DATA DICTIONARY

Variable changes / creation:

FORMAT:

*original variable(s) -> new variable(s)*

*For new variable 1*

* *Description of changes/new values*

*For new variable k*

* *Description of changes/new values*

SEX -> female

* ‘male’ = 0; ‘female’ = 1
* 3 ‘non-bi’ values are treated as NA

HEIGHT; WEIGHT -> bmi

* (WEIGHT/2.205)/( HEIGHT/39.37)^2

INOUT -> inpatient

* ‘outpatient’ = 0; ‘inpatient’ = 1

DIABETES -> diabetes; insulin

For diabetes:

* ‘NO’ = 0; else 1

For insulin:

* ‘INSULIN’ = 1; else 0

One-hot encode

1. PRNCPTX
2. TRANST
3. ANESTHES
4. SURGSPEC
5. DYSPNEA
6. FNSTATUS2
7. WNDCLAS
8. ASACLAS
9. DISCHDEST
10. WOUND\_CLOSURE

Group? Pull out values of interest? (Any retained category will be one-hot encoded)

1. PRSEPIS
2. PODIAGTX
3. PODIAGTX10
4. PODIAG\_OTHER
5. PODIAG\_OTHER10
6. ANESTHS\_OTHER
7. COL\_\*

data\_cleaning.json

**cols\_to\_drop**: A list of variable names to remove from the data.

These variables are removed because they have too many missing values, are redundant to other data, or are otherwise uninformative for the purposes of analysis.

Also included in this list are variables that contain information about a patient after their initial admission. These “future” information cannot be used to predict “present” readmissions.

Example:

PGY and NOTHGRAFL are not recorded past 2015 NSQIP

**ensure\_before\_readmission**: A dictionary of tuples representing complications that may or may not have occurred during a patient’s initial admission to the hospital.

Each dictionary element has the following format:

*“key”: {“day\_col”, “cols\_to\_drop”}*

Where:

*key* = variable name representing a binary instance of a particular complication

*day\_col* = variable name representing number of days until particular complication occurs

*cols\_to\_drop* = list of variables related to *key* that are not used in the current analysis. Note: these variables may contain helpful information for more involved analysis.

Example:

*"REOPERATION1": {*

*"day\_col": "RETORPODAYS",*

*"cols\_to\_drop": [*

*"REOPORCPT1",*

*"REOPORICD91",*

*"REOPOR1ICD101"*

*]*

*}*

*REOPERATION1* represents whether or not there was at least 1 unplanned reoperation.

*RETORPODAYS* is the *“day\_col”* and represents the number of days between the initial admission and the first unplanned reoperation if it exists. For each patient, if this number is less than *READPODAYS1*, the reoperation will be considered part of their initial admission and retained. If the number is greater or equal to *READPODAYS1*, the reoperation will be considered part of a later readmission and will not be considered for the given patient.

*REOPORCPT1*, *REOPORICD91*, and *REOPOR1ICD101* contain information related to the first unplanned readmission, but are extraneous and subsequently dropped. Note: they may contain helpful information for more in-depth analysis.

Other notes:

RACE\_NEW is not for the model but for descriptive stats

To do:

In a .csv:

Column 1 = procedure name

Include procedures that are at least 1% of the population

Column 2 = number of patients with procedure (proportion)

Look for mutual exclusivity for data being present

Are concurrent and other procedures variables that are simply used after a certain time period? Can a patient have both a “concurrent” and an “other” procedure?

Might be worth looking into whether or not concurrent procedures and other procedures names overlap at all

Analytics on number of days between particular complication and readmittance (plus we should also know number of people with that complication who were *not* readmitted)

BLEEDIS/BLEEDDIS