Capture and coding of industry and occupation measures: Findings from eight National Program of Cancer Registries states

MaryBeth B. Freeman, MPH[1], **Lori A. Pollack, MD, MPH**[1], **Judy R. Rees, BM, Bch, PhD**[2], **Christopher J. Johnson, MPH**[3], **Randi K. Rycroft, MSPH, CTR**[4], **David L. Rousseau, BS**[5], and **Mei-Chin Hsieh, MSPH, CTR**[6] **On behalf of the enhancement of NPCR for comparative effectiveness research team**

[1]Cancer Surveillance Branch, Division of Cancer Prevention and Control, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, Georgia [2]New Hampshire State Cancer Registry and the Geisel School of Medicine at Dartmouth, Department of Epidemiology, Hanover, New Hampshire [3]Cancer Data Registry of Idaho, Boise, Idaho [4]Colorado Central Cancer Registry, Denver, Colorado [5]Rhode Island Cancer Registry, Providence, Rhode Island [6]Louisiana Tumor Registry and Epidemiology Program, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, Louisiana

Abstract

Background—Although data on industry and occupation (I&O) are important for understanding cancer risks, obtaining standardized data is challenging. This study describes the capture of specific I&O text and the ability of a web-based tool to translate text into standardized codes.

Methods—Data on 62 525 cancers cases received from eight National Program of Cancer Registries (NPCR) states were submitted to a web-based coding tool developed by the National Institute for Occupational Safety and Health for translation into standardized I&O codes. We determined the percentage of sufficiently analyzable codes generated by the tool.

Correspondence: MaryBeth B. Freeman, MPH, Centers for Disease Control and Prevention, 4770 Buford Hwy. MS F-76. Atlanta, GA 30341-3717. yrj8@cdc.gov.

Results—Using the web-based coding tool on data obtained from chart abstraction, the NPCR cancer registries achieved between 48% and 75% autocoding, but only 12–57% sufficiently analyzable codes.

Conclusions—The ability to explore associations between work-related exposures and cancer is limited by current capture and coding of I&O data. Increased training of providers and registrars, as well as software enhancements, will improve the utility of I&O data.

Keywords

cancer; industry; NPCR; NIOCCS; occupation

1 | INTRODUCTION

There are well-known associations between occupational exposures and cancer risk, such as mesothelioma and asbestos fiber exposure in insulation workers, plumbers, and welders.[1–3] The National Institute of Occupational Safety and Health (NIOSH) estimates that occupational exposures contribute to 40 000 new cancer cases and 20 000 cancer deaths annually in the United States.[4] Furthermore, a review of previous literature estimated that 2–8% of cancer may be attributable to occupational exposures, although the true burden may be higher.[5]

Congress established the National Program of Cancer Registries (NPCR) in 1992—a national cancer surveillance system administered by the Centers for Disease Control and Prevention (CDC)—to collect information on cancer diagnoses, initial treatment, and outcome. In addition, NPCR cancer registries are directed to collect information on the "industrial oroccupational history ofthe individuals with thecancers, to the extent such information is available from the same record."[6] To this end, NPCR requires each cancer registry to collect cancer patients' usual occupation ("type of job patient engaged in for the —greatest number of working years") and usual industry ("type of business or industry where patient worked in his or her usual occupation"). [7] These data are collected through medical record abstraction (consolidated from hospitals, outpatient facilities, death certificates, and other sources) without any direct contact with patients or their families.

Although NPCR cancer registries are required to collect industry and occupation (I&O) data, there is not a standard practice for collection of these variables.[8] Certified tumor registrars (hereinafter referred to as registrars) collect I&O from text narratives within medical records that have been entered by a variety of individuals. Often, these data may be missing, incomplete, lacking detail, and/or more likely to be ascertained in patients with cancers that have known association to occupational exposures (ie, differential misclassification).[9] Furthermore, coding I&O data manually is a time and resource intensive process that is prohibitive in many cases; the availability of a tool that could reduce the time of manual coding would be helpful in a myriad of studies of occupation, not just occupational cancer.

In 2010, the CDC received additional funding from the American Recovery and Reinvestment Act to enhance standard NPCR data collection to support Comparative Effectiveness Research in 10 NPCR cancer registries.[10] A major objective of this project

was to explore the feasibility of collecting new data variables and to improve the quality of variables with known deficiencies, such as I&O. This project specifically focused on four cancer sites: breast, colon, rectum, and chronic myelogenous leukemia (CML) to better understand the role of biomarkers and related treatment through comparative effectiveness research.

This study focuses on the methods and quality of I&O collection within the Comparative Effectiveness Research project. Our objectives are to describe data capture of I&O through chart abstraction in the cancer registries that received additional funding and training through NPCR to enhance data collection, as well as to explore the capabilities of the NIOSH Industry & Occupation Computerized Coding System (NIOCCS), a web-based system that translates I&O text into standardized I&O codes.

2 | MATERIALS AND METHODS

The data from this study include all cases combined of breast, colon and rectum cancers and CML diagnosed in Alaska, Colorado, Idaho, Louisiana, North Carolina, New Hampshire, Rhode Island, and Texas, and within two specified county groupings within California and Florida.[10] The Comparative Effectiveness Research project was approved by the CDC Institutional Review Board. Patient consent was not needed because the submitted data were de-identified before being received by the CDC. At the time of the Comparative Effectiveness Research project, only Florida and Idaho submitted I&O codes; California and Rhode Island submitted only I&O text; and Alaska, Colorado, Louisiana, North Carolina, New Hampshire, and Texas submitted both I&O codes and text.

NIOCCS is a web-based system that translates I&O text into standardized I&O codes. This system is freely available for "use by researchers, government agencies and other organizations that collect or evaluate information using I&O."[11] I&O text data can be inputted through a "slim file format" (unique identifier, industry title, and occupation title) or an "expanded file format" (slim format plus employer company name, job duties, employer city, state and zip code, age, education level, and two user defined fields), and fields must be delimited by a Tab or Pipe character. NIOCCS auto-codes text data based on two confidence levels: medium (70%) and high (90%). As these confidence levels are defined as "only matched candidates where NIOCCS has [70/90]% or greater confidence of accuracy will be automatically coded," there is some variation in the results by confidence level.[11] The NIOCCS coding scheme has been described previously.[12] The difference between the confidence levels assigned is based upon a number of factors including a word-swapped factor, synonym factor, and the weight associated with the process of auto-coding. NIOCCS also has "computer-assisted coding" where users are provided "information and functions" to help select the correct code; however, this tool requires an understanding of I&O coding.[11] NIOCCS includes potential coding outputs for different I&O classification schemes (Census 2000, 2002, 2010) as well as the associated North American Industry Classification System and/or Standard Occupational Classification codes. The North American Association of Central Cancer Registries data dictionary recommends that I&O coding be completed at the registry level due to the need for specially trained and qualified personnel.[13]

At the time of the Comparative Effectiveness Research project's data collection, NPCR cancer registries were encouraged to recode textual I&O data into census numeric codes, using NIOCCS (version 1). Registries received financial and technological support during the Comparative Effectiveness Research project, including assistance from NIOSH through a presentation at an annual professional conference, an available online training, and a hard copy instruction booklet.

We examined the data received from the cancer registries for missing values and then submitted the available text data from eight of the ten registries to the NIOCCS tool (version 2) at CDC. Text data from Florida and Idaho were not available because they coded I&O at the cancer registry before submitting to CDC. Both medium (70%) and high (90%) NIOCCS confidence levels were used to auto-code I&O text. Neither the "computer assisted coding" process nor manual coding were performed for this project due to time and staffing limitations.

SAS v. 9.3 (32) was used to determine the percentage of the NIOCCS auto-coded narratives that defined an occupation (eg, nurse) and industry (eg, healthcare); we deemed these sufficiently analyzable.[14] We defined "insufficient codes" as codes that were unknown (which includes those cases missing I&O text), retired, never worked or military. The auto-coding ability of the NIOCCS tool was compared by confidence level and by state. Finally, we examined the auto-coded data for the percentage of sufficiently analyzable codes by confidence level and state.

3 | RESULTS

Registrars abstracted information from the medical records of 62 525 persons with newly diagnosed breast, colon, or rectum cancer or CML. The demographic and tumor characteristics of the patients have been described previously.[10]

Tables 1 and 2 illustrate the overall completeness of the data submitted by each cancer registry. Only four registries (Florida, Idaho, Louisiana, and North Carolina) of the eight registries that submitted I&O code had coded values for more than 75% of cases, while for each of the eight registries that submitted I&O text, some text was provided for 79–100% of cases. Overall, code data were missing occupation in 38 035 cases (50%) and missing industry in 38 456 cases (51%), while text was missing in 18 047 cases (24%) and 19 799 cases (26%), respectively.

Of the available I&O text, NIOCCS was unable to auto-code 21 586 cases (35%) for industry and 20 849 cases (33%) for occupation (Fig. 1). Auto-coded data include codes for sufficiently analyzable data as well as unknown, retired, never worked, and military. The percentage of text that was auto-coded at the high confidence level ranged from 48% to 75%. Auto-coding results for occupation text were similar to those for industry text. Using a medium confidence level, the percentage of occupation and industry text fields auto-coded by NIOCCS ranged from 56% to 80% (data not shown).

Figure 2 shows that in many cancer registries, the I&O data were auto-coded as "unknown" by the NIOCCS tool (eg, 66% industry/63% occupation, 73/72%, 64/62%, and 58/57% in

Alaska, California, North Carolina, and Texas, respectively at the high confidence level). Overall, the NPCR registries achieved between 12% and 57% sufficiently analyzable codes using NIOCCS set at the high confidence levels; two registries (New Hampshire and Rhode Island) achieved greater than 40%. The percentage of cases auto-coded as retired, never worked, or military varied by registry from 0% to 18%. At the medium confidence level, the percentage of sufficiently analyzable data ranged from 15% to 67% (data not shown).

4 | DISCUSSION

This study examined the capture of I&O from medical record abstraction for cancer surveillance and demonstrated that a freely available tool can assist in assigning I&O codes from text to enable analysis. Among the auto-coded data, occupational results mirrored industry results. For both I&O, data for many cases were missing, unknown, or otherwise insufficient for analysis (43–87% of auto-coded cases).

NPCR registries that were included in our paper received additional funding to enhance routine cancer registry practice. Thus, participating registries were incentivized to capture I&O through abstraction. One of the main difficulties facing cancer registries concerning I&O data is that medical records often have insufficient documentation for I&O fields, especially for elderly patients who have retired. Furthermore, registry staff reported that I&O coding requires extensive manual review and processing, on top of the intense data collection, consolidation, and cleaning involved for the CDC's regular reporting requirements for registries. The participating cancer registries identified the need for ongoing training for registrars so that they can collect better quality text information on I&O that could more easily be coded by registries, or auto-coded by the NIOCCS tool.

Ten years prior to the Comparative Effectiveness Research project, researchers in Massachusetts examined collection of I&O within their cancer registry. This study revealed that detailed medical record review improved either the presence or detail of I&O information for 32% of the 1 020 cases reviewed.[9] The researchers also noted the lack of consistency in documentation within the medical record of information related to I&O. Furthermore, they cited the need for training and time allocation for hospital registrars to continue detailed record reviews.

Among the Comparative Effectiveness Research project cancer registries, New Hampshire had the highest percentage of sufficiently analyzable codes auto-coded by the NIOCCS tool. Several years before the Comparative Effectiveness Research project, the New Hampshire State Cancer Registry provided statewide training to support better capture of I&O data, and found that I&O data quality could be "substantially improved by means of minimal training provided to cancer registrars to highlight the importance of these data."[15] The increased capture of I&O data highlighted by the New Hampshire study was also seen in the results from this study.

The Louisiana Tumor Registry (LTR) and Texas Cancer Registry (TCR) collaborated to examine their I&O data from diagnosis years 2010 and 2011 using the NIOCCS tool. Both cancer registries completed manual coding as well as used the NIOCCS auto-coding. The

findings from their study were similar to our results, with "44.2% of TCR records and 31.1% of LTR records" that were missing or unknown.[16] This study further highlighted the importance of high quality I&O text data to help "maximize the efficiency of NIOCCS."[16]

Additional analyses of the combined I&O data showed there are some specific factors that can be considered in research studies since standards for a minimum number of sufficient codes do not currently exist. We found that limiting our population to ages 18–64 years reduced the number of "retired" text fields, decreased the percentage of cases auto-coded, and, among the auto-coded data, decreased the percentage of insufficient codes by 10%, on average (data not shown). Similarly, researchers could consider restricting the study to ages 18–64 years since more accurate I&O information may be recorded in their patient file given that this population is more likely to be working. However, because many solid tumors linked to occupational exposures occur at advanced ages post-retirement, this would also limit the ability to identify a portion of occupational cancer cases. Finally, NIOCCS has the flexibility of two confidence levels that can be used for the auto-coding process so that researchers can balance capture versus confidence for their analyses.

Despite the limitations noted above, previous analyses of cancer registry I&O data have shown similar results to those seen in etiologic studies.[3] Furthermore, an untried approach to addressing the previously mentioned gaps is the use of job exposure matrices, which assign occupational exposure levels based upon I&O data and might be able to overcome some of the reported limitations of cancer registry I&O data.[17] However, the usefulness of job exposure matrices are subject to the accuracy of the underlying I&O data they are being applied to.

The use of I&O data for cancer studies may be limited by the elicitation and recording of data. Improvement in collection of I&O data could occur through revision of forms, training of healthcare providers to elicit, and record this information routinely (rather than only when the illness suggests an occupational component—ie, leading to differential misclassification), and training medical personnel to probe more about lifetime occupation when provided "retired" in response to questions about occupation. As shown in the New Hampshire study, cancer registrars could improve their ability to routinely capture I&O data through training on standardization of placement of this information within the medical record.[15] Additionally, successful incorporation of I&O as structured data in electronic health records could improve the data available for use in registries. Improvements in the NIOCCS tool could help registries to more easily and quickly convert I&O text into usable code. A new coding engine, restructured knowledgebase and other underlying databases have been developed for the next release of NIOCCS (expected summer 2017). Early test results for cancer registry data show an increase in the number of records autocoded. Finally, individuals from the National Cancer Institute have developed an algorithm called Standardized Occupation Coding for Computer-assisted Epidemiologic Research (SOCcer) that maps job titles into standardized occupation classification (SOC) codes, which may also be useful in epidemiological studies of occupational exposures.[18]

When the information obtained by registries improves, there will remain limitations regarding identifying the carcinogenic exposure, level of exposure and time latency between

exposure, and cancer diagnosis. Increased knowledge about occupations and exposures, however, is an important public health issue to protect workers and identify potentially harmful exposures.[3]

The strengths of this study include the number of cancer cases and registries included, which reflect 27.3% of the U.S. population,[10] and the identification of specific factors for consideration when analyzing I&O with cancer cases. The limitations of this study are as follows. We could not compare NIOCCS auto-coded data with manually coded data due to staffing limitations; the Comparative Effectiveness Research project focused on four cancer sites; only one year of data was analyzed; and data were missing or unknown in many cases. Analyzing the I&O data by cancer site could highlight the differences in capture between those with and without known occupational exposure linkages. However, among the data for this project, there was no significant difference between the percentages of auto-coded cases even though CML is related to occupational exposure and breast, colon, and rectum are not (data not shown).[19]

One future direction of I&O data collection could include the use of natural language processing in electronic health records, which has been successfully used to advance cancer care.[20] Another direction includes linkages of data between cancer registries and occupational registries as has been done to improve race and ethnicity data through linkages with Indian Health Services records.[21]

In conclusion, while there are known associations between occupational and industrial exposure to cancer, the ability to explore such associations is limited by the capture and coding of I&O data. Ultimately, emphasis on training of providers and registrars, as well as future software enhancements, will improve the utility of I&O data and further occupational cancer research.

Acknowledgments

administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Centers for Disease Control and Prevention.

References

1. Roelofs CR, Kernan GJ, Davis LK, Clapp RW, Hunt PR. Mesothelioma and employment in massachusetts: analysis of cancer registry data 1988–2003. Am J Ind Med. 2013; 56:985–992. [PubMed: 23788253]

2. Tsai RJ, Luckhaupt SE, Schumacher P, Cress RD, Deapen DM, Calvert GM. Acute myeloid leukemia risk by industry and occupation. Leuk Lymphoma Early Online. 2014; 55:2584–2591.

3. Tsai RJ, Luckhaupt SE, Schumacher P, Cress RD, Deapen DM, Calvert GM. Risk of cancer among firefighters in california, 1988–2007. Am J Ind Med. 2015; 58:715–729. [PubMed: 25943908]

4. National Institute for Occupational Safety and Health (NIOSH). [accessed November, 2016] Cancer, Reproductive, Cardiovascular, and Other Chronic Disease Prevention Program. Centers for Disease Control and Prevention website [Internet]. Updated May 26, 2017Available at: https://www.cdc.gov/niosh/programs/crcd/description.html

5. Purdue MP, Hutchings SJ, Rushton L, Silverman DT. The proportion of cancer attributable to occupational exposures. Ann Epidemiol. 2015; 25:188–192. [PubMed: 25487971]

6. Cancer Registries Amendment Act of 1992, Public Law 102–515, 106 Stat. 3372.

7. Department of Health and Human Services Centers for Disease Control and Prevention National Institute for Occupational Safety and Health. A Cancer Registrar's Guide to Collecting Industry and Occupation. 2011. DHHS (NIOSH) Publication No. 2011-173

8. Hutton MD, Burnett CA. Occupation of cancer patients: a challenge to healthcare facilities. J Ahima. 1996; 67:64–68.

9. Levy J, Brooks D, Davis L. Availability and quality of industry and occupation information in the Massachusetts cancer registry. Am J Ind Med. 2001; 40:98–106. [PubMed: 11439402]

10. Chen VW, Eheman CR, Johnson CJ, et al. Enhancing cancer registry data for comparative effectiveness research (CER) project: overview and methodology. J Registry Manag. 2014; 41:103–112. [PubMed: 25419602]

11. Centers for Disease Control and Prevention National Institute for Occupational Safety and Health (NIOSH). [accessed May, 2016] User Manual: NIOSH Industry and Occupation Computerized Coding System (NIOCCS) [Internet]. [Updated September, 2014]. Available at: https://www.cdc.gov/niosh/topics/coding/pdfs/nioccs_user_manual_2014_sept.pdf

12. Schmitz M, Forst L. Industry and occupation in the electronic health record: an investigation of the national institute for occupational safety and health industry and occupation computerized coding system. JMIR Med Inform. 2016; 4:e51–e59.

13. Thornton, M., editor. [accessed May, 2016] North American Association of Central Cancer Registries Standards for Cancer Registries Volume II Version 16 Data Standards and Data Dictionary [Internet]. 2015. [Updated November, 2015Available at: https://www.naaccr.org/data-standards-data-dictionary/

14. SAS version 9.3 (32). Copyright © 2002–2010 by SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

15. Armenti KR, Celaya MO, Cherala S, Riddle B, Schumacher PK, Rees JR. Improving the quality of industry and occupation data at a central cancer registry. Am J Ind Med. 2010; 53:995–1001. [PubMed: 20860053]

16. Weiss NS, Cooper SP, Socias C, Weiss RA, Chen VW. Coding of central cancer registry industry and occupation information: the texas and louisiana experiences. J Registry Manag. 2015; 42:103–110. [PubMed: 27028094]

17. Calvert GM, Rice FL, Boiano JM, Sheehy JW, Sanderson WT. Occupational silica exposure and risk of various diseases: an analysis using death certificates from 27 states of the United States. Occup Environ Med. 2003; 60:121–128.

18. Russ DE, Ho KY, Colt JS, et al. Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. Occup Environ Med. 2016; 73:417–424. [PubMed: 27102331]

19. Luckhaupt SE, Deapen D, Cress R, Schumacher P, Shen R, Calvert GM. Leukemia among male construction workers in California, 1988–2007. Leuk Lymphoma. 2012; 53:2228–2236. [PubMed: 22563817]

20. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. JAMA Oncol. 2016; 2:797–804. [PubMed: 27124593]

21. Espey DK, Wiggins CL, Jim MA, Miller BA, Johnson CJ, Becker TM. Methods for improving cancer surveillance data in American Indian and Alaska Native populations. Cancer. 2008; 113:1120–1130. [PubMed: 18720372]
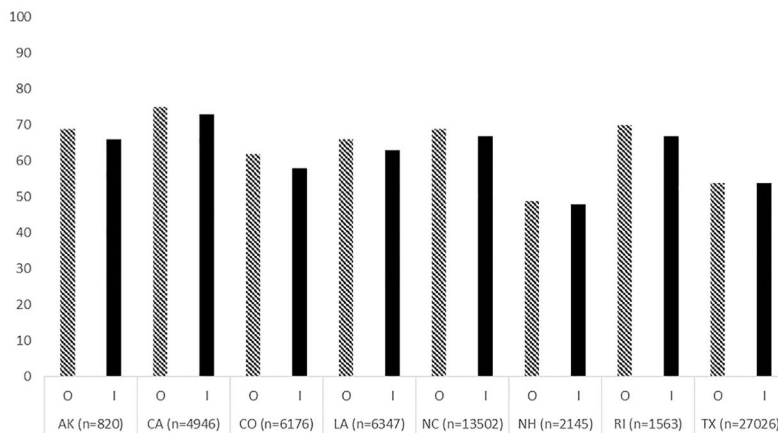
FIGURE 1.
Percent of occupation (O) and industry (I) text fields auto-coded[1] by the NIOCCS tool at high (90%) confidence level, eight cancer registries. [1]"Auto-coded" data include codes for specific industries and occupations, as well as unknown, retired, never worked, and military. AK, Alaska; CA, California; CO, Colorado; LA, Louisiana; NC, North Carolina; NH, New Hampshire; RI, Rhode Island; TX, Texas. Florida and Idaho were excluded because text data were unavailable
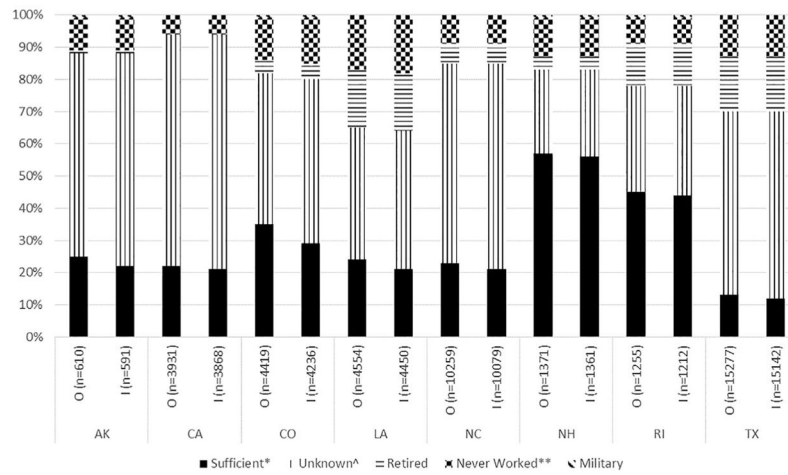
FIGURE 2.
Comparison of types of occupation (O) and industry (I) data auto-coded[1] by registry at high (90%) confidence level of the NIOCCS tool, eight cancer registries. *Sufficient includes codes that define an occupation (eg, nurse) and industry (eg, healthcare). ^Unknown includes cases missing I&O text. **Never worked includes homemaker, student, volunteer, and never worked. [1] I&O text that was not auto-coded is not included in this figure. AK, Alaska; CA, California; CO, Colorado; LA, Louisiana; NC, North Carolina; NH, New Hampshire; RI, Rhode Island; TX, Texas. Florida and Idaho were excluded because text data were unavailable

TABLE 1

Completeness of occupation code and text data submitted by 10 states to the Centers for Disease Control and Prevention

Registry	Occupation codes (% of cases)		Occupation text (% of cases)		
	Available	Missing	Available	Missing	Total cases
Alaska	22 (2.7)	798	820 (100)	0	820
California[a]	–	4946	4921 (99)	25	4946
Colorado	58 (0.9)	6118	6043 (98)	133	6176
Florida[b]	11048 (100)	0	–	11048	11048
Idaho	1917 (100)	0	–	1917	1917
Louisiana	6345 (99)	2	6339 (99)	8	6347
North Carolina	10398 (77)	3104	11217 (83)	2285	13502
New Hampshire	7 (0.3)	2138	2145 (100)	0	2145
Rhode Island	–	1563	1563 (100)	0	1563
Texas	7660 (28)	19366	24395 (90)	2631	27026
Total		38035		18047	75490

[a] 13 counties of the Sacramento region of California.

[b] Five metropolitan counties of Miami, Florida.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE 2

Completeness of industry code and text data submitted by 10 states to the Centers for Disease Control and Prevention

Registry	Industry codes (% of cases)		Industry text (% of cases)		Total cases
	Available	Missing	Available	Missing	
Alaska	24 (3)	796	820 (100)	0	820
California[a]	–	4946	4919 (99)	27	4946
Colorado	46 (0.7)	6130	6019 (97)	157	6176
Florida[b]	11048 (100)	0	–	11048	11048
Idaho	1917 (100)	0	–	1917	1917
Louisiana	6345 (99)	2	5036 (79)	1311	6347
North Carolina	10599 (79)	2903	11136 (82)	2366	13502
New Hampshire	14 (0.7)	2131	2145 (100)	0	2145
Rhode Island	–	1563	1563 (100)	0	1563
Texas	7041 (26)	19985	24053 (89)	2973	27026
Total	38456		19799		75490

[a]13 counties of the Sacramento region of California.

[b]Five metropolitan counties of Miami, Florida.