Machine learning to support the capture and use of data

Facilitator: Samantha Chao, MPH, Public Health Informatics Institute | May 20, 2021
Finding your way around Zoom

- The audience is muted.
- Type your questions into the chat or use reactions to communicate with presenters.
PLUNGED INTO THE FUTURE

Efforts in surveillance data modernization (survival) in Los Angeles County under COVID-19
Framing

• Issues:
  – Volume of data
  – Surveillance data system age
  – Limited import utility in surveillance system
  – Proliferation of laboratories not yet on ELR
  – Expectation of data reporting and interest
    • Agency leadership, media and the public

• General approach:
  – Acceptance, standardization, automation
  – Data intake, integration, analysis, reporting
Web-based Reporting
Tools

Intake
- R
- nintex
- alteryx

Matching, Analysis, Reporting
- S
- R

ETL
- SQL
- Python

Integration
- RPA bot
- sunquest
- IRIS SYSTEM
Aspirations for the future

- Big data infrastructure and tools for communicable disease surveillance
  - Hadoop data lake—engineering is in progress
- Surveillance system import API
  - Vendor development in progress for multi-threaded RESTful API
- Data science
  - Hired four data scientists and a python engineer
  - Develop and deploy models using data science framework
  - Professional development for epidemiologists
- Extend novel intake processes to other diseases
- Refine standardization and automate matching processes
- New supplemental data source identification and integration
Questions?

Rebecca Fisher, MPH MA
Chief Epidemiologist
Epidemiology and Data Unit
Acute Communicable Disease Control
Los Angeles County Department of Public Health
E-mail: rfisher@ph.lacounty.gov
Applying Machine Learning to Coding Work Information for Public Health Surveillance

Stacey Marovich, MHI, MS, PMP
Lead Health Informatics Scientist
CDC/NIOSH

Public Health Informatics Institute Data Modernization Workshop
May 20, 2021
Background

- Health and work are interrelated but work information is inadequately collected in public health activities.

- Critical need for industry and occupation (I&O) to be collected and coded into structured categories to facilitate care and inform interventions.
What is industry and occupation?

- **Industry** – The kind of business where a person works
  Examples: *Grocery Store, Hospital, Bank, Auto Manufacturer*

- **Occupation** – The kind of work a person does to earn a living
  Examples: *Bank Teller, Nurse, Inspector, Waiter*
## Industry and occupation coding challenges

<table>
<thead>
<tr>
<th>Industry</th>
<th>Occupation</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automobiles</td>
<td>Administrative</td>
<td>Inadequate information</td>
</tr>
<tr>
<td>Hole Sail</td>
<td>CRC Rep</td>
<td>Misspellings, abbreviations</td>
</tr>
<tr>
<td>9 thru 12</td>
<td>Math Teacher</td>
<td>Unusual terms</td>
</tr>
<tr>
<td>Jones &amp; Sons</td>
<td>Secretary</td>
<td>Company names for industry</td>
</tr>
<tr>
<td>Welder</td>
<td>Construction</td>
<td>Reversed entries</td>
</tr>
<tr>
<td>Metal Smith – Jewelry Designer – Studio</td>
<td>Make handmade jewelry and small objects</td>
<td>Lengthy/complicated responses</td>
</tr>
</tbody>
</table>
NIOCCS: What is it?

- NIOSH developed the **NIOSH Industry and Occupation Computerized Coding System** (NIOCCS), a web-based software tool, to translate industry and occupation (I&O) narratives to standardized I&O codes.

- Reduces high cost of manually coding I&O information while simultaneously improving uniformity of the codes.

- NIOCCS has coded over 67 million records to date.
NIOCCS: past, present, and future

- Version 1 was released in late 2012.
- Versions 1 through 3 used rules-based logic for industry/occupation auto-coder.
  - Nearly 17 million records already coded through the system
- Version 4 uses machine learning (ML) for the industry/occupation auto-coder.
NIOCCS: who uses it?

- NIOCCS has nearly 900 user accounts.
  - Over 140 accounts requested so far for new version of NIOCCS

- Global user base includes U.S. and 21 foreign countries.

- Used by variety of organizations that collect and/or evaluate information using I&O, including:
  - Occupational researchers
  - Federal government agencies
  - State/local health departments
  - Academic institutions
  - Cancer registries
Machine learning in I&O coding

- NIOCCS Version 4 applies machine learning methods to auto-code I&O narratives into standardized coding classification schemes (NAICS, SOC, Census).

- Building blocks of the NIOCCS I&O auto-coder are:
  - **Word vectors** – technique used to encode text for use in machine learning algorithms
  - **Neural networks** – machine learning algorithms used to classify I&O text into I&O codes

- Technologies used:
  - Python, Scikit Learn, Keras, TensorFlow, and pre-trained word vectors

- Training data:
  - ~1 million industry and occupation records coded by professional coders
  - Mostly comprised of free text I&O responses from death certificates
  - Split 80/10/10 between training/cross validation/testing
Key benefits of applying machine learning to I&O coding

- **Speed**
  - Orders of magnitude faster execution times
- **Easier to maintain and more adaptable**
  - Transition from rules-based to data-driven
- **Improve uniformity of coding**
  - Reducing human factor
- **Support more occupational safety and health (OSH) research**
  - Increase NIOSH and others’ ability to code more data for OSH research
- **Auto-codes 100 percent of records**

<table>
<thead>
<tr>
<th>NIOCCS Version</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-Coded Rate</td>
<td>59%</td>
<td>66%</td>
<td>84%</td>
<td>100%</td>
</tr>
<tr>
<td>Accuracy Rate</td>
<td>~90%</td>
<td>~90%</td>
<td>93%</td>
<td>&gt;90%</td>
</tr>
</tbody>
</table>
Tools for coding I&O data

- Batch Coding
  - NIOCCS – Code a file function

- Single Record Coding
  - CDC EPI Info
    - I&O template released in May 2020 in Version 7.2.4
    - Custom, localized implementations
    - NIOCCS web service
Tools for coding I&O data

- NIOCCS web service
  - Uses same machine learning models as NIOCCS
  - Accessed for free via a simple web call
  - No login or account required
  - Non-proprietary/system agnostic
    - JSON output
  - No data is stored/retained on the NIOSH server
  - Example code and more information available at:
Acknowledgements

**NIOCCS System Development Team**
Stacey Marovich (NIOSH), Project Officer
Kelly Vanoli (NIOSH), Lead Developer
Jennifer Cornell (NIOSH), Technical Information Specialist
Jeff Purdin (Contractor), Developer/Data Manager
Pam Schumacher (Contractor), I&O Coding SME

**Industry and Occupation Data Coding Team**
Matt Hirst (Contractor), Coding Team Lead
Liz Smith (Contractor), QC Coder
Katrina Bicknave (Contractor), Coder
Sue Burton (Contractor), Coder
Crystal Greene (Contractor), Coder
Krista Purdin (Contractor), Coder
Rebecca Purdin (Contractor), Coder
Joetta Thornton (Contractor), Coder
Surprese Watts (Contractor), Coder

---

**Contact Us**

email: NIOCCS@CDC.GOV

https://csams.cdc.gov/nioccs/
Next steps

• Continue the discussion on the message board (link provided in the chat)

• Ten-minute break: 1:30–1:40 PM EST

• Next session: 1:40 – 1:55 PM EST
  • *Data modernization planning toolkit overview*
Thank you.