Coding Verbatim Responses Using an Auto-coding Program Based on a Two-step Matching Process:
National Hospital Ambulatory Medical Care Survey Emergency Department Data, 2015

Akintunde Akinseye
Brian W. Ward
Division of Health Care Statistics
National Center for Health Statistics

FCSM 2018 Research and Policy Conference
Washington, DC
March 7, 2018
The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the National Center for Health Statistics or the Centers for Disease Control and Prevention.
Outline

- Background
- Objectives
- Proposed Approach
- Results
- Summary & Next Steps
Background
Background: Scope

- The National Hospital Ambulatory Medical Care Survey (NHAMCS) is a probability survey that assesses the utilization of ambulatory medical care services in hospital emergency departments, outpatient departments, and ambulatory surgery locations.

- Started in 1992 to complement the National Ambulatory Medical Care Survey (NAMCS)

- Provides nationally-representative estimates of visits
Verbatim information is manually abstracted from medical records into narrative text variables.

Verbatim variables are coded using a specific standardized scheme:
- E.g., Reason for Visit uses a unique NCHS coding scheme.

Verbatim variables are sent to a contractor for coding; then returned to NCHS for processing/production.
Background (cont.): Verbatim Variables

- Currently in NHAMCS, medical coding occurs for 5 areas:
  1. Reason for Visit (RFV)
  2. Medical Diagnoses
  3. Cause of Injury
  4. Medical Procedures
  5. Medications Ordered or Prescribed at Visit

- For each area, potential for multiple variables:
  1. Reason for Visit – 5 variables
  2. Medical Diagnoses – 5 variables
  3. Cause of Injury – 3 variables
  4. Medical Procedures – 9 variables
  5. Medications – 30 variables
Objectives
Objectives

- Examine if there is a more efficient way to code verbatim data “in-house”

- Develop a new approach for coding the RFV verbatim information

- Assess accuracy and coverage of the new approach using 2015 NHAMCS Emergency Department (ED) data
Proposed Approach
Proposed Approach: Three-stage Model with Two-step Matching Process

- Stage 1: Percent Word Matching
- Stage 2: Direct Matching
- Stage 3: Manual Coding of Unmatched Records

Semi-Automated Steps
Proposed Solution (cont.): Three-stage Model

Data optimization steps:
- Fixing typos
- Fixing abbreviations
- Getting rid of excess words

Coded data from prev. yrs.

Percent Matching
Direct Matching

Coded dataset

RFV data dictionary: Tabular/Index

Manual coding of unmatched records
Precursor Step 1: RFV Data Dictionary

- Tabular and Index RFV Data Dictionary (RFV DD) were consolidated into a single data dictionary.

- All RFV codes sub-divided into 9 categories (or modules) based on coding scheme.
### Precursor Step 1: RFV Data Dictionary

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200.0</td>
<td>Abnormal involuntary movements</td>
<td>Includes: Jerking, Shaking, Tics, Tremors, Twitch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Excludes: Eye movements (see 1325.0-1325.4), Eyelid twitch (1340.4)</td>
</tr>
<tr>
<td>1205.0</td>
<td>Convulsions</td>
<td>Includes: Febrile convulsions (Code fever also), Fits, Seizures, Spells</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Excludes: Fainting (1030.0)</td>
</tr>
<tr>
<td>1207.0</td>
<td>Symptoms of head, NEC</td>
<td>Excludes: Headache, pain in head (1210.0)</td>
</tr>
<tr>
<td>1210.0</td>
<td>Headache, pain in head</td>
<td>Includes: Post-traumatic (also code 5575.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Excludes: Migraine (2365.0), Sinus headache (1410.1), Symptoms of head, NEC (1207.0)</td>
</tr>
<tr>
<td>1220.2</td>
<td>Increased sensation (hyperesthesia)</td>
<td>Includes: Burning legs, Burning, tingling sensation, Needle and pins</td>
</tr>
<tr>
<td>1220.3</td>
<td>Abnormal sensation (paresthesia)</td>
<td>Includes: Prickly feeling, Stinging</td>
</tr>
<tr>
<td>1220.4</td>
<td>Other disturbances of sense, including smell and taste</td>
<td>Includes: Falling sensation, Giddiness (dizziness), Lightheadedness, Loss of sense of equilibrium or balance, Room spinning</td>
</tr>
<tr>
<td>1225.0</td>
<td>Vertigo - dizziness</td>
<td>Includes: Fainting (1030.0)</td>
</tr>
<tr>
<td>1230.0</td>
<td>Weakness (neurologic)</td>
<td>Includes: Drooping, facial or NOS, Right- or left-sided weakness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Excludes: General weakness (1020.0)</td>
</tr>
</tbody>
</table>
Precursor Step 2: Optimization steps

- Fixing common abbreviations and misspellings/typos
- Removing “excess” words
- RFV DD enrichment *(not yet completed)*
Stage 1: Percent Word Matching

- Using word-based matching, verbatim entries were compared with RFV DD, and “total % match” was calculated.

- Threshold at which codes are retained is scalable.
  - Demonstration thresholds set at: any percentage, 50%, 80%, and 90%.
Stage 1 (cont.): Percent Word Matching

- Formula:
  - \( \% \text{ Match} = \frac{\text{Total Matched Words}}{\text{Total Words}} \)

<table>
<thead>
<tr>
<th>Verbatim RFV</th>
<th>RFV DD</th>
<th>Percent Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Back Pain</td>
<td>Low Back Pain</td>
<td>4/6 = 66.7%</td>
</tr>
<tr>
<td>Lower Back Pain</td>
<td>Back Pain</td>
<td>4/5 = 80.0%</td>
</tr>
<tr>
<td>Lower Back Pain</td>
<td>Hand Pain</td>
<td>2/5 = 40.0%</td>
</tr>
<tr>
<td>Lower Back Pain</td>
<td>Vomiting</td>
<td>0/4 = 0.0%</td>
</tr>
<tr>
<td>Ankle Pain</td>
<td>Ankle Pain</td>
<td>4/4 = 100.0%</td>
</tr>
</tbody>
</table>
Stage 2: Direct Matching

- Verbatim entries and corresponding RFV codes compiled into a library using previously-coded data
  - 2013-2014 NHAMCS ED coded RFV data
  - Potential to include 2012 and prior data
Stage 3: Manual Coding of Unmatched Records

- Review of unmatched verbatim by certified medical coders

- Assigned codes would be available to update and expand the coding library for subsequent data years

- Medical coders facilitate updates to library
Results
Results

- **Aim:** Assess the accuracy and coverage of Stages 1 and 2 in Model

- 2015 NHAMCS ED data (post-optimization)

- RFV1-RFV5 variables
  - 43,565 verbatims
  - 46,299 codes (assigned by contractor)
Results (cont.)

- Two separate stages:
  I. Stage 1: Only DHCS internal RFV Data Dictionary used
  II. Stage 2: Created from 2013-2014 NHAMCS ED files

- Word matching threshold(s) for Stage 1 set at:
  • Any percent, 50%, 80%, and 90%

- Results compared to “final” 2015 data (coded by contractor)
Results (cont.)

**ED 2015 RFV1-RFV5**

- 62,880 no entries, per the 5 variable matrix of RFV1-RFV5
- 43,565 verbatim entries

<table>
<thead>
<tr>
<th>Any word matching</th>
<th>coverage</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>39,318</td>
<td>90.3</td>
</tr>
<tr>
<td>incorrect</td>
<td>8,568</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>50%+ word matching</th>
<th>coverage</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>36,456</td>
<td>83.7</td>
</tr>
<tr>
<td>incorrect</td>
<td>6,643</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>80%+ word matching</th>
<th>coverage</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>27,970</td>
<td>64.2</td>
</tr>
<tr>
<td>incorrect</td>
<td>2,965</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>90%+ word matching</th>
<th>coverage</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>24,487</td>
<td>56.2</td>
</tr>
<tr>
<td>incorrect</td>
<td>2,066</td>
<td></td>
</tr>
</tbody>
</table>
Summary & Next Steps
Summary: Conclusion

❑ This approach indicates that there is potential for a more efficient way to code verbatim data “in-house” using SAS

❑ Assessment:
  • ≈92% accuracy (*Note: contractor held to 95% threshold*)
  • ≈56% coverage
Summary (cont.): Advantages/Disadvantages

- **Advantages**
  - Cost/time savings
  - Percent word matching is scalable
  - Potential to be utilized with other coding schemes

- **Disadvantages**
  - Time required to build/improve coding library
  - Potential under-coding of data
Next Steps (cont.): Key Questions for NCHS

- Is this Three-Stage Model a worthy pursuit?
- What Stage(s) should be adopted?
- Should manual coding (Stage 3) continue to be performed by a contractor, or completed in-house?
- Is there a way to reduce the amount of verbatim entries that end up in Stage 3?
  - E.g., enriching RFV DD, examining characteristics of uncoded verbatim RFVs, etc.
References

• NHAMCS RFV data dictionary

• English and Medical Dictionary: OpenOffice.org
  http://extensions.services.openoffice.org/dictionary.

• Medical abbreviations
  http://www.medabbrev.com/
Thank you!

Akintunde Akinseye
aakinseye@cdc.gov