Analyzing ICD-10 Diagnosis Codes with Stata

By Tiankai Wang, PhD, and Jackie Moczygemba, MBA, RHIA, CCS, FAHIMA

Due to the volume, granularity, and complexity of ICD-10 diagnosis codes, it is valuable for data analysts to know how to use statistical software to conduct ICD-10 research efficiently. This article will provide an illustration of how to use ICD-10-related Stata commands, known as icd10 commands, to analyze ICD-10 diagnosis codes in healthcare. A sample set of 2013 US mortality data from the World Health Organization (WHO) will be utilized. An instructional overview of the commands "icd10 check," "icd10 clean," "icd10 generate," and "icd10 lookup" will be provided. By applying icd10 commands, reimbursement analysts and epidemiologists are able to produce more meaningful healthcare reports.

Benefits of Analyzing ICD-10 Codes

On October 1, 2015 the United States transitioned to the 10th revision of the International Classification of Diseases set of diagnoses for coding medical encounters and inpatient procedure codes, known as ICD-10-CM/PCS. ICD-10 impacts almost all providers, health payment systems, and clinical processes. By analyzing ICD-10 diagnosis codes, healthcare data analysts can better explain costs, as well as assist with more accurate planning and selection of strategic initiatives for business development. The American Hospital Association notes other benefits, including expanded ability to perform public health surveillance, improved ability to measure the quality of healthcare services, and strengthened capability to distinguish improvements in medicine and associated technology.[1]

However, the increased granularity of ICD-10 versus that of ICD-9—such as the addition of laterality-specific codes—presents a challenge. Moreover, there are thousands of new codes, which can create enormous complexities and confusion.[2] A 2015 study published in the *American Journal of Emergency Medicine* analyzed a subset of a 2010 Illinois Medicaid database of emergency department ICD-9-CM codes. The researchers were interested in determining the accuracy of mapping tools to better prepare emergency physicians for the conversion to ICD-10-CM. This study revealed that 27 percent of 1,830 codes represented convoluted multidirectional mappings. Their analysis of the convoluted transitions revealed that eight percent (23 percent of the convoluted transitions) were clinically incorrect.[3] This finding underpins the challenges faced when migrating to a classification system with increased complexity in diagnosis codes. Proficiency in using statistical software to conduct ICD-10 research would be of tremendous value to healthcare professionals, researchers, and epidemiologists.

The most popular data analysis software systems in 2015 were SPSS, R, SAS, and Stata.[4] In SPSS, R, and SAS there are user-written commands to analyze ICD-10 codes. Those commands were developed by software users, such as professors in universities. The reliability and accuracy of the user-written packages are variable. In addition, most user-written commands lack the necessary update services. That is a serious issue in ICD-10 analysis because the Centers for Medicare and Medicaid Services frequently updates ICD-10 codes.

Currently, Stata is the only data analysis software system that provides specific commands to analyze ICD-10. Theses icd10 commands are designed to verify, clean, and analyze ICD-10 diagnosis codes. Different from other data analysis software systems, the Stata icd10 commands are official (i.e., the commands are developed by StataCorp LLC).[5] That increases the commands' reliability and ensures the commands are updated with current ICD-10 codes. For these reasons, the authors reference the use of Stata as the tool to analyze ICD-10 codes in this article. The newly released Stata version 15 uses ICD-10 2017 codes. However, the authors are using Stata version 14 in this article due to license restriction.

When data are gathered from multiple sources, they may not be fully standardized. There can also be reporting errors. Stata icd10 commands are designed to address these common challenges with secondary data. Stata icd10 commands include the following data management commands (icd10 check, icd10 clean, and icd10 generate) and interactive utilities (icd10 lookup). This article will illustrate how to use Stata icd10 commands to analyze ICD-10 diagnosis codes for healthcare research.

Figure 1: Check and Drop Invalid ICD-10 Diagnosis Codes

```
. icd10 check Cause, generate(invalid) year(2014)
(Cause contains no missing values)

Cause contains invalid codes:

    1.  Invalid placement of period                    0
    2.  Too many periods                               0
    3.  Code too short                                 0
    4.  Code too long                                  0
    5.  Invalid 1st char (not A-Z)                     0
    6.  Invalid 2nd char (not 0-9)                     2
    7.  Invalid 3rd char (not 0-9)                     0
    8.  Invalid 4th char (not 0-9)                     0
   99.  Code not defined                             350
                                                  ------
        Total                                        352
```

Figure 2: List first 10 original observations Cause, Sex, and Deaths, separating by Cause without the observation numbers

```
. list Cause Sex deaths in 1/10, sepby(Cause) noobs
```

Cause	Sex	deaths
A009	1	1
A020	1	8
A020	2	4
A021	1	16
A021	2	7
A022	1	2
A029	1	2
A029	2	1
A030	2	1
A039	1	3

Figure 3: Format ICD-10 Codes for Standard Format

```
. icd10 clean Cause, dots
(5607 changes made)

. list Cause Sex deaths in 1/10, sepby(Cause) noobs
```

Cause	Sex	deaths
A00.9	1	1
A02.0	1	8
A02.0	2	4
A02.1	1	16
A02.1	2	7
A02.2	1	2
A02.9	1	2
A02.9	2	1
A03.0	2	1
A03.9	1	3

Figure 4: List the Top Causes of Male Deaths in the US in 2013

```
. gsort -deaths

. list Cause Sex deaths in 1/10 if Sex==1, sepby(Cause)
```

	Cause	Sex	deaths
1.	I25.1	1	89340
3.	C34.9	1	85214
6.	I21.9	1	65974
10.	J44.9	1	50233

Figure 5: Look Up the Descriptions of ICD-10 Diagnosis Codes

```
. icd10 lookup I25.1

    I25.1  Atherosclerotic heart disease

. icd10 lookup F03

    F03    Unspecified dementia

. icd10 lookup P07.2

    P07.2  Extreme immaturity
```

Figure 6: Using Generate and Tabulate to Find Respiratory Illness Deaths

```
. icd10 generate resp= Cause, range (J10/J898)

. tab resp [fweight=deaths]
```

resp	Freq.	Percent	Cum.
0	2,202,569	89.91	89.91
1	247,063	10.09	100.00
Total	2,449,632	100.00	

Usage of Stata Icd10 Commands

To illustrate the usage of the commands, the most recent year (2013) of WHO mortality data of the United States is utilized. The WHO mortality data can be accessed at http://apps.who.int/healthinfo/statistics/mortality/causeofdeath_query/. The original sample data contains 6,215 observations and 33 variables.

First, verification of ICD-10 diagnosis codes for validity is needed. The icd10 check verifies that a variable contains defined ICD-10 diagnosis codes and provides a summary of any problems encountered. The syntax is "icd10 check *varname*, generate (*newvar*) year (201x)"; *varname* is the variable containing ICD-10 diagnosis codes and *newvar* is the new variable generated to indicate whether the code in *varname* is valid or invalid. A 0 indicates a valid ICD-10 code, numbers 1 through 8 indicate the reason the code is not valid, and a value of 99 is used for an undefined code. The year (201x) indicates the edition of codes used; for example, year (2014) indicates that the verification is based on ICD-10 diagnosis codes of the 2014 edition. Before conducting further analysis, the researcher will need to drop the observations with the invalid ICD-10 diagnosis codes. Figure 1 above shows the codes and outputs.

The first line, "icd10 check Cause, generate(invalid) year(2014)", is the icd check command, indicated by the dot preceding it. "Cause" is the *varname* in this dataset and "invalid" is a new variable name generated by the authors/analysts of this article. "2014" indicates that the authors/analysts use ICD-10 2014 codes in Stata version 14. In this process, Stata generates a variable named *invalid* to each observation and assigns a value either from 0 to 8, or 99. The summary of the invalid codes is presented. Among the invalid codes, two observations' "Causes" are "TOT," which is coded as "6" (i.e., "Invalid 2nd char (not 0-9)"), and 350 observations' "Causes" are not defined in the ICD-10 2014 edition because the fourth character is 9 (e.g., W00.9 (Unspecified fall due to ice and snow)). W00.9 is not a billable or specific ICD-10-CM diagnosis code as there are three codes below W00.9 that describe this diagnosis in greater detail. The command "drop if invalid==6 | invalid ==99" is to drop the observations with *invalid* equal to 6 or 99. In this process, the above 352 observations were dropped for further analysis due to invalid ICD-10 diagnosis codes identified in the prior icd10 check command.

Figure 2 above displays the original data contents. The command "list Cause Sex deaths in 1/10, sepby(Cause) noobs" requests to list the first 10 observations' *Cause*, *Sex*, and *deaths* (the frequency of deaths), separating them with lines by each *Cause* without listing the observations numbers. The variable *Cause* contains the ICD-10 diagnosis codes indicating the

causes of deaths. "1" in the variable *Sex* stands for males, and "2" stands for females. Notice that the original format of *Cause* is one letter followed by three numbers (e.g., A009). This is not a standard format of ICD-10. The command line "icd10 clean *varname*, dots" can be used to standardize the format. Figure 3 above shows the standard ICD-10 format after cleaning.

With the cleansed ICD-10 data, further analysis can be conducted. For example, a researcher may want to identify the most frequent causes of male deaths in the United States in 2013. The researcher can sort *deaths* in descending order with *gsort*, then list the first 10 observations specifying *Sex* equal to 1 (i.e., male). The output is in Figure 4 above. Thus, the researcher can see that the top cause of death is ICD-10 code I25.1 with 89,340 deaths. Accordingly, the researcher would like to know the meaning of I25.1 in the ICD-10 classification codes. This is where Stata icd10 commands are of tremendous value to the researcher. The command "icd10 lookup" displays descriptions of the diagnosis codes specified on the command line. In this example, the output of the command line "icd10 lookup I25.1" is "I25.1, Atherosclerotic heart disease." The researcher is now informed that the top cause of male deaths in the US in 2013 is atherosclerotic heart disease. With similar Stata codes specifying different conditions, researchers will construe that the top cause of female deaths in the US in 2013 is "F03, Unspecified dementia" with 88,408 deaths, and the top cause of both male and female newborn (i.e., Age=0) deaths is "P07.2, Extreme immaturity" with 3,168 deaths. The outputs are shown in Figure 5 above. The data analysis software system SPSS provides a dataset with the ICD-10 diagnosis codes description, but it needs manual lookup or merging of the ICD-10 description dataset to the target dataset. The inherent "icd10 lookup" command in Stata improves the efficiency in ICD-10 diagnosis codes lookup and avoids potential errors when using manual processes or merging datasets.

Another advantage of Stata icd10 commands is the "icd10 generate" command. For example, the researcher may want to identify all deaths due to respiratory illness, which covers 275 different ICD-10 diagnosis codes. A variable would need to be created and set equal to one of the 275 codes. Without icd10 commands, the researcher would need three command lines to generate this variable. Using the icd10 generate command with the range( ) option, the researcher can easily generate the variable with one command line. The Stata code is "icd10 generate resp= Cause, range (J10/J898)." In this process, the observations with ICD-10 codes ranging from J10 to J898 are coded to "1" in the new variable *resp*. Next, tabulate *resp* with the frequency weight option "[fweight=deaths]". The output is shown in Figure 6 above. The researcher will see that about 10 percent of all deaths in the US in 2013 were from respiratory illnesses.

The above icd10 commands can be applied to reimbursement analysis and epidemiological statistics to make working with ICD-10 diagnosis codes easier for researchers.

Stata Commands Save Time, Enhance Analytics Power

In summary, this article illustrated the usage of Stata icd10 commands with the 2013 US mortality sample data from WHO. The command "icd10 check" verifies that a variable contains defined ICD-10 diagnosis codes. The command "icd10 clean" formats the ICD-10 diagnosis codes to standard formats. The command "icd10 generate" generates new variables more efficiently. The command "icd10 lookup" displays descriptions of the diagnosis codes specified on the command line.

Stata icd10 commands provide powerful tools to verify, clean, and analyze ICD-10 diagnosis codes. With the built-in data dictionary, icd10 lookup can improve efficiency of codes lookup, and provide code descriptions automatically. Therefore, using icd10 commands can save data analysts time and avoid errors that often happen in manual processes or merging datasets. Furthermore, with the other Stata commands, researchers have the ability to complete many other tasks such as graphing summary data for data visualization and conducting regression analysis to identify the potential associations among factors.

Notes

1. AHA Central Office. "About ICD-10 Coding." www.ahacentraloffice.org/codes/ICD10.shtml.
2. Manchikanti, Laxmaiah; Frank Falco; and Joshua Hirsch. "Ready or not! Here comes ICD-10." *Journal of Neurointerventional Surgery* 5, no. 1 (2013): 86-91. http://jnis.bmj.com/content/neurintsurg/5/1/86.full.pdf.
3. Krive, Jacob et al. "The complexity and challenges of the International Classification of Diseases, Ninth Revision, Clinical Modification to International Classification of Diseases, 10th revision, Clinical Modification transition in EDs." *American Journal of Emergency Medicine* 33, no. 5 (May 2015): 713-718. www.ajemjournal.com/article/S0735-6757(15)00133-3/abstract.
4. Muenchen, Robert. "The Popularity of Data Science Software." r4stats.com. http://r4stats.com/articles/popularity/.

5. StataCorp LLC. “Commands for working with ICD codes.” www.stata.com/features/overview/icd/.

Tiankai Wang (tw26@txstate.edu) is associate professor and Jackie Moczygemba (jackiem@txstate.edu) is associate professor and chair in the health information management department at Texas State University, based in San Marcos, TX.
