Code Reviewer: Neil Kamdar

Code Co-Authors: Tom Valley, Brahmajee Nallamothu, Jessica Lehrich

Date: 8/20/2019

Title: Intensive care use and mortality among patients with ST elevation myocardial infarction: retrospective cohort study

Lines 1-87

-Code appears to be selecting STEMI across multiple years using an array of diagnosis codes and then concatenating all files together

-Important that the program documentation includes the level of analysis in the files. Consider that the program should show specifically which years of data in the comments, and whether this is patient-level, hospital-level, discharge inpatient records, etc. Since some administrative datasets also include diagnosis codes in long form and wide form alike, documentation should be supplied regarding this dataset being in wide form and that selection of diagnosis codes in any position on the inpatient claim be indicated.

Lines 88 – 120

-Please include rationale behind index admission logic. It was unclear if the index admission was just the first time listed in the dataset, or that it was representing some type of gap. For instance, can there be clarification in documentation of the definition of an index admission? Probably quite important to know in prose, since it is ambiguous in the code, what is meant by index admission in this context?

-Data functional form appears to be transposed to wide form for use in the array, and some length of stay calculations were performed. Is this what is meant by gap’num’? Is every hospitalization gap in days (LOS) between discharge date and admit date being calculated?

-Appears that OSH transfers occur when the subsequent hospital is different from the first indicated in the list for every admission, and that the length of stay between admission and discharge is 0.

Lines 121-183

-Depending on who is reviewing this, might be useful to provide some information regarding the merge step on line 132; it would also appear that transfers are dropped once they have been flagged, and merged to the original discharge record level file.

-Author is also excluding psych ICU, missing zip code records for the bene\_ids, and U.S. territories. Other data management steps are performed which flag if someone has a missing ICU indicator code and coronary code. Can the author elaborate on this step at line 150?

-Data management steps here may or may not require additional explanation. Difficult to ascertain if these variables will be used later or the manner of their use. It would appear that some decisions were made to identify those records that were intermediate care based on a combination of codes related to ICU indicator code and coronary codes being missing or with a non-missing and flagged value.

Line 198 – At this step, it would be important for a full review to show how the data management steps led to the final dataset. American Hospital Association data is being included (should be specified at the beginning of the code, as well as the years of interest. It is evident that years 2014 and 2015 are being included.

Lines 199-300

-Additional data management activities are being performed mainly via recoding of already existing variables.

-Zip code centroids are included, but this should also be documented at the beginning of the program file, and also a structure and format of the file should be specified. The source of the data (i.e. U.S. Census Bureau data) would be preferred to indicate the year, website where found, etc.

-Hospital zip code information and zip code centroids are merged with appropriate latitude and longitude. Authors should specify the purpose for these steps.

-Additional data management of discharge destination, discharge disposition of death, days alive, 30-day mortality, etc.

-NCHS data requires a description of what it is, and how it supplements the analysis.

Lines 302-530

-It would appear that some of this code could get consolidated to generate the flags. This is simply an efficiency step and could shorten the code.

-It was unclear if any of the diagnosis groupings were based simply on those coded at for the discharge record, or were some of these coded for a history of disease? For instance, it appears that respiratory failure and shock are selected at the time of the hospitalization; however, is there consideration for comorbidity history?

-Use of Elixhauser comorbidities is unclear on line 532. Can the authors provide clarification on the origin of the variable or command? Is this based on an index representing a history, or just present on admission, etc.? Some documentation would help clarify.

Lines 531 – 560

-Some data management activities on income. If the dataset for income has already been defined a priori, then no need to document at this step; however, if not used until this step, prefer that the program header indicates its use and purpose.

Lines 561 – 590

-Develop the quartiles of hospital ICU rates and flag if the ICU rate is in the top quartile or not; then merge back to the original discharge record file.

Lines 591 – 683

-Somewhat difficult to follow these steps. Can the author provide additional guidance on these steps? It would appear that data management activities are being performed which create the eventual instrumental variable (although, the code reviewer might not know this if they blinded themselves to the manuscript’s methods). If this is meant for the construction of the instrumental variable to obtain the differential distance calculation, then some explanation would be necessary on the constant values applied, and also the functional form of the data. When crossing the patient location with the hospital location, it would be helpful to see how the authors put these together.

-Can the authors explain the constant on line 655 for 300 mile exclusions? This appears arbitrary unless the centroid of the zip code of the hospital to the centroid of the zip code of the patient was usually below 300, and that these were outliers based on a distribution of distances. The code does not appear to provide any descriptive calculation on how the constant was selected.

-Line 678 is a bit unclear (and perhaps justified that it is ambiguous since the data management steps require the patient to have the highusedist and mindist attached to every patient)

Lines 684-701

-It appears that differential distance is calculated then for every patient. It is difficult to determine if these distances are based on where the patient actually goes, or whether this is based on the patient’s general proximity to centers of care? It would appear that, regardless of where the patient goes for their inpatient care, they would be associated with two possible hospitals (both of which might be different from where the patient initially went). Is this true? Can further explanation and documentation be included here regarding the rote mechanics of this calculation? Some of the code reviewer’s conclusions infer this to the case, but it is tenuous to confirm.

Lines 702-761

-Additional data management activities of hospital-level variables.

-Tertile of Medicaid were considered here; however, unsure the basis for the creation?

Lines 761-783

-Critically important, as it appears that the endogeneity test being performed is trying to ascertain a valid instrument. Is this correct? If so, it is important for the reviewer to know the final functional analytic form of the file. Can the authors provide additional information here? Also, does this generate the Durbin score statistic? Are we trying to determine if the instrument is endogenous?

-Appears that a logistic regression is run with differential distance as a covariate; was this applicable because of being a valid instrument from lines 767-774?

Lines 784-906

-Data analyses and descriptive statistics. No suggestions for clarification.

Lines 907-926

-Stored covariate list into a new variable, and included it in the 30-day mortality outcome model with robust standard errors and reported margins

-Lines 920 and 924 seem like the opposite steps from earlier. Is this true? For instance, it appears that diffdist was tested using Woolridge’s score test and then logistic regression performed with dd10 as a covariate. Dd10 was derived from diffdist, so I presume that testing diffdist as an appropriate instrument would yield identical or very similar results to dd10 (referencing lines 761-783). Therefore, in this subsequent analysis from lines 920-926, the steps appear reversed. Logistic regression is performed without dd10, and the instrumental variable (ivregress) is checked with diffdist using Woolridge’s score test. Perhaps clarification or tidying up the code a bit would help clarify any misconceptions?

Lines 928-1044

-Subgroup analysis of 30-day mortality with diffdist instrument tested. No other suggestions as long as ivregress statement is provided with additional explanation. Was the diffdist instrument reasonable for subgroups? Was that the motivation for running ivregress and NOT logistic commands?

Lines 1046-1050

-Can the authors provide clarification on these steps?

-GLMs were ran that include diffdist as a covariate, with clustering by hospital for both ICU as the dependent and death30d (30-day mortality).