SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

Assessing Completeness of Clinical Data through Claims Data Matching

using SAS® Enterprise Miner™

USERS PROGRAM

Assessing Completeness of Clinical Data through Claims Data Matching

Catherine Olson, Thomas Horstman

Optum

ABSTRACT

Research using electronic health records (EHR) is emerging, but questions remain about its completeness, due in part to physicians' lack of time to enter data in all fields. This project used SAS® Enterprise Miner™ to predict completeness of a sampling of clinical data, using claims data as the standard "source of truth" against which to compare it. A method for assessing and predicting the completeness of clinical data is presented, along with tips for preparing sample data for use in SAS Enterprise Miner; for preparing sample data set for modeling, including effective use of the Input Data, Data Partition, Filter, and Replacement nodes; and for building predictive models using Stat Explore and Decision Tree.

DATA

This analysis used data from a company research database that includes de-identified administrative claims data and de-identified clinical EHR records for 37 anonymous provider groups. No identifiable protected health information was extracted or accessed. The resulting sample consisted of 1,248,102 members who met the study criteria:

- Commercial and Medicare members had enrollment periods in health plans with pharmacy and medical benefits between 1/1/2007 and 12/31/2014.
- Members had at least one paid pharmacy claim with a valid NDC code, or one paid medical claim with a valid diagnosis or procedure code.
- Excluded members with missing values for age, gender or geographic information in claims data.
- Health plan members had both administrative claims data and clinical EHR data.
- Time span of a patient represented in the clinical data, defined by dates of clinical activity, overlapped with at least one of the enrollment periods.
- Selected patients' most recent enrollment periods with at least 12 months of continuous health plan enrollment from the date of the first observed paid claim (medical or pharmacy) with valid coding.

METHODS

Seven completeness measures based on diagnosis or procedure codes were calculated at the member and study population level to analyze the completeness of clinical data: AHRQ, 3 Digit & 5 Digit ICD9, CPT4/HCPCS and ICD9 Procedure Codes. Measures were calculated as a percent of unique codes or coding categories present in members' claims that were also present in their clinical data. The denominator was the number of unique categories per member in the claims and the numerator was the number of unique categories in the clinical data which matched the claims.

METHODS CONTINUED

Members were also assigned to demographic categories based on their clinical records:

- **Age groups.** <=17, 18-24, 25-44, 45-64, 65+, Unknown.
- **Race/Ethnicity.** White or Caucasian, Black or African American, Hispanic, Asian, Unknown or Other.
- **Gender.** Male, Female, Unknown.
- **Insurance Type.** Commercial, Medicare, Unknown or Other.

In addition, a member's provider and other clinical record indicators (such as whether a member had immunization records or notes) were identified from the clinical data and assigned at the member level. Lastly, a "match level" was assigned based on demographic information common to both claims and clinical data to identify patients and link their data. The match level definitions used varying combinations of names, birthdates, social security numbers, residential states and zip codes.

A final member level analytic dataset was prepared in BASE SAS which included a member's 7 completeness measures, demographics, provider, match level and other clinical indicators. This final analytic dataset was loaded into SAS® Enterprise Miner™.

The following tools were used in loading and preparing the data in SAS® Enterprise Miner™ for modeling:

- Input Data node was used to identify each target (the completeness measure) and input variables (such as a member demographics), modify variable levels (binary, ordinal, nominal, Interval, unary), and to drop variables not needed for the model.
- Filter node was used to remove specific values from the sample. Two providers were removed after subject matter experts determined they did not have sufficient data for the analysis.
- Replacement node was used to stratify variables for insurance and race.
- Stat explore node was used to validate that the data set identified from the above steps was accurate, and to find the mean completeness measures for the overall population.
- Data partition node was used to partition the data into training, validation and test sets for the models.

Multivariate analysis was used to identify subpopulations whose demographic and other characteristics predict higher completeness scores than the overall study population. Initial decision trees in SAS® Enterprise Miner™ were used to identify a subset of clinical attributes that had the highest importance in the decision analysis.

Assessing Completeness of Clinical Data through Claims Data Matching

Catherine Olson, Thomas Horstman

Optum

METHODS CONTINUED

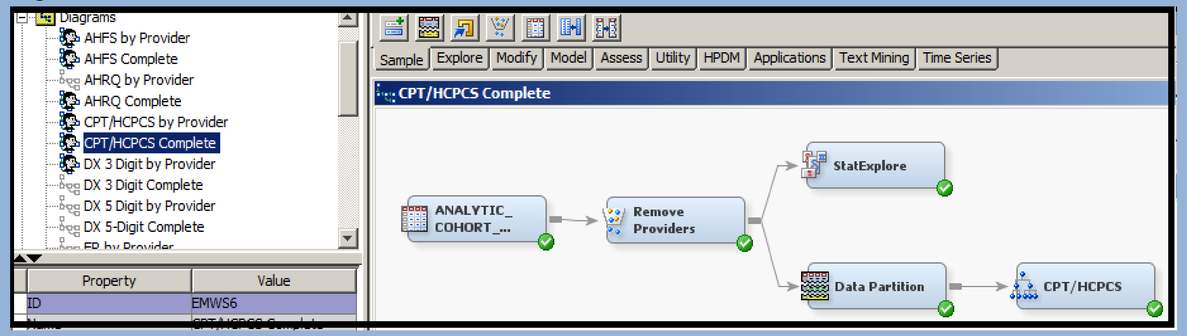
These initial trees used all member level clinical attributes, demographics, match levels and providers as the input variables, and a separate tree was generated for each completeness measure as the target variable. The subset of clinical attributes identified by these initial trees was further stratified using bivariate analysis, initial tree results and domain knowledge to understand the appropriate groupings, and to remove input variables from the model. Final simplified models were then run with the subset of variables.

RESULTS

Iteration 1 – Initial Trees

Initial trees were generated for each of the 7 completeness measures. Figure 1 shows the overall steps for generating the initial CPT/HCPCS decision tree model.

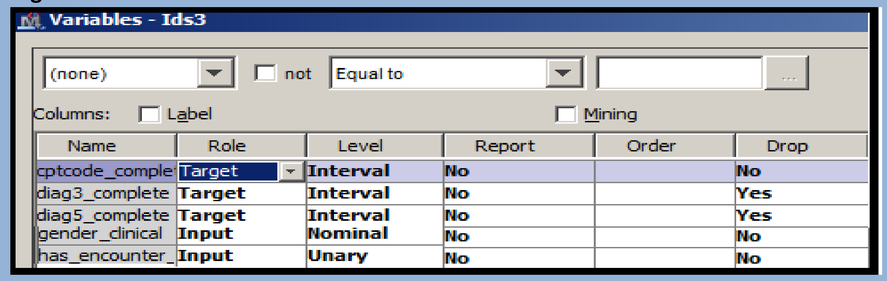
Figure 1



Input node: Figure 2 shows the example of how CPT/HCPCS was identified as the target variable, and other completeness measures were dropped from the model. In addition, the input node was used to verify all other variables' attributes were represented appropriately for the decision tree model. For example, CPT/HPCS was identified as an interval value, and gender was not dropped and was specified as Nominal.

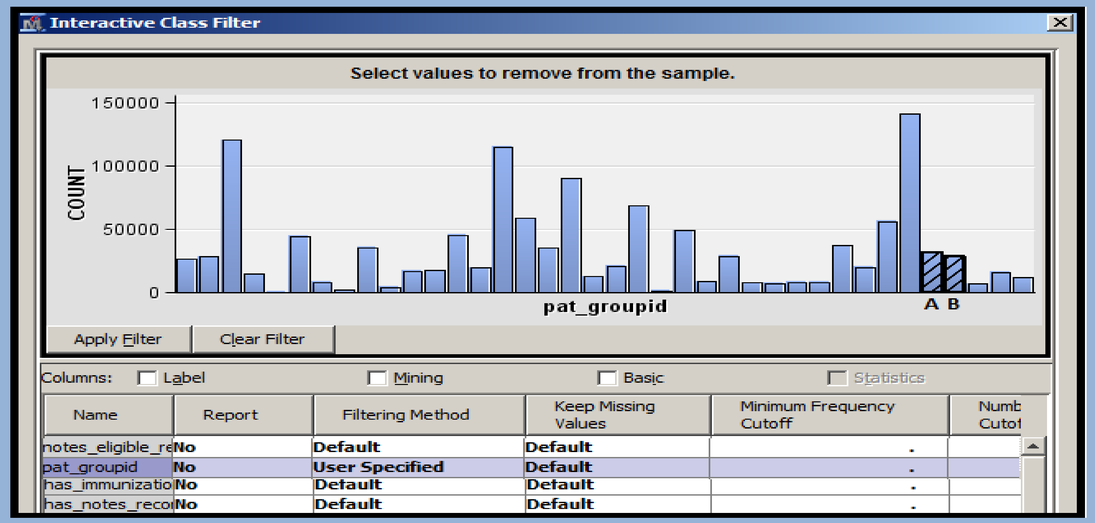
RESULTS CONTINUED

Figure 2



Filter Node: A partial interactive class filter is shown in Figure 3. Two providers were removed after subject matter experts determined they did not have sufficient source data for the analysis. The highlighted bars represent providers A & B, which were selected to be removed from the model (pat_groupid was the variable name for a member's providers).

Figure 3



Assessing Completeness of Clinical Data through Claims Data Matching

Catherine Olson, Thomas Horstman

Optum

RESULTS CONTINUED

Stat Explore: A partial results window is shown in Figure 4. This analysis determined the overall mean completeness score was 20% for the entire population ***(NOTE: 20% is an example result only).***

Figure 4

Output

```
40    (maximum 500 observations printed)
41
42    Data Role=TRAIN
43
44                                                  Standard
45        Variable          Role          Mean     Deviation
46
47    cptcode_complete     TARGET     0.200058     0.258059
```

Data Partition Node: Partitioned the data into training (40%), validation (30%) and test (30%).

Decision Tree Node: All 7 initial completeness decision trees identified provider group as the attribute with the highest importance. Therefore, provider group was the first split rule in every initial decision tree. (Provider groups were anonymized for the study.) In addition, all initial decision trees identified insurance, match level, race/ethnicity, age category and gender as the next highest split rules. Figure 5 is an example output for CPT completeness measure.

Figure 5

Output

	Variable Name	Label	Splitting Rules	Importance
50	pat_groupid		1	1.0000
52	race_clinical		1	0.5169
53	match_level_clinical		5	0.3758
54	agecat_clinical		4	0.3688
55	insurance_type_clinical		5	0.3461
56	gender_clinical		4	0.2150

RESULTS CONTINUED

Iteration 2 – Simplified Trees

Because all initial decision trees identified provider, insurance, match level, race/ethnicity, age category and gender as the highest split rules, they were identified as potential attributes for the final models. However, because there were no meaningful differences in the bivariate analysis results for gender and age group, these attributes were excluded.

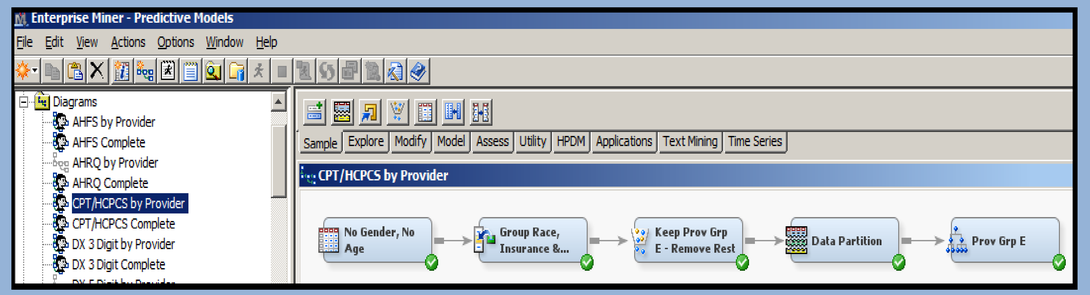
Insurance, race/ethnicity, and match level were further stratified as follows:

- Insurance. Medicare and Commercial were grouped into one variable (INS_YES), and "other" (i.e. not Medicare or Commercial) was grouped into another variable (INS_OTH). This corresponds to bivariate analysis showing that insurance types of Medicare and Commercial have higher completeness rates than insurance type of "other".
- Race/ethnicity. Asian, Black, Hispanic and White were grouped into one variable (RACE_KNOWN), and unknown race/ethnicity was grouped into another variable (RACE_UNKNOWN). This corresponds to bivariate analysis showing that race values that are known have higher completeness than when race/ethnicity is unknown.
- Match Levels. Match levels were grouped with other match levels of similar attributes.

Because provider group had the highest importance in every completeness measure, and there was a significant variation in each of the provider groups with no common patterns to pool the data, a decision tree was created for each provider group and completeness measure.

Figure 6 is an Example for 1 provider group (E) and 1 completeness measure (CPT/HCPCS) for Iteration 2.

Figure 6



Assessing Completeness of Clinical Data through Claims Data Matching

Catherine Olson, Thomas Horstman

Optum

RESULTS CONTINUED

Input Node: Removed age category and gender from the model using the input node where Drop = Yes (Figure 7).

Figure 7

Variables - Ids2

(none) not Equal to

Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
agecat_clinical	Input	Nominal	No		Yes	.	.
cptcode_comple	Target	Interval	No		No	.	.
gender_clinical	Input	Nominal	No		Yes	.	.

Replacement Node: Grouped race, insurance and match levels into new replacement values as described above. Replacement values for insurance and race are shown below (Figure 8).

Figure 8

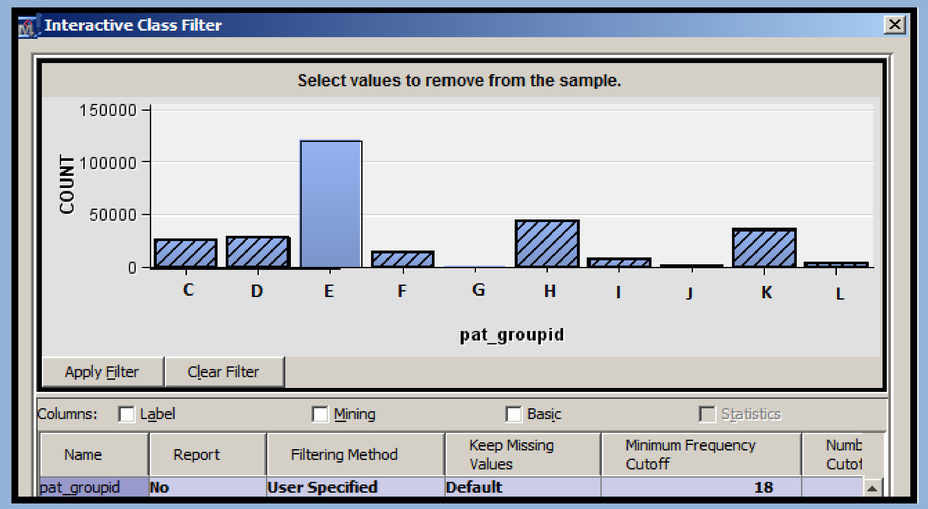
Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value
insurance_type_clinical	COM	INS_YES	623568	C	COM
insurance_type_clinical	OTH	INS_OTH	446013	C	OTH
insurance_type_clinical	MCR	INS_YES	178521	C	MCR
race_clinical	White or caucasian	RACE_YES	859098	C	White or caucasian
race_clinical	Unknown race	RACE_UNK	238071	C	Unknown race
race_clinical	Black or african american	RACE_YES	75592	C	Black or african american
race_clinical	Hispanic	RACE_YES	45657	C	Hispanic
race_clinical	Asian	RACE_YES	29684	C	Asian

RESULTS CONTINUED

Filter Node: Filtered out all provider groups except provider E (Figure 9).

Figure 9



Data Partition Node: Partitioned the data into training (40%), validation (30%) and test (30%).

Decision Tree Node: Figure 10 is an example of decision trees for one provider (E) and one completeness measure (CPT/HCPCS).

Provider Group E has a subpopulation that predicts highest CPT/HCPCS completeness when insurance and race are defined, and when match level l_4 is used as the criteria to link claims to clinical data based on a member's demographic data.

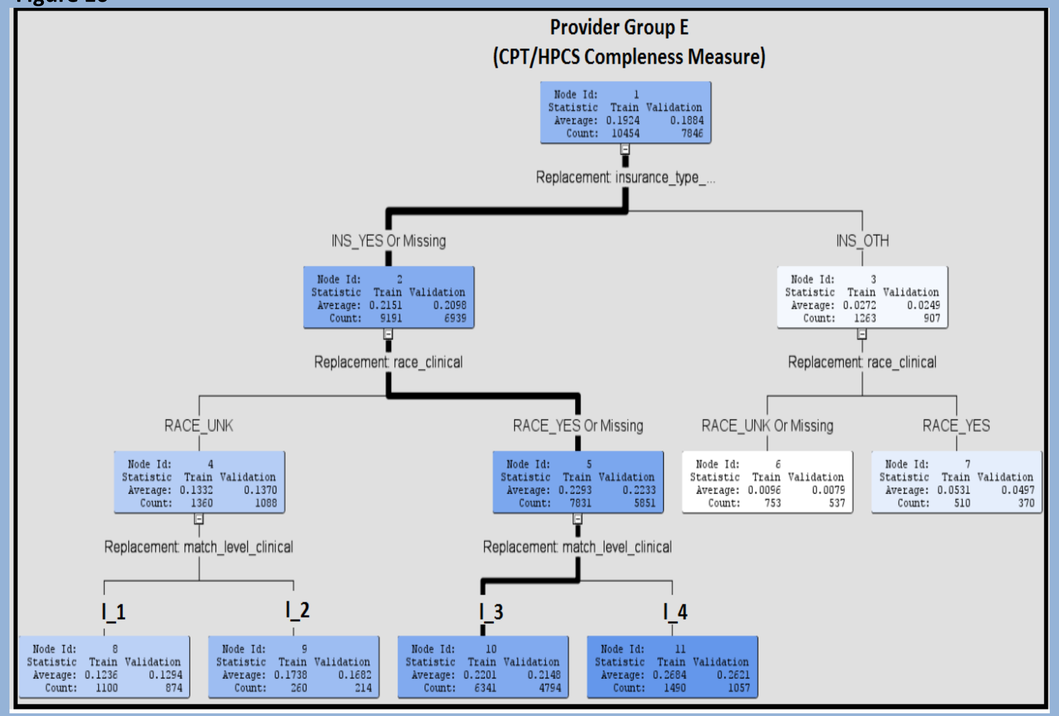
Assessing Completeness of Clinical Data through Claims Data Matching

Catherine Olson, Thomas Horstman

Optum

RESULTS CONTINUED

Figure 10



CONCLUSIONS

Each of the decision trees (for all providers and completeness measures) demonstrated that, when defining subpopulations with insurance, race, and match levels for each provider group, their completeness scores were greater than the entire population – even though values for each of these attributes vary among provider groups.

A completeness measure using administrative claims is a useful tool for assessing the completeness of clinical EHR data. Clinicians and provider groups could use these findings to identify and address reasons why their EHR records are lacking in certain demographic and other areas, and whether their practice habits result in some variations in record completeness. A next step in completeness research could be to re-evaluate this comparison of claims and EHR data over time, to determine if the quality and completeness of EHR data improved in each year of the study period.

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL