Automated harmonisation of variables names and values from several datasets prior to conducting batch statistical analyses

Barcelona, 23rd October 2014
Xavier Bosch-Capblanch
Swiss TPH, Basel (Switzerland)
THE NEED AND THE PROBLEM

THE FIRST ATTEMPTS

THE SOLUTION: HARMONISATION

WAY FORWARD

PMID: 21595905; PubMed Central PMCID: PMC3123542.

Coverage = Number of children age A vaccinated with V / Number of children age A targetted by V

Pakistan

Offical country estimates  Administrative data
EPI  DHS secondary data
DHS primary data  WHO/UNICEF estimates
Global Immunization 1980-2012, DTP3 coverage
global coverage at 83% in 2012

Bangladesh 2000 (337 variables)

Variable: day vaccination DTP3 (month, year)
Value = 3 for variable DTP3 vaccination

Value = 3 for variable DTP3 vaccination

Coverage = Number of children age A vaccinated with V

Coverage = Number of children age A targeted by V
INCONVENIENCES

- Errors
- Code with hundreds of lines:
  - surveys \( \times \) variables + surveys \( \times \) variables \( \times \) values

\[
\text{If survey = Afghanistan 2000 AND OldVar = h7}
\]
\[
\quad \text{rename h7 to DPT3}
\]
\[
\quad \text{recode DTP3 1=2, 2=3, 3=0}
\]

- At survey 68: complains from collaborators
- At survey 174: nonsense, who am I? what is life?
- At survey 362: addictions

THE SOLUTION: HARMONISATION

- Use `Datasetsvars.dta`
- For each survey \( S \) in `Datasetsvars.dta`
  - For each variable of interest \( V \) for survey \( S \) in `Datasetsvars.dta`
    - `rename OldVar to NewVar`
    - `recode OldValues to NewValues`
    - Go to next variable of interest
    - Go to next survey

Folder structure

- **Setup**
  - 0. Start.do: starting 'Project' folder
  - 1. Setup.do: Stata environment
- **Main.do**: main loop (2)
- **Harmonisation**
  - 0. Start.do
  - Code
  - Logs
  - Outputs
- **Data**
  - `DatasetsVars.dta`
  - `Exceptions.dta`
  - `Thesaurus.dta`
  - `Surveys`
  - `Other`
- **Graphs**
- **Tables**
- **Text (.csv)**
Do Code C with Dataset D

**MAIN LOOP**

Do Include / exclude datasets and variables (DatasetsVars.dta)

- Folder: code  
  Code C = 1_abc.do, 2_cde.do..., n_xyz.do
  Do Code C once

- Folder: datasets  
  Dataset D = 1_abc.dta, 2_cde.dta..., n_xyz.dta
  Do Code C with Dataset D

Final procedures
- Export Outputs / tables -> Outputs / txt (xls)
- Close logs
- ...

**Initial procedures**
- Folders...
- Open logs
- ...

**Harmonisation (overview)**
- Any available dataset
- Import (and repair country and year) datasets
- Update DatasetsVars

1. Variables
- Search and identify variables
- Decisions on searches outputs
- Produce output for manual checks

2. Value labels
- Search and identify values
- Decisions on searches outputs
- Produce output for manual checks

3. Rename and recode
- ...

**1. Harmonisation of variables**

**STRATEGY B**

<table>
<thead>
<tr>
<th>Old labels</th>
<th>Normalised</th>
<th>Thesaurus</th>
<th>CHK [Day DTP 3]</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>day of dpt III</td>
<td>DAY DTP 3</td>
<td>DAY DTP 3</td>
<td>3 = 3</td>
<td>Best match</td>
</tr>
<tr>
<td>dia vacinação</td>
<td>DIA</td>
<td>VACCINACAO DTP 3</td>
<td>DTP 3</td>
<td>4 &gt; 3</td>
</tr>
<tr>
<td>dtp 3</td>
<td>DTP 3</td>
<td>DTP 3</td>
<td>2 &lt; 3</td>
<td>OUT</td>
</tr>
</tbody>
</table>

**STRATEGY C**

Old Variables used in other surveys for New Variable, which exist in current dataset:

abc, def, ghi, jkl...

```
local nT = 0
foreach cTerm of local cVarNewKTerms (if for 'dtp', '3', 'year' in 'year dtp3')
    local nT = `nT' + 1
    quiet ds
    foreach cVar in `r(varlist)' (dtp1 dtp2 dtp3...)
        local cLabel : variable label `cVar' // dtp 3 day vaccinacao
        local nFoundInThisVar = 0
        foreach cSyn of local Synos`nT'
            local nFoundInThisVar = cond(strpos(" `cLabel' ", "`cSyn' ") > 0 | `nFoundInThisVar' == 1, 1, 0)
        if `nFoundInThisVar' == 0 drop `cVar' // Drops var if none of the synonymous of the term is found.
    }
    ds
```

I drop existing 'old' variables which do not have term 1 (DTP); from the remaining, I drop those which do not have term 2 (DTP)...

```
local nT = 0
foreach cTerm of local cVarNewKTerms (if for 'dtp', '3', 'year' in 'year dtp3')
    local nT = `nT' + 1
    quiet ds
    foreach cVar in `r(varlist)' (dtp1 dtp2 dtp3...)
        local cLabel : variable label `cVar' // dtp 3 day vaccinacao
        local nFoundInThisVar = 0
        foreach cSyn of local Synos`nT'
            local nFoundInThisVar = cond(strpos(" `cLabel' ", "`cSyn' ") > 0 | `nFoundInThisVar' == 1, 1, 0)
        if `nFoundInThisVar' == 0 drop `cVar' // Drops var if none of the synonymous of the term is found.
    }
    ds
```

I drop existing 'old' variables which do not have term 1 (DTP); from the remaining, I drop those which do not have term 2 (DTP)...
NO DUPLICATES: N = 2,898

<table>
<thead>
<tr>
<th>Search</th>
<th>Type</th>
<th>SData</th>
<th>SVarNew</th>
<th>SLabelNew</th>
<th>SKeyTerms</th>
<th>Clear</th>
<th>SVarOld</th>
<th>SLabelOld</th>
<th>Case</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-</td>
<td>mics afghanistan-mic/BCG</td>
<td>BCG</td>
<td>month</td>
<td>BCG</td>
<td>month</td>
<td>BCG</td>
<td>month</td>
<td>month</td>
<td>im3bm</td>
<td>month</td>
</tr>
<tr>
<td>B-</td>
<td>mics centralafricanre BCGby</td>
<td>BCG</td>
<td>year</td>
<td>BCG</td>
<td>year</td>
<td>BCG</td>
<td>year</td>
<td>year</td>
<td>im2by</td>
<td>year</td>
</tr>
<tr>
<td>B-</td>
<td>mics cuba -mics_4-ch BCGby</td>
<td>BCG</td>
<td>year</td>
<td>BCG</td>
<td>year</td>
<td>BCG</td>
<td>year</td>
<td>year</td>
<td>im3by</td>
<td>year</td>
</tr>
<tr>
<td>B-</td>
<td>dhs am-dhs_5-ch</td>
<td>BCG</td>
<td>year</td>
<td>BCG</td>
<td>year</td>
<td>BCG</td>
<td>year</td>
<td>year</td>
<td>s2bcgy</td>
<td>year</td>
</tr>
</tbody>
</table>

ALL RECORDS: N = 15,153

2. Harmonisation of values labels

<table>
<thead>
<tr>
<th>Old value labels</th>
<th>New value label</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = old value label for 0</td>
<td>0 = old value label for 1</td>
<td>Match: 2 -&gt; 4</td>
</tr>
<tr>
<td>1 = old value label for 1</td>
<td>1 = yes</td>
<td>No match; empty 1 = yes</td>
</tr>
<tr>
<td>44 = in health centre</td>
<td>-</td>
<td>No match: manual review</td>
</tr>
</tbody>
</table>

Old value labels: 0 = abc; 1 = def; 2 = history; 4 = recall; 6 = no label; 7 = inconsistent; 8 = don't know; 9 = missing

New value labels: N = xyz; 4 = recall; 1 = yes

Normalization Thesaurus: N = old value label for N
// MATCHING VALUE LABELS. For each Value label (old), check whether there is an equivalent in any of the Value label (new).
foreach cValLblOld_i of local lValLblOld { // Test each value label old... e.g. 'caretaker recall'
    // Each one of the Value label (new) to be tested against the old one above.
    foreach cValLblNew_i of local lValLblNew {
        if !missing(mValLblNew[rownumb(mValLblNew,"`cValLblNew_i'"),2]) continue // Has already been found.
        local cValLblNew_i2 = subinstr("r(xR)" , "_", " ", .)
        // E.g. 'mother' against 'caregiver', then recall; and 'recall' against 'caregiver', then recall.
        fNormTxt "`cValLblNew_i'"
        local nTermsFound = 0
        foreach cValLblNew_i_t of local cValLblNew_i2 { // All terms of value label new i have to be in an old
            do "`cDoSynonymous'"
            local cValLblNew_i_ts = "r(xR)"
            local cValLblOld_i2 = subinstr("cValLblOld_i", "_", " ", .)
            local nIsTermFound = 0
            foreach cValLblOld_i_t of local cValLblOld_i2 { // e.g. 'mother' 'recall'
                if strpos(" `cValLblNew_i_ts' ", " `cValLblOld_i_t' ") > 0 {
                    local nIsTermFound = `nIsTermFound' + 1
                    continue, break
                }
            }
            local nTermsFound = `nTermsFound' + `nIsTermFound'
        }
    }
} // If this point is reached, there are value labels to handle (cValLbl is not missing)

NO DUPLICATES: N = 1,101

3. Rename and recode

Value labels dataset

RENAME
RECODE

When all cleared (= 1)

Dataset OldVar NewVar
mocambique-mics_3-ch n4 Sex
afghanistan-mics_4-ch n4 Sex
cuba-mics_4-ch n4 Sex
centralafricanrepublic-mics_3-ch n4 Sex

NO DUPLICATES: N = 1,101

ALL RECORDS: N = 7,891
WAY FORWARD

1. Database of surveys
2. Database of coverage from survey reports and web-based estimates
3. Harmonisation platform integrated into WUEIC engine
4. Use variables distribution
5. Availability to third parties / cross-platforms
Thanks for your infinite patience!