10 features (Count12mChronicSign, Count6mChronicStage3, DaysMajorER, Count6mERChronic, Count12mOV, CountDaysSuppOpiates3m, CountDaysSupp3m, CountTherClassChronic6m, DaysAnyAdm, CountAdm3mAll) plus the flags for the three logic based rules categories.

1. Output File

In the end, the model will create an output file where the final category is given for each patient. There are more additional variables that can be made available for users in the regular output file or only when debugging option would be enabled. Also, as is the case with other Flexible Analytics products, customized patient level variables could be added as desired like zip code or patient’s gender, etc.

Table 8 – Rule Confidence

|  |  |
| --- | --- |
| Rule Index | Rule Confidence |
| 1 | 0.933 |
| 2 | 0.600 |
| 3 | 0.667 |
| 4 | 0.865 |
| 5 | 0.667 |
| 6 | 0.667 |
| 7 | 0.851 |
| 8 | 0.727 |
| 9 | 0.750 |
| 10 | 0.765 |
| 11 | 0.667 |
| 12 | 0.909 |
| 13 | 0.667 |

One of the additional variable in the output is the Rule Confidence given in Table 8. Each of the 13 rules from the decision tree model described in Table 6 has an associated confidence that describes how much was the accuracy of the rule as measured during the model training. More information is given in Appendix 2.

The full list of variables available in the output is given in Table 9 below along with an example. We highlighted in yellow the minimum information to save in the output file. In orange, we indicated the possible other data elements that we recommend for consideration in the regular output file. Finally, in green we have the 10 features needed for scoring. We recommend having all of them for debugging purposes when the option is enabled by the user.

Table 9 – Example of Population Classification Output

|  |  |
| --- | --- |
| ENROLID | Sy530jG2 |
| Final Category | Rebalance |
| Custom1 (i.e. Gender) ,… | Male |
| Decision Tree Model Category | Support |
| Decision Tree Rule Index | 7 |
| Decision Tree Rule's Confidence | 0.851 |
| Flag Treatment Navigation | 0 |
| Flag Rebalance | 1 |
| Flag Surveillance | 0 |
| Flag Logic Based Rule Overwrite | Yes |
| Count12mOV | 4 |
| Count6mERChronic | 0 |
| DaysMajorER | 365 |
| CountAdm3mAll | 0 |
| DaysAnyAdm | 365 |
| CountDaysSupp3m | 140 |
| CountDaysSuppOpiates3m | 0 |
| CountTherClassChronic6m | 4 |
| Count6mChronicStage3 | 0 |
| Count12mChronicSign | 2 |

# Model Validation and Impact Analysis

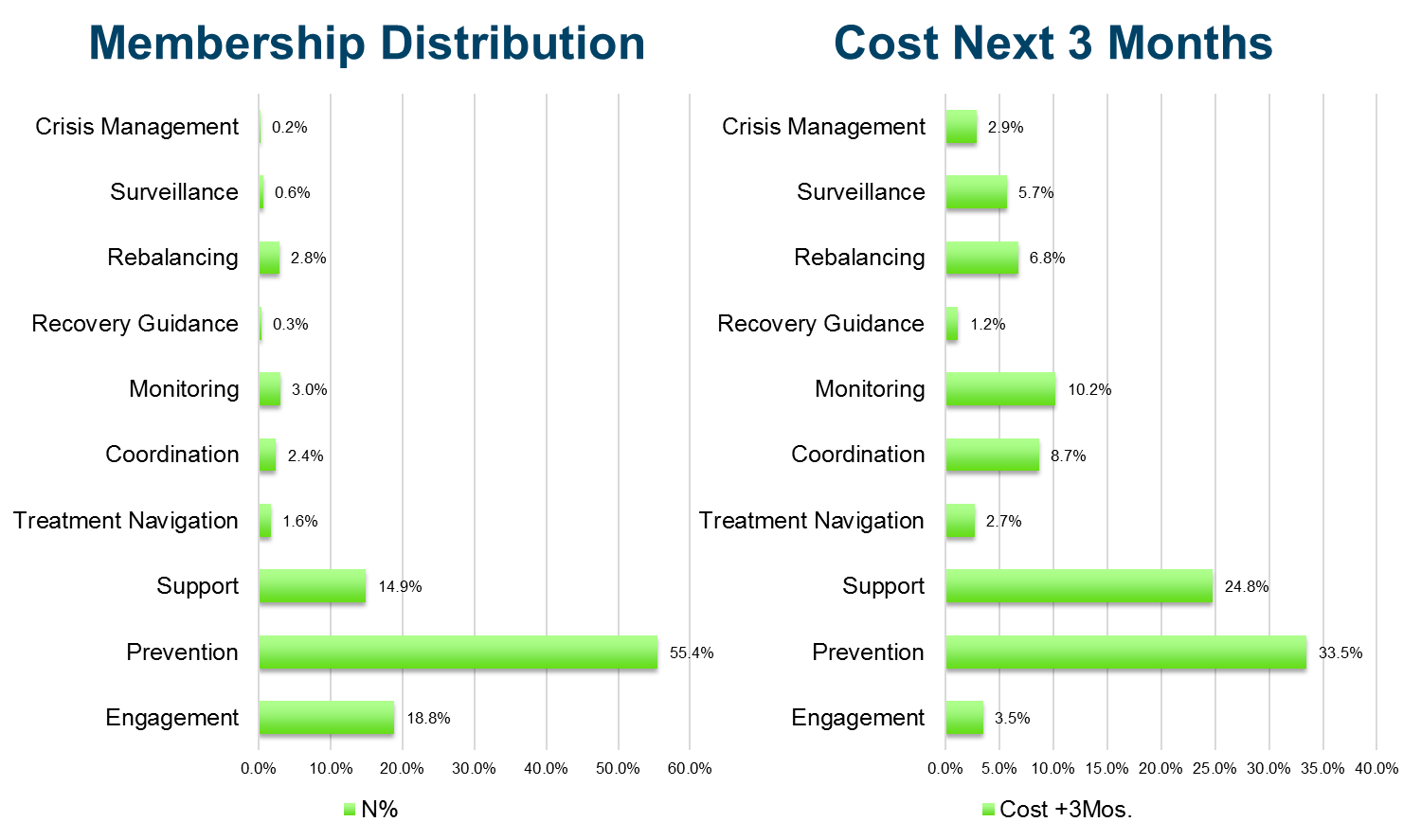
As explained in the [Section 3.2](#Section3_2), there is no objective measurement for the model actual outcome. It is based on clinical expertise and the effort to create the ground truth for a large sample of patients is prohibitive. For this reason, Population Classification model validation cannot be done at the patient level, but only at the population level.

In our analysis, we ran the Population Classification model through a 10% MarketScan sample. So, the model validation at the macro level is based on a sample representative for the MarketScan population. The results from the analysis can also be used as benchmarks for future impact analysis or comparisons with other populations from different clients.

The sample we selected had claims data from 2013 and had approximately 3.7 million members from MarketScan. The cutoff date for running the Population Classification was December 31st, 2013. We also selected the claims for the same members for 2014 and 2015 to track the healthcare expenditures in the year after the classification was done. One of the reasons we defined the 10 categories to align with the care management needs is because we expect that healthcare spending patterns are correlated with the care management needs and actions.

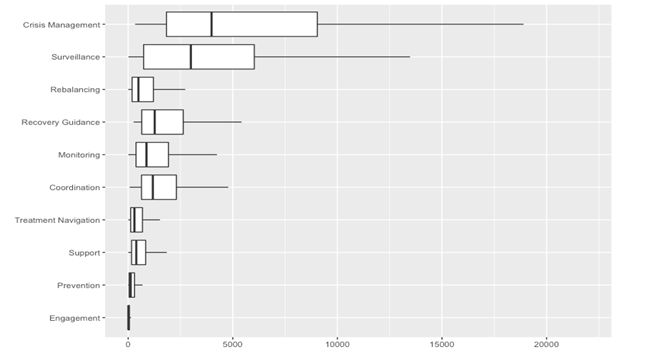
The first analysis was to check the 70-30 reverse rule for the membership vs cost distribution. More exactly we looked at the membership distribution and we observed that 74.2% of the members are classified into the lower ranked categories Prevention and Engagement. But the distribution for total cost of medical and drugs for the same sample for the next 3 months after cutoff date is reversed and only 37% of the costs are captured by the members who were classified in Prevention and Engagement. The full picture of the side by side distribution comparisons is given below.

Figure 1 – Compare Membership and Cost Distribution in the Population Classification Categories



The second analysis was to look at the distribution of the total cost (including medical and drugs) per member per month (PMPM) over a longer period - 24 months in the years 2014 and 2015 following the cutoff date. The pattern was confirmed in this case too, as you can see in the figure below. The future financials align with the expectation that higher ranked categories are also more expansive for the patients.

Figure 2 – Boxplot for Total Cost (Medical and Drugs) PMPM over 24 months period after the cutoff date



Finally, the last analysis was to confirm the different patterns in the healthcare expenditure over time. For example, we expect to see that the patients classified in Recovery Guidance have a high peek at around the cutoff date but then their healthcare spending would level back off to normal. In the figure below we can see the time series patterns for all 10 categories over a span of 2 years centered at the cutoff date of December 31st, 2013. We used month intervals and PMPM average for the total cost, including medical and drugs.

Figure 3 – Time Series Healthcare Spending by Population Classification Category

# Appendix 1 - Reference Tables

**Table A1.1 – Mapping DXCAT to Type of Diagnosis Category (Acute vs. Chronic)**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > GeneralAssignments](https://ibm.box.com/s/mj8rfw0487i39d4qk06ydnjha8f8gu1i)

**Table A1.2 – List of Maintenance Drugs**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > MaintenanceDrugs](https://ibm.box.com/s/idf9542f6q8o8z55cff3plm5mk6q4y7h)

**Table A1.3 – List of Opiate Drugs**

Source: Janet Young

THERCLS: 60 or 61

or

GENERID: 108156, 118478, 124984, 124986, 124988, 128184, 128947, 129199, 129311, 129314, 129315, 130023, 131028, 131029, 131146 or 131486

**Table A1.4 – OPEG Procedure Group Lookup (version v2015\_2)**

Source: OPEG

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > OPEG\_ProcedureGroup\_Lookup\_v2015\_2](https://ibm.box.com/s/15kt3e6so43993jgz4b9ilulg5n3pw3e)

**Table A1.5 – List of Significant Conditions**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > SignificantConditions](https://ibm.box.com/s/r5dsti1dtsa5jxnsvf2osvsrb1xgq5dx)

**Table A1.6 – List of CPT codes to use for exclusion criteria in features construction**

Source: George Sirbu

PROC1: "36415","G0001","36400","36405","36406","36410" or "36416"

Logic: CPT codes correspondent to Venipunctures.

**Table A1.7 – List of codes for specialist providers to use for exclusion criteria in feature construction.**

Source: George Sirbu

STDPROV: 200,202,204,206,240,245,320,400

Logic: Exclude the codes that are for OBGYN, Internal Medicine, Pediatrics/Geriatrics (general) or General/Family Practice and not specific to a specialty.

**Table A1.8 – List of codes for Oxygen Treatment**

Source: Janet Young

PROC1: "E0431","E0434","E0439","E0441","E0442","E0443","E0444","E1390","E1392","K0738"

Logic: CPT codes for Oxygen Treatment

**Table A1.9 – List of DXCAT Mental Health and Substance Abuse (MHSA) codes**

Source: Janet Young

DXCAT: “PSY02”,”PSY03”,”PSY80”

**Table A1.10 – List of DXCAT Maternity codes**

Source: Janet Young

DXCAT: “GYN01”,”GYN02”,”GYN03”,”GYN06”,”GYN09”,”GYN10”,”GYN12”,”GYN27”,”GYN29”,”GYN30”

**Table A1.11 – List of SVCSCAT** **to use for exclusion criteria in features construction**

Source: George Sirbu

Last 2 characters of SVCSCAT : "30", "32", "37", "38", "51", "52", "53", "54", "55", "56", "59", "61", "62", "63", "64", "65", "66", "67", "68" or "69"

Logic: We need to exclude DXCATs from outpatient claims that come from diagnosis tests (labs, radiology) as they are not reliable. The correct diagnosis is usually assigned by a physician after the tests are completed.

**Table A1.12 – Mapping DXCAT to the Body Systems**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > BodySystem](https://ibm.box.com/s/n0r6e5dwmkwzo4blf6ul6jksd2tkksb2)

**Table A1.13 – List of DXCAT and DXSTAGE specific to Treatment Navigation Category**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > TreatmentNavigation](https://ibm.box.com/s/0x4scx5rtbso6ipe8rs2kv1jvvaatd2t)

**Table A1.14 – List of DXCAT specific to Rebalancing Category, no exclusion for previous history**

Source: Janet Young

DXCAT: “CVS10”, “CVS11”

**Table A1.15 – List of DXCAT specific to Rebalancing Category based on ER claims**

Source: Janet Young

DXCAT: “PSY01”-“PSY99”

**Table A1.16 – List of Procedure Groups (PROCGRP) specific to Rebalancing Category**

Source: Janet Young

PROCGRP: '0090', '0095', '2370', '1045', '4580', '4585'

**Table A1.17 – List of DRGs specific to Rebalancing Category**

Source: Janet Young

DRG: '280', '281', '282', ‘283', '284', '285', '239', '241', '474', '475', '476', '616', '617', '618', '619', '620', '621', '003', '004', '011', '012', '013' or '240';

**Table A1.18 – List of DXCAT specific to Rebalancing Category (need check against prior history)**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > RebalanceNewDxCat](https://ibm.box.com/s/kkftzu45yhve49r1orx5hfqbi42b9hrs)

**Table A1.19 – List of Drugs (by THERCLS) matching conditions in Rebalancing Category for exclusion based on past history**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > RebalanceNewDxCatwoRx](https://ibm.box.com/s/w38ewyoh67x3e7b6ynmt7jda6hcdiafv)

**Table A1.20 – List of DRG specific to Rebalancing Category (need check against prior history)**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > RebalanceNewDRG](https://ibm.box.com/s/j5v96rom5fdrmhhma3p44tsxhv60foho)

**Table A1.21 – List of Dx (ICD) codes specific to Rebalancing Category (need check against prior history)**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > RebalanceNewDx](https://ibm.box.com/s/ovys63jplhvag15vwdh5jfsnit63trff)

**Table A1.22 – List of General Active Cancer DXCAT specific to Surveillance**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > SurveillanceGeneralActiveCancerDxCat](https://ibm.box.com/s/9gqvrlt6fr74v6j3fepav6rwwyetmtga)

Logic: List of active cancer DXCAT

**Table A1.23 – List of General Active Cancer PROCGRP specific to Surveillance**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > SurveillanceGeneralActiveCancerDxCat](https://ibm.box.com/s/9gqvrlt6fr74v6j3fepav6rwwyetmtga)

Logic: List of active cancer Procedure Groups

**Table A1.24 – List of Specific Active Cancer combinations of DXCAT/PROCGRP specific to Surveillance**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > SurveillanceSpecificActiveCancer](https://ibm.box.com/s/e2g4urqiwizdolmynt7grj2gc6s70xq0)

Logic: List of active cancer combinations of Disease Categories and Procedure Groups

**Table A1.25 – List of Specific Active Cancer combinations of DXCAT/PROCGRP specific to Surveillance**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > SurveillanceMiscellaneous](https://ibm.box.com/s/u3zpwn2i72xvsweao4ioikkc3i8wvc1z)

Logic: List of miscellaneous DRG that are indication of significant conditions requiring intensive care.

**Table A1.26 – List of active ChemoTherapy drugs (NDC)**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > SurveillanceChemoActive](https://ibm.box.com/s/zsspixj0ruiz5tyqbhxjmexe0lhubfac)

Logic: List of active Chemotherapy drugs

**Table A1.27 – List of chronic ChemoTherapy drugs (NDC)**

Source: Janet Young

[IBM.box.com > Population Classification > SAS code > Reference data from SAS code > SurveillanceChemoChronic](https://ibm.box.com/s/tljnuyz8vk6ovt0u7gmq9yqrwazex5kd)

Logic: List of chronic Chemotherapy drugs

**Table A1.28 – List of Neonate DRGs**

Source: Janet Young

DRG: “790”, ”791”, “792”, “793”, “794”, “795”

# Appendix 2 - Model Building Methodology Details

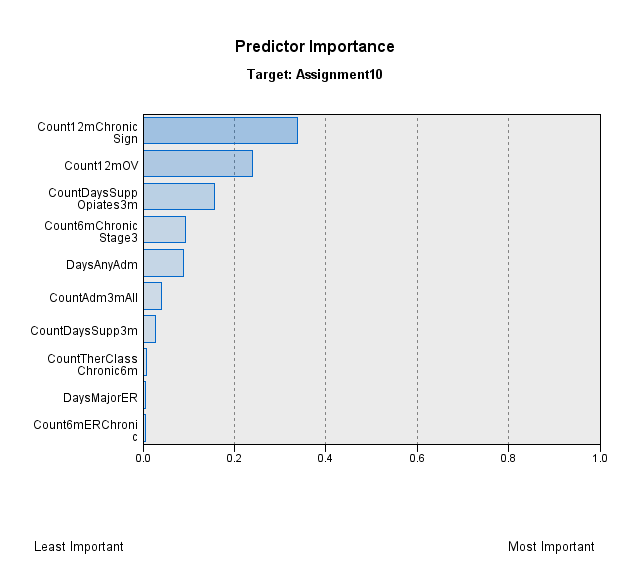
The decision tree model was built in SPSS Modeler version 18.0 using a C5 algorithm. The details for the model build settings are given below. Although this is the final version of the model there were many other variations with some changed parameters in order to find the ideal compromise between the accuracy and the interpretability of the model. We indicate in the details column when we varied the parameters in the different iterations of the model.

|  |  |
| --- | --- |
| **Build Settings** | **Details** |
| Use partitioned data: false | Given the very small sample available for training, we used cross validation but no partitions. |
| Calculate predictor importance: true |  |
| Calculate raw propensity scores: false |  |
| Calculate adjusted propensity scores: false |  |
| Use weight: false |  |
| Output type: Decision tree |  |
| Group symbolics: false |  |
| Use boosting: false | In some iterations we also used boosting. |
| Cross-validate: true |  |
| Number of folds: 4 |  |
| Mode: Expert |  |
| Pruning severity: 80 | In some iterations we ranged the pruning severity from 70 to 90 by increaments of 5. |
| Minimum records per child branch: 3 | In some iterations we ranged the minimum records from 2 to 5 in increments of 1. |
| Winnow attributes: false |  |
| Use global pruning: true |  |
| Use misclassification costs: false | In some iterations we used weighting to give more importance to the misclassification of higher ranked categories like “Crisis Management” into lower ranked categories like “Engagement”. |
| **Training Summary** |  |
| Algorithm: C5 |  |
| Model type: Classification |  |
| Application: IBM® SPSS® Modeler 18 |  |
| **Analysis** |  |
| Tree depth: 7 |  |
| Cross Validation Mean: 72.8 |  |
| Cross Validation Standard Error: 2.3 |  |

In Section 3.3, Table 6 we present the decision tree in the summarized form. In the table below we present the thirteen (13) rules for each classification. For each rule is listed the number of observations in the final node and the confidence in the rule as given by the ratio of the number of patients in the correspondent category out of total number of the patients in the final node. The Rule Index is the same as the one given in [Table 6](#Table6) and the Rule Confidence is the same as the one given in [Table 8](#Table8).

|  |  |  |
| --- | --- | --- |
| **Rules for Engagement - contains 1 rule(s)** | | |
|  | Rule 1 for Engagement (15; 0.933) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign <= 0 |
|  |  | and Count12mOV <= 0 |
|  |  | and CountDaysSupp3m <= 37 |
|  |  | then Engagement |
| **Rules for Prevention - contains 2 rule(s)** | | |
|  | Rule 2 for Prevention (10; 0.6) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign <= 0 |
|  |  | and Count12mOV <= 0 |
|  |  | and CountDaysSupp3m > 37 |
|  |  | then Prevention |
|  | Rule 4 for Prevention (155; 0.865) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign <= 0 |
|  |  | and Count12mOV > 0 |
|  |  | and DaysAnyAdm > 32 |
|  |  | then Prevention |
| **Rules for Support - contains 1 rule(s)** | | |
|  | Rule 7 for Support (67; 0.851) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign > 0 |
|  |  | and DaysAnyAdm > 38 |
|  |  | and Count12mChronicSign <= 4 |
|  |  | and DaysMajorER > 255 |
|  |  | and CountTherClassChronic6m <= 6 |
|  |  | then Support |
| **Rules for Coordination - contains 2 rule(s)** | | |
|  | Rule 8 for Coordination (11; 0.727) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign > 0 |
|  |  | and DaysAnyAdm > 38 |
|  |  | and Count12mChronicSign <= 4 |
|  |  | and DaysMajorER > 255 |
|  |  | and CountTherClassChronic6m > 6 |
|  |  | then Coordination |
|  | Rule 9 for Coordination (4; 0.75) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign > 0 |
|  |  | and DaysAnyAdm > 38 |
|  |  | and Count12mChronicSign > 4 |
|  |  | then Coordination |
| **Rules for Monitoring - contains 3 rule(s)** | | |
|  | Rule 6 for Monitoring (3; 0.667) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign > 0 |
|  |  | and DaysAnyAdm > 38 |
|  |  | and Count12mChronicSign <= 4 |
|  |  | and DaysMajorER <= 255 |
|  |  | then Monitoring |
|  | Rule 10 for Monitoring (17; 0.765) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m > 50 |
|  |  | then Monitoring |
|  | Rule 11 for Monitoring (3; 0.667) | |
|  |  | if Count6mChronicStage3 > 0 |
|  |  | and CountAdm3mAll <= 0 |
|  |  | then Monitoring |
| **Rules for Recovery guidance - contains 3 rule(s)** | | |
|  | Rule 3 for Recovery guidance (3; 0.667) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign <= 0 |
|  |  | and Count12mOV > 0 |
|  |  | and DaysAnyAdm <= 32 |
|  |  | then Recovery guidance |
|  | Rule 5 for Recovery guidance (3; 0.667) | |
|  |  | if Count6mChronicStage3 <= 0 |
|  |  | and CountDaysSuppOpiates3m <= 50 |
|  |  | and Count12mChronicSign > 0 |
|  |  | and DaysAnyAdm <= 38 |
|  |  | then Recovery guidance |
|  | Rule 13 for Recovery guidance (3; 0.667) | |
|  |  | if Count6mChronicStage3 > 0 |
|  |  | and CountAdm3mAll > 0 |
|  |  | and Count6mERChronic > 0 |
|  |  | then Recovery guidance |
| **Rules for Crisis Management - contains 1 rule(s)** | | |
|  | Rule 12 for Crisis Management (11; 0.909) | |
|  |  | if Count6mChronicStage3 > 0 |
|  |  | and CountAdm3mAll > 0 |
|  |  | and Count6mERChronic <= 0 |
|  |  | then Crisis Management |

The figure below represents the features importance in the decision tree model. For example, the most important feature for discriminating the seven categories is the count of Significant Chronic Conditions in the past year (Count12mChronicSign).



Finally, the accuracy and the full confusion matrix from training the model is given in the table below. Recall that there were 345 members overall selected in the sample out of whom 305 were in the seven categories that were used for the classification with the decision tree model.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Overall Accuracy** | |  |  |  |  |  |  |  |
|  | Correct | 255 | 83.61% |  |  |  |  |  |
|  | Wrong | 50 | 16.39% |  |  |  |  |  |
|  | Total | 305 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **Confusion Matrix** | |  |  |  |  |  |  |  |
|  |  | **Prediction** | | | | | | |
|  |  | Coordination | Crisis Management | Engagement | Monitoring | Prevention | Recovery guidance | Support |
|  | Coordination | 11 | 0 | 0 | 4 | 4 | 0 | 4 |
|  | Crisis Management | 1 | 10 | 0 | 1 | 0 | 0 | 0 |
|  | Engagement | 0 | 0 | 14 | 0 | 4 | 0 | 0 |
| **Actual** | Monitoring | 1 | 0 | 0 | 17 | 3 | 3 | 1 |
|  | Prevention | 1 | 0 | 1 | 0 | 140 | 0 | 5 |
|  | Recovery guidance | 0 | 1 | 0 | 1 | 0 | 6 | 0 |
|  | Support | 1 | 0 | 0 | 0 | 14 | 0 | 57 |

# Appendix 3 - MarketScan Variables Descriptions

**Table A3.1 – Description for variables needed from MarketScan datasets for features creation**

|  |  |  |
| --- | --- | --- |
| Variable Short Name | Dataset | Label/Description |
| ENROLID | CCAEI | Unique member identifier |
| PDXCAT | CCAEI | Principal Diagnosis Category. It is based on Disease Staging |
| ADMTYP | CCAEI | Admission Type. It can be surgical, medical, maternity or mental health |
| ADMDATE | CCAEI | Date of Admission |
| DISDATE | CCAEI | Date of Discharge |
| ENROLID | CCAED | Unique member identifier |
| NDCNUM | CCAED | National Drug Code number |
| SVCDATE | CCAED | Date Service Incurred |
| DAYSUPP | CCAED | Days Supply |
| GENERID | CCAED | Generic Product ID |
| THERCLS | CCAED | Therapeutic Class |
| ENROLID | CCAEO | Unique member identifier |
| SEQNUM | CCAEO | Sequence Number (unique key for all records) |
| DXCAT | CCAEO | Disease Staging Diagnosis Category |
| MSCLMID | CCAEO | MarketScan Claim ID. It may not be unique key as some claims have multiple lines (records). |
| FACPROF | CCAEO | Facility/Professional Claim Indicator |
| PROC1 | CCAEO | Procedure Code |
| PROCGRP | CCAEO | Procedure Code Group |
| SVCSCAT | CCAEO | Medstat Service Sub-Category Code |
| STDPROV | CCAEO | Standardized Provider Type |
| FACHDID | CCAEO | Facility Header Record ID |
| SVCDATE | CCAEO | Date Service Incurred |
| ENROLID | CCAEF | Unique member identifier |
| FACHDID | CCAEF | Facility Header Record ID |
| CASEID | CCAEF | Case ID for inpatient admissions |
| SVCDATE | CCAEF | Date Service Incurred |
| TSVCDAT | CCAEF | Date Service Ending |

**Table A3.2 – Description for variables needed from MarketScan datasets for running Disease Staging**

|  |  |  |
| --- | --- | --- |
| Variable Short Name | Dataset | Label/Description |
| ENROLID | CCAEO/CCAES | Unique member identifier |
| FACPROF | CCAEO/CCAES | Facility/Professional Claim Indicator |
| SEQNUM | CCAEO/CCAES | Sequence Number (unique key for all records) |
| AGE | CCAEO/CCAES |  |
| SEX | CCAEO/CCAES |  |
| PROC1 | CCAEO/CCAES | Procedure Code |
| DXVER | CCAEO/CCAES | Diagnosis Code (ICD) Version Indicator |
| DX1-DX4 | CCAEO/CCAES | Diagnosis Code (ICD) filed 1 through 4. |
| STDPLAC | CCAEO/CCAES | Place of Service |