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

Disclaimer

- 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

Background (cont.): Data Collection & Processing

- 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

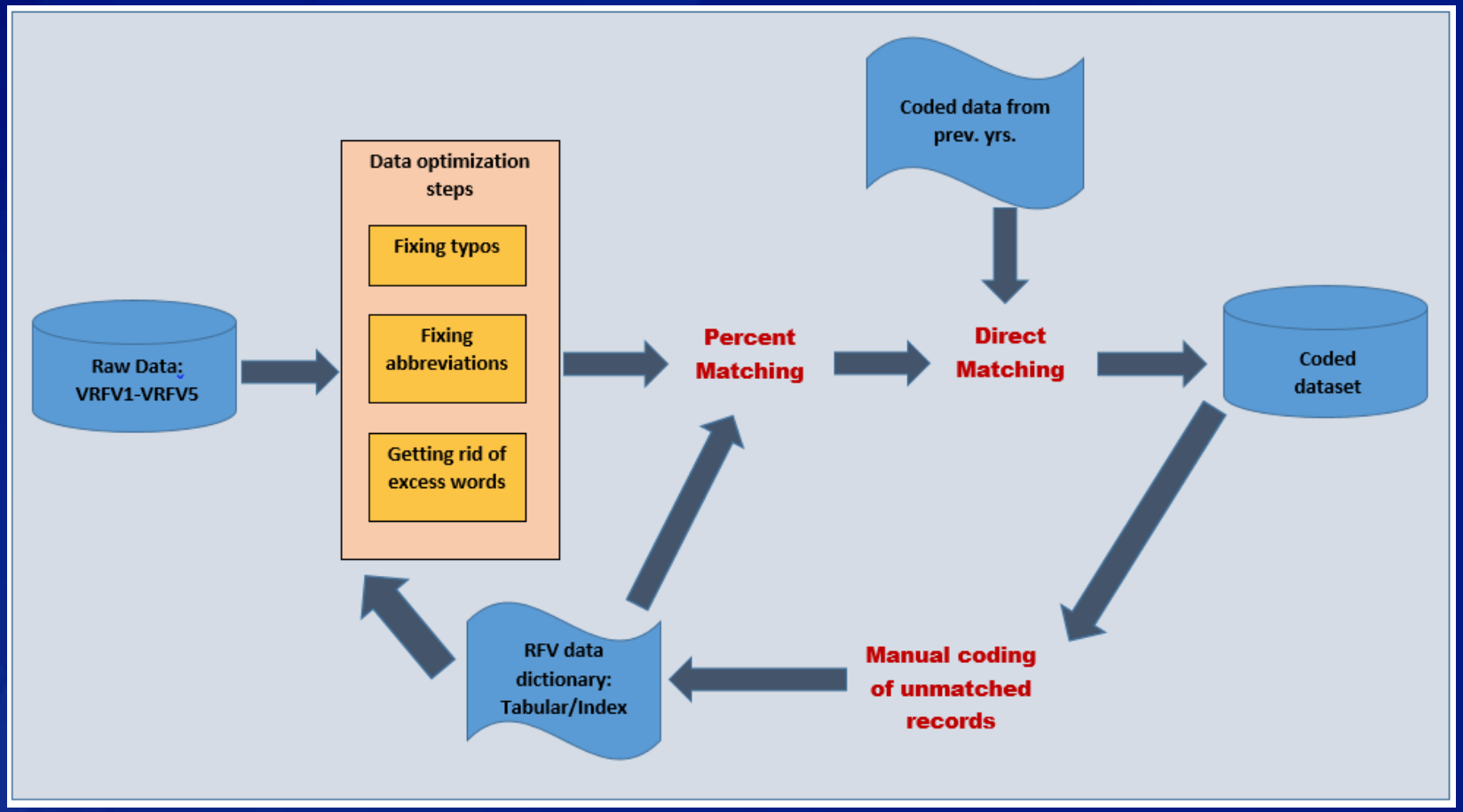
Proposed Approach: Three-stage Model with Two-step Matching Process

- Stage 1: Percent Word Matching
- Stage 2: Direct Matching

Semi-Automated Steps

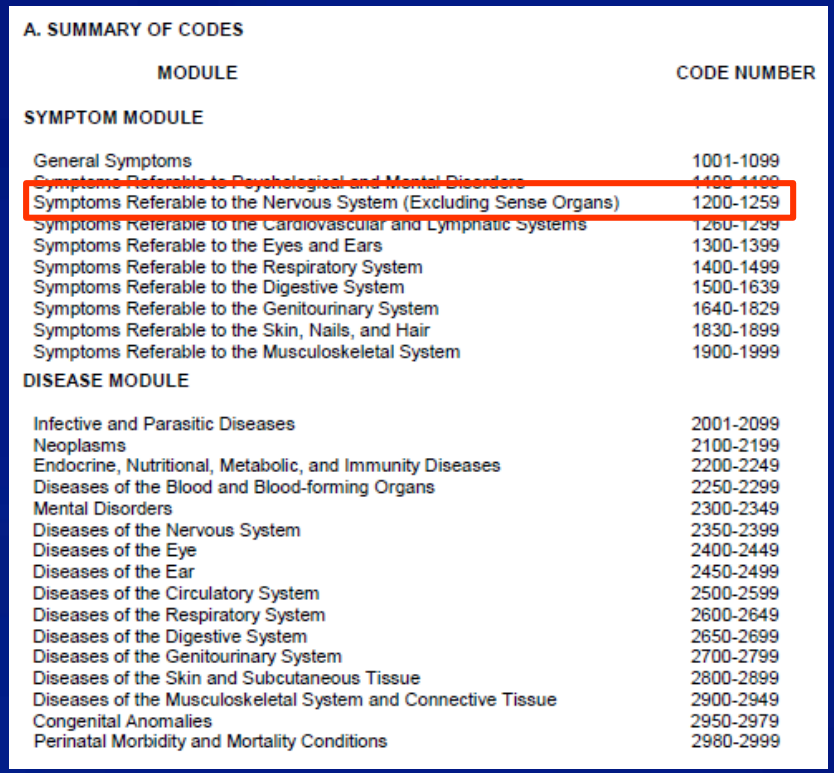
- Stage 3: Manual Coding of Unmatched Records

Proposed Solution (cont.): Three-stage Model



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



A. SUMMARY OF CODES

MODULE	CODE NUMBER
SYMPTOM MODULE	
General Symptoms	1001-1099
Symptoms Referable to Psychological and Mental Disorders	1100-1199
Symptoms Referable to the Nervous System (Excluding Sense Organs)	1200-1259
Symptoms Referable to the Cardiovascular and Lymphatic Systems	1260-1299
Symptoms Referable to the Eyes and Ears	1300-1399
Symptoms Referable to the Respiratory System	1400-1499
Symptoms Referable to the Digestive System	1500-1639
Symptoms Referable to the Genitourinary System	1640-1829
Symptoms Referable to the Skin, Nails, and Hair	1830-1899
Symptoms Referable to the Musculoskeletal System	1900-1999
DISEASE MODULE	
Infective and Parasitic Diseases	2001-2099
Neoplasms	2100-2199
Endocrine, Nutritional, Metabolic, and Immunity Diseases	2200-2249
Diseases of the Blood and Blood-forming Organs	2250-2299
Mental Disorders	2300-2349
Diseases of the Nervous System	2350-2399
Diseases of the Eye	2400-2449
Diseases of the Ear	2450-2499
Diseases of the Circulatory System	2500-2599
Diseases of the Respiratory System	2600-2649
Diseases of the Digestive System	2650-2699
Diseases of the Genitourinary System	2700-2799
Diseases of the Skin and Subcutaneous Tissue	2800-2899
Diseases of the Musculoskeletal System and Connective Tissue	2900-2949
Congenital Anomalies	2950-2979
Perinatal Morbidity and Mortality Conditions	2980-2999

Precursor Step 1: RFV Data Dictionary

1200.0 Abnormal involuntary movements

Includes:
Jerking
Shaking
Tics
Tremors
Twitch

Excludes:
Eye movements (see 1325.0-1325.4)
Eyelid twitch (1340.4)

1205.0 Convulsions

Includes:
Febrile convulsions (Code fever also)
Fits
Seizure disorders
Seizures
Spells

Excludes:
Fainting (1030.0)

1207.0 Symptoms of head, NEC

Excludes:
Headache, pain in head (1210.0)

1210.0 Headache, pain in head

Includes:
Post-traumatic (also code 5575.0)

Excludes:
Migraine (2365.0)
Sinus headache (1410.1)
Symptoms of head, NEC (1207.0)

1220.2 Increased sensation (hyperesthesia)

1220.3 Abnormal sensation (paresthesia)

Includes:
Burning legs
Burning, tingling sensation
Needles and pins
Prickly feeling
Stinging

1220.4 Other disturbances of sense, including smell and taste

1225.0 Vertigo - dizziness

Includes:
Falling sensation
Giddiness (dizziness)
Lightheadedness
Loss of sense of equilibrium or balance
Room spinning

1230.0 Weakness (neurologic)

Includes:
Drooping, facial or NOS
Right- or left-sided weakness

Excludes:
General weakness (1020.0)

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:
  - ***% Match = (Total Matched Words)/Total Words***

Verbatim RFV	RFV DD	Percent Match
Lower **Back Pain**	Low **Back Pain**	4/6 = 66.7%
Lower **Back Pain**	**Back Pain**	4/5 = 80.0%
Lower Back **Pain**	Hand **Pain**	2/5 = 40.0%
Lower Back Pain	Vomiting	0/4 = 0.0%
Ankle Pain	**Ankle Pain**	4/4 = 100.0%

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		
			(%)
Any word matching	39,318	coverage	90.3
correct	30,750	**accuracy**	**78.2**
incorrect	8,568		
50%+ word matching	36,456	coverage	83.7
correct	29,813	**accuracy**	**81.8**
incorrect	6,643		
80%+ word matching	27,970	coverage	64.2
correct	25,005	**accuracy**	**89.4**
incorrect	2,965		
90%+ word matching	24,487	coverage	56.2
Correct	22,421	**accuracy**	**91.6**
incorrect	2,066		

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

ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHAMCS

- 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