Description

This entry provides a brief introduction to the basic concepts of the International Classification of Diseases (ICD). If you are not familiar with ICD terminology, we recommend that you read this entry before proceeding to the individual command entries.

This entry also provides an overview of the format of the codes from each coding system that Stata’s icd commands support. Stata supports ninth revision codes (ICD-9) and tenth revision codes (ICD-10). For ICD-9, Stata uses codes from the United States’s Clinical Modification, the ICD-9-CM. For ICD-10, Stata uses the World Health Organization’s (WHO’s) codes for international morbidity and mortality reporting. We encourage you to read this entry to ensure that you choose the correct command and that your data are in the proper format for using the icd suite of commands.

Finally, this entry provides information about using the icd commands with multiple diagnosis or procedure codes at one time. None of the commands accepts a varlist, so we illustrate methods for working with multiple codes.

If you are familiar with ICD coding and the icd commands in Stata, you may want to skip to the command-specific entries for details about syntax and examples.

Commands for ICD-9 codes

- icd9 ICD-9-CM diagnostic codes
- icd9p ICD-9-CM procedure codes

Commands for ICD-10 codes

- icd10 ICD-10 diagnosis codes

Remarks and examples

Remarks are presented under the following headings:

- Introduction to ICD coding
- Terminology
- Diagnosis codes
- Procedure codes
- Working with multiple codes

Introduction to ICD coding

The icd commands in Stata are implemented for two specific coding systems: ICD-9-CM and ICD-10.

The International Classification of Diseases (ICD) coding system was developed by and is copyrighted by the World Health Organization (WHO). The ICD coding system is used for standardized mortality reporting and, by many countries, for reporting of morbidity and coding of diagnoses during health care encounters. Since 1999, the ICD system has been under its tenth revision, ICD-10 (World Health Organization 2011). These codes provide information only about diagnoses, not about procedures.
The United States and some other countries have also developed country-specific coding systems that are extensions of WHO’s system. These systems are used for coding information about health care encounters. In the United States, the coding system is referred to as the International Classification of Diseases, Clinical Modification. These codes are maintained and distributed by the National Center for Health Statistics (NCHS) at the U.S. Centers for Disease Control and Prevention (CDC) and by the Centers for Medicare and Medicaid Services (CMS).

**Terminology**

The `icd9` and `icd10` entries assume knowledge of common terminology used in the ICD-9-CM documentation from the NCHS or CMS or in the ICD-10 revision manuals from WHO. The following brief definitions are provided as a reference.

**Edition.** The ICD-9-CM and ICD-10 each have editions, which represent major periodic changes. ICD-9-CM is currently in its sixth edition (National Center for Health Statistics 2011). ICD-10 is currently in its fifth edition (World Health Organization 2011).

**Version.** In the ICD-9-CM coding system, a new version is published each year and is issued with a sequential version number. The current version is 32.

**Update.** In the ICD-10 coding system, an update may occur each year. The update is not issued with a number but may be identified by the year in which it occurred.

**Category Code.** A category code is the portion of the ICD code that precedes the period. It may represent a single disease or a group of related diseases or conditions.

**Valid Code.** A valid code is one that may be used for reporting in the current version of the ICD-9-CM or current update to the ICD-10 edition. What constitutes a valid code changes over time.

**Defined Code.** A defined code is any code that either is currently valid, was valid at a previous time, or has meaning as a grouping of codes. See [D] icd9 and [D] icd10 for information about how the individual commands treat defined codes.

**Diagnosis Codes**

Let’s begin with the diagnostic codes processed by `icd9`. An ICD-9-CM diagnosis code may have one of two formats. Most use the format

\[
\{0–9, V\} \{0–9\} \{0–9\} [ . ] [0–9] [0–9]
\]

while E-codes have the format

\[
E\{0–9\} \{0–9\} \{0–9\} [ . ] [0–9]
\]

where braces, `{ }`, indicate required items and brackets, `[ ]`, indicate optional items.

ICD-9-CM codes begin with a digit from 0 to 9, the letter V, or the letter E. E-codes are always followed by three digits and may have another digit in the fifth place. All other codes are followed by two digits and may have up to two more digits.

The format of an ICD-10 diagnosis code is

\[
\{A–T, V–Z\} \{0–9\} \{0–9\} [ . ] [0–9]
\]
Each ICD-10 code begins with a single letter followed by two digits. It may have an additional third digit after the period.

Diagnosis codes must be stored in a string variable (see \[D]\ data types). For codes from either revision, the period separating the category code from the other digits is treated as implied if it is not present.

Technical note

There are defined five- and six-character ICD-10 codes. However, these codes are not part of the standard four-character system codified by WHO for international morbidity and mortality reporting and are not considered valid by icd10. See \[D]\ icd10 for additional details about these codes and options for using icd10 with them.

Technical note

ICD-10 codes U00–U49 are reserved for use by WHO for provisional assignment of new diseases. Codes U50–U99 may be used for research purposes to identify subjects with specific conditions under study for which there is no defined ICD-10 code (World Health Organization 2011).

If you are working in one of these specialized cases, see the technical note in Creating new variables under Remarks and examples of \[D]\ icd10.

Procedure codes

The ICD-9-CM coding system also includes procedure codes. The format of ICD-9-CM procedure codes is

\[
\{0–9\}\{0–9\}[.][0–9][0–9]
\]

Procedure codes must be stored in a string variable.

Working with multiple codes

Oftentimes, multiple diagnoses or procedures are recorded for each observation. None of the icd commands accepts a varlist, but you can still work with multiple-diagnosis or multiple-procedure records. To use the icd commands with more than one diagnosis or procedure variable at a time, you must either first *reshape* your data or use a loop; see \[D]\ reshape and \[P]\ forvalues.

Example 1: Summarizing information from multiple variables

In example 1 of \[D]\ icd9, we add a variable indicating whether each diagnosis code was invalid or undefined. Here, we use the same extract from the National Hospital Discharge Survey (NHDS).

It is often more useful to add a single variable that summarizes the results from several diagnosis or procedure variables. For example, we may wish to add a variable indicating whether a particular diagnosis code or range of codes appeared in any field. Summary variables can be created from the results of the check subcommand with option generate() or the generate subcommand with option range().
Suppose that we want a single variable that contains the number of improperly formatted or undefined codes that each discharge had. To illustrate, we use the nhds2010 dataset, keeping the variables for discharge identifier (recid), patient age, and patient sex, as well as the three diagnosis variables. We list the first three observations below.

```
. use http://www.stata-press.com/data/r14/nhds2010
   (Adult same-day discharges, 2010 NHDS)
. keep recid age sex dx1 dx2 dx3
. list in 1/10, noobs

  age  sex  dx1  dx2  dx3  recid
  -------------
    85  Female  4414  99811  14275   84
    23   Male  25013   3572  -2506  105
    63   Male  51909   1489  -V146  255
    43  Female  9678  E8528    8  651
    25  Female   V271  64421  16564  696
    57  Female  5409  V1582  2V106  779
    61  Female  27651  V1087  7V436  814
    60   Male  9951    462  -2724  826
    22   Male  42789  5409  -2780  833
    49   Male  5770  29181  14255  863
```

The data are in wide form, so we specify `reshape long` with stub dx because our diagnosis codes are in dx1, dx2, and dx3. The observation identifier, recid, is specified in i(). `reshape` creates the new variable dxnum for us.

```
. reshape long dx, i(recid) j(dxnum)
   (note: j = 1 2 3)
```

The output shows that dxnum has 3 values, so we know that all three diagnosis variables were recognized by `reshape`.

```
Data   wide   ->   long
       Number of obs.        2210   ->   6630
       Number of variables     6   ->     5
       j variable (3 values)  ->    dxnum
       xij variables:        dx1  dx2  dx3  ->    dx
```
. list in 1/9, sepby(recid) noobs

<table>
<thead>
<tr>
<th>recid</th>
<th>dxnum</th>
<th>age</th>
<th>sex</th>
<th>dx</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>1</td>
<td>85</td>
<td>Female</td>
<td>4414</td>
</tr>
<tr>
<td>84</td>
<td>2</td>
<td>85</td>
<td>Female</td>
<td>99811</td>
</tr>
<tr>
<td>84</td>
<td>3</td>
<td>85</td>
<td>Female</td>
<td>14275</td>
</tr>
<tr>
<td>105</td>
<td>1</td>
<td>23</td>
<td>Male</td>
<td>25013</td>
</tr>
<tr>
<td>105</td>
<td>2</td>
<td>23</td>
<td>Male</td>
<td>3572</td>
</tr>
<tr>
<td>105</td>
<td>3</td>
<td>23</td>
<td>Male</td>
<td>-2506</td>
</tr>
<tr>
<td>255</td>
<td>1</td>
<td>63</td>
<td>Male</td>
<td>51909</td>
</tr>
<tr>
<td>255</td>
<td>2</td>
<td>63</td>
<td>Male</td>
<td>1489</td>
</tr>
<tr>
<td>255</td>
<td>3</td>
<td>63</td>
<td>Male</td>
<td>-V146</td>
</tr>
</tbody>
</table>

Notice that our data on recid, age, and sex are retained and duplicated for each new observation. If you are working with a large dataset, you may wish to drop variables other than a merge key and your diagnosis (or procedure) variables to conserve space and speed up reshape.

After we reshape, we create prob using icd9 check, an indicator for whether there was a problem with a given diagnosis code. We then use egen to create anyprob, the total number of codes that had a problem within each recid. See [D] egen for information about summary functions.

. icd9 check dx, generate(prob)
  (dx contains 358 missing values)

dx contains invalid codes:
  1. Invalid placement of period 0
  2. Too many periods 0
  3. Code too short 177
  4. Code too long 0
  5. Invalid 1st char (not 0-9, E, or V) 875
  6. Invalid 2nd char (not 0-9) 128
  7. Invalid 3rd char (not 0-9) 0
  8. Invalid 4th char (not 0-9) 0
  9. Invalid 5th char (not 0-9) 36
 10. Code not defined 778

  Total 1,994

. generate anyprob=prob>0

. by recid, sort: egen numprobs=total(anyprob)

. list recid dxnum dx anyprob numprobs in 1/9, sepby(recid) noobs

<table>
<thead>
<tr>
<th>recid</th>
<th>dxnum</th>
<th>dx</th>
<th>anyprob</th>
<th>numprobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>1</td>
<td>4414</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>84</td>
<td>2</td>
<td>99811</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>84</td>
<td>3</td>
<td>14275</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>1</td>
<td>25013</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>2</td>
<td>3572</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>3</td>
<td>-2506</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>255</td>
<td>1</td>
<td>51909</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>255</td>
<td>2</td>
<td>1489</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>255</td>
<td>3</td>
<td>-V146</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Before we *reshape*, we drop *prob* and *anyprob* because they are specific to diagnosis variables. By construction, *numprobs* is constant within *recid*, so we do not specify it when we *reshape*.

```
. drop prob anyprob
. reshape wide dx, i(recid) j(dxnum)
   (note: j = 1 2 3)
```

Data long -> wide

<table>
<thead>
<tr>
<th></th>
<th>long</th>
<th>-&gt;</th>
<th>wide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of obs.</td>
<td>6630</td>
<td>-&gt;</td>
<td>2210</td>
</tr>
<tr>
<td>Number of variables</td>
<td>6</td>
<td>-&gt;</td>
<td>7</td>
</tr>
<tr>
<td>j variable (3 values)</td>
<td>dxnum</td>
<td>-&gt;</td>
<td>(dropped)</td>
</tr>
<tr>
<td>xij variables:</td>
<td>dx</td>
<td>-&gt;</td>
<td>dx1 dx2 dx3</td>
</tr>
</tbody>
</table>

```
. list in 1/3, noobs
```

<table>
<thead>
<tr>
<th>recid</th>
<th>dx1</th>
<th>dx2</th>
<th>dx3</th>
<th>age</th>
<th>sex</th>
<th>numprobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>4414</td>
<td>99811</td>
<td>14275</td>
<td>85</td>
<td>Female</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>25013</td>
<td>3572</td>
<td>-2506</td>
<td>23</td>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td>255</td>
<td>51909</td>
<td>1489</td>
<td>-V146</td>
<td>63</td>
<td>Male</td>
<td>1</td>
</tr>
</tbody>
</table>

The three diagnosis variables are restored to the dataset. However, we have added only a single variable showing the total number of codes with problems for each record.

### Example 2: Adding multiple variables from ICD codes

Now suppose that rather than creating a summary variable like we did in example 1, we want a new variable for each diagnosis variable in our dataset. For example, in example 1 of [D] *icd9*, we *icd9 check* each diagnosis code separately, requiring us to type the command three times. While this is not burdensome for three codes, the full NHDS includes fourteen diagnosis codes, for which we almost certainly would not want to type separate commands.

The easiest way to accomplish this is with a loop. We use *forvalues* because our codes all end in a number.

```
   (Adult same-day discharges, 2010 NHDS)
. forvalues i=1/3 {
   2.   icd9 check dx'i', generate(dx'i'_prob)
   3. }
   (dx1 contains defined ICD-9-CM codes; no missing values)
   (dx2 contains defined ICD-9-CM codes; 179 missing values)
   (dx3 contains 179 missing values)
```

**dx3 contains invalid codes:**

1. Invalid placement of period 0
2. Too many periods 0
3. Code too short 177
4. Code too long 0
5. Invalid 1st char (not 0-9, E, or V) 875
6. Invalid 2nd char (not 0-9) 128
7. Invalid 3rd char (not 0-9) 0
8. Invalid 4th char (not 0-9) 0
9. Invalid 5th char (not 0-9) 36
10. Code not defined 778

Total 1,994

This is exactly what we obtain in example 1 of [D] *icd9*. 
If our variables had not been numbered sequentially, we could have either renamed them or used `foreach; see `[P] foreach.`

The methods shown above will work for any of the `icd9`, `icd9p`, or `icd10` data-management commands.

References


Also see

[D] `icd9` — ICD-9-CM diagnosis and procedure codes

[D] `icd10` — ICD-10 diagnosis codes