Original Article

Misclassification of Sex in Central Cancer Registries

Recinda L. Sherman, MPH, PhD, CTR[a]; Francis P. Boscoe, PhD[b]; David K. O'Brien, PhD, GISP[c]; Justin T. George, MPH[d]; Kevin A. Henry, PhD[e]; Laura E. Soloway, PhD[f]; David J. Lee, PhD[g]

Abstract: Background: Intrarecord edits on site–sex combinations are a standard tool to identify errors in the coding of sex in cancer registry data. However, the percentage of sex-specific cancers, like cervix, is low (20% of total invasive cases). Visual review and follow-back to improve the quality of the sex coding is labor intensive and typically only performed as a special project on subsets of data. The New York State Cancer Registry (NYSCR) created an edit for identifying potential sex misclassification in cancer registry data and has made its components available for use through the North American Association of Central Cancer Registries (NAACCR). The edit uses the most popular male and female first names based on decade of birth to identify potentially miscoded cases. This paper provides a summary of 3 independently conducted assessments of the sex edit at the central cancer registry level and includes a focus on misclassification of sex for breast cancer. Methods: The sex edit was applied in 3 state cancer registries: Alabama, Alaska, and Florida. Alabama applied the edit to their entire database for 1996–2004 (N = 190,614) and compared the results to external databases available to most cancer registries. Alaska applied the edit to their entire database (N = 46,645) and were able to compare the results to 2 unique, state-based databases (Alaska Permanent Fund Dividend database and State Troopers database). Florida applied the sex edit to a sample of sites (n = 953,074) with particular attention to breast cancer. Results for breast cases were compared to results from an a priori quality control project on Florida male breast cancer cases. Using the Florida data, issues specific to male breast cancer were evaluated. Results: In Alabama, 45% of 977 cases flagged as potentially miscoded sex were determined to be miscodes. In Alaska, 19% of 88 cases flagged as potentially miscoded sex were determined to be miscodes but the percent of miscoded cases identified by the edit more than doubled in the most recent years of data. For the Florida male breast cancer comparison, the sex edit correctly identified 729 of 903 cases known to be miscoded (81%) and was unable to assign a potential sex on the remaining 174 cases—but did not incorrectly flag any cases as miscodes. Implications: The sex edit is a useful tool for identifying cases that require further review to confirm the reported sex code is correct. However, it only assesses 69%–84% of cases based on name and, of those flagged, only 19%–45% are true misclassifications. But for breast cancer, a site with a skewed male to female ratio, the verified misclassification rate was 100% of the male breast cancer cases flagged as potential females. The proper application of the sex edit can improve the quality of the sex variable and can greatly reduce the impact of miscoded sex on gender-skewed sites like male breast cancer.

Key words: *automated edit, bias, breast cancer, cancer registries, data quality*

Introduction

Missing data in cancer registries is due to either the absence of the data in the clinical assessment, such as an unstaged case due to contraindicated comorbidities, or the failure of the surveillance system to capture the information.[1] In the case of a demographic variable like sex, it is rarely missing but may be miscoded due to a clerical error either during patient intake or when the certified tumor registrar (CTR) abstracts the case. Although there is little documentation of the impact of missing or miscoded data on research results,[1] it is likely that miscoded data, even due to a random clerical error versus the more problematic systemic error that can occur with software problems, can have profound impacts on research results.

For instance, a CTR will generally abstract a comparable number of male and female cases. A miscoded sex for a sex-specific tumor, like cervical or prostate, can be easily identified using an automated edit, but a clerical error on the sex code for other cancer sites may never be identified once the case is sent to the cancer registry. If the miscode rate is low, say 1 in 500, and the miscode is a random clerical error and not a systemic error, there will be essentially no impact on rates, except in the cases of cancers with a skewed male to female ratio.

Breast cancer, for instance, is 0.5%–1% male with about 2,000 male breast cancer cases diagnosed annually in the United States.[2] With approximately 200,000 new female breast cancers a year and 2,000 male, and a theoretical random miscode rate of 1 in 500, 400 women would be errantly coded as men but only 4 men would be miscoded as women. Male breast cancer cases would actually be 2,396 and the female cases would be 199,602. That is a 20% increase in the crude male breast cancer rate but only a 0.2% decrease in the crude female breast cancer rate.

Such miscodes can produce invalid rates as well as bias research results. But visual review and follow-back to improve the quality of the sex coding is labor intensive and so is typically only performed as a special project on subsets

Address correspondence to Recinda L. Sherman, MPH, CTR, North American Association of Central Cancer Registries, 2121 West White Oaks Drive, Suite B, Springfield, IL 62704-7412. Telephone: (217) 698-0800 ext. 6. Fax: (217) 698-0188. Email: rsherman@naaccr.org.

[a]North American Association of Central Cancer Registries. [b]New York State Cancer Registry. [c]Alaska State Cancer Registry. [d]Alabama State Cancer Registry. [e]Temple University, Department of Geography. [f]New York State Department of Health. [g]University of Miami.

of data. Most registries use intrarecord edits on site–sex combinations as a standard tool to identify errors in the sex code in cancer registries, which only apply to sex-specific sites. But the New York State Cancer Registry (NYSCR) created an edit for identifying potential sex miscodes for all sites and has made its components available for use through the North American Association of Central Cancer Registries (NAACCR).[3] The edit uses the most popular male and female first names based on decade of birth to identify potentially miscoded cases. For specific names that are gender-specific to the opposite gender in the United States compared to other countries (eg, Jean, Carmen, Andrea, Angel), the edit is not used if the person was born outside the United States.

Florida was one of the first states to apply the NYSCR sex edit to their registry. The Florida registry is one of the largest cancer registries in the world, with more than 115,000 incident cancer cases collected annually. The Florida registry relies heavily on automated processes to ensure data integrity including automated site–sex edits. Manual quality control projects to improve data accuracy are considered impractical for standard registry operations in Florida.

In 2002, researchers in Florida were concerned that the rates of breast cancer among Floridian men were higher than for men nationally. The data indicated that breast cancer incidence rates were increasing at a faster, statistically significant, rate in Florida males compared to the SEER-9* males.[4] Studying such a high-risk population could be important in advancing etiologic knowledge about the disease. However, before drawing research conclusions, potential spuriousness of results that can occur due to underlying data errors, such as miscoded sex, should be evaluated.

This paper describes the application of the NYSCR automated sex edit to improve the coding of sex in registry data using 3 example states: Alaska, Alabama, and Florida. Extensive detail is given to the issue of sex miscodes of breast cancer cases in Florida.

Methods

New York State Sex Edit

The sex edit was developed by the NYSCR.[3] It evaluates names that are highly correlated with gender and flags suspicious name/sex combinations. Many names have been gender-specific for centuries (eg, Elizabeth and Charles), but occasionally the gender associations of names change over time (eg, Rosario was a typical male name in 1900 but became typically female in 1940 forward). The edit uses the Social Security Administration database of the 1,000 most popular male and female names for each decade from 1890–2008. Names with at least a 49:1 ratio of one sex to the other were branded as sex-specific. This list of sex-specific names is matched against the names in the cancer registry and potentially miscoded name–sex combinations can be identified for review.

Multi-State Assessment

The sex edit was evaluated for use in 3 uncoordinated efforts in the Alabama, Alaska, and Florida central cancer registries as part of state-specific quality control (QC) registry operations.

The Alabama assessment tested the edit against the database for cases diagnosed from 1996–2004. If a case was flagged as a potentially miscoded sex, descriptive text from all the original source records was first reviewed. If the text review was inconclusive, the prefix, suffix and spouse name fields were also reviewed for confirmation of sex. If no determination of sex could be made, the patient's vital status was checked. If the patient was alive, sex was confirmed when possible based on an external, prior linkage with state Medicaid data. If the patient was deceased, sex was confirmed using state and national death files. If the sex code still remained unresolved, the code was determined based on staff judgment using primary site and name.

The Alaska assessment tested the edit against the entire database from 1996–2009. If a case was determined to be a potentially miscoded sex, descriptive text from all the original source records was first reviewed. If the text review was inconclusive, the patient's sex was manually looked up in the Alaska Permanent Fund Dividend (PFD) database. The PFD database contains demographic information, updated annually, on approximately 95% of Alaska residents who submit an application to share the interest on royalties paid by the petroleum industry to the Alaska state government for transporting oil thorough the Trans-Alaska Pipeline. If the PFD review was inconclusive, then the record was reviewed using the Alaska Department of Public Safety's State Troopers database. The Troopers database includes demographic and driver's license–related information for individuals with a driver's license or other state issued identification, who have been fingerprinted, or who have had contact with state or local law enforcement.

The Florida assessment tested the edit against breast, thyroid, liver, and colorectal cancers diagnosed from 1981–2008. The Florida assessment also compared the results of the sex edit with the results of a quality control (QC) project conducted in 2003 on the accuracy of reported sex of male breast cancer cases. The QC project was prompted by concerns regarding an increase in male breast cancer among Floridian men, and it assessed male breast cancer cases diagnosed in Florida from 1981–2000. For the QC project, cases determined by visual review to be female names were followed back with a letter to the hospital to confirm the sex of the patient.

Results

Alabama

In Alabama, 190,164 cases were evaluated; about 0.5% (977) were flagged as potentially miscoded sex and 44%

*SEER (Surveillance, Epidemiology, and End Results) registries are population-based cancer registries funded by the National Cancer Institute. SEER data was used as the comparison to more closely match the available years of diagnosis in Florida. NPCR (National Program of Cancer Registries) funded by the Centers for Disease Control and Prevention covers the majority of the nation but start with diagnosis year 1995. The SEER-9 registries are Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco–Oakland, Seattle–Puget Sound, and Utah.

(429) were confirmed as miscodes and corrected in the registry data. Sixty-six percent of potentially miscoded cases were reported as female. Of the 429 changed, registry personnel could find corroborating evidence for the change for 283 (66%) and the remainder were changed based on visual review by the registry personnel. Of the 548 cases where sex remained unchanged, registry personnel could find corroborating evidence in support of the reported sex for 388 (71%). The remainder could not be confirmed, so the codes were left unchanged and the cases will eventually be re-reviewed with additional years of death data. There were 332 of 738 (45%) non-Hispanic whites confirmed as having miscoded sex, but only 4 of 14 (29%) Hispanics and 81 of 204 (40%) non-Hispanic blacks were confirmed as having miscoded sex. These groups are more likely to have unique names or names that have a less common gender affiliation, such as Angel being a common name for Hispanic males but a common name for non-Hispanic white females. There were an additional 4 of 4 (100%) cases of unknown sex that were updated with a known sex after being flagged with a potentially miscoded sex. The number of cases for which a potential sex could not be determined by the edit was not recorded.

Alaska

Of the 46,645 consolidated cancer cases in the Alaska databases, 16% (7,303) could not be assigned a potential sex by the edit because their first names were not gender-specific or were not common enough to be ranked. There were 88 cases that were flagged as a potentially miscoded sex and underwent manual review. During manual review, it was determined that several names were either misspelled or were nicknames and were corrected to their formal first name. The corrected names were appropriate for the sex, such as Louis vs Lois and Marty vs Martha. Using either the accompanying text in the source abstracts, the PFD database or Troopers database, sex was confirmed for all 88 cases. Of the 88 cases flagged with potentially miscoded sex, 19% (17) were confirmed to be misclassified and their coded sex was corrected in the registry data. An assessment by year indicated that the percent of cases truly misclassified by sex is higher for more recent years with 31% of the potentially miscoded sex cases identified as true miscodes for cases diagnosed in the last 2 years, 2008–2009. Increasing demands on the CTR may be resulting in increased clerical error, but it is likely that a small registry like Alaska is able to identify most miscoded sex cases through visual review and use of the data over time.

Florida

A data quality project was undertaken in Florida to evaluate the sex coding of breast among males. The first name of male breast cancer patients diagnosed from 1981–2000 were visually reviewed. A total of 904 of approximately 3,800 male cases of breast cancer were identified as potentially female based on first name. All but 3 were confirmed female by the hospitals, and the sex code was corrected in the registry data.

Figure 1. Number of Reported Male Breast Cancer Cases Identified as Miscodes Over Time in Florida

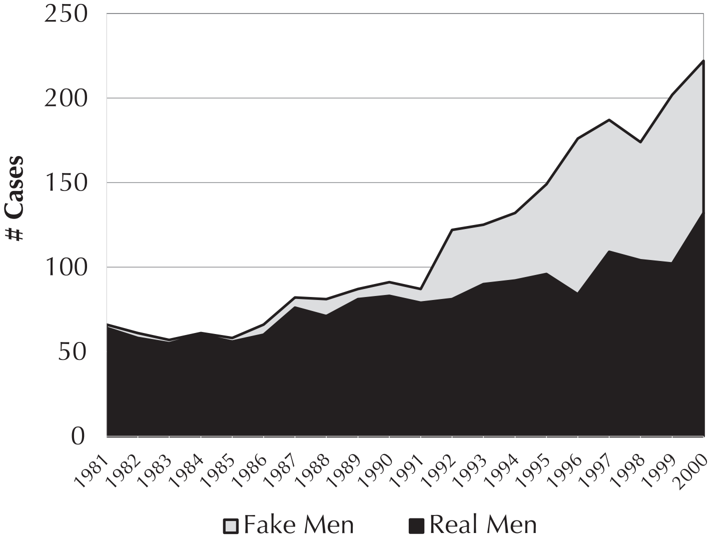


Figure 1 illustrates the number of breast cases that were misclassified as men ("fake men") by year of diagnosis from 1981–2000. It is clear that sex misclassification for breast is more problematic with later diagnosis years. This is likely due to changes in International Classification of Diseases for Oncology (ICD-O) coding. Prior to the 1990s, the ICD-O classification system was similar to the International Classification of Diseases (ICD)-9 classification with separate codes for female (174) and male (175) breast cancers. Starting with ICD-O-2 in the early 1990s, breast cancer became a single code (C50) regardless of sex. This level of misclassification can significantly inflate breast cancer rates in males because it is a rare cancer while only negligibly altering rates in females.

The sex edit was tested against the original 904 cases manually followed back to hospitals in the 2003 QC project. The edit correctly identified 729 (81%) of the "fake men" as female plus 1 of the 3 "real men" breast cancer cases as male. The remaining cases were not assessable because the name was not gender-specific. Although the edit could not determine a potential sex code for 175 of the cases, the edit did not misclassify the sex of any of the male breast cancer cases.

Most (648,769, or 68%) of the 953,074 cases in the site-specific evaluation of the sex edit agreed with the edit's potential sex. About 31% were indeterminate: 298,888 had non–gender-specific first names, 68 had a missing year of

Table 1. Percent Identified as Potential Miscoded Sex and Percent Not Assessed by Edit, in Florida

	% Potentially Miscoded		*% Not Determined*
	Reported as:		
Site	Male	Female	
Breast	21.00%	0.20%	31%
Thyroid	1.30%	0.40%	29%
Liver	0.30%	1.10%	29%
Colorectal	0.50%	0.60%	33%

birth, 595 had a reported year of birth born prior to 1890, and 90 cases were coded in the registry as hermaphrodite or transgender. There were 4,519 (0.5%) cases that were identified as potential sex miscodes. Additionally, 145 cases were coded as unknown sex in the registry but the edit identified a potential sex.

Results varied by site (Table 1). Over a fifth, 21%, of breast cancer patients reported as male were identified as potentially miscoded sex (0.2% for breast cancer cases reported as female). Breast cancer is about 100 times more common in women than men, so the count of potential miscodes for each sex were close; 1,076 cases reported as male were identified as miscodes and 984 reported as female were identified as miscodes. For thyroid, a site 3 times more common in women than in men, 1.3% of the thyroid cases reported as male were identified as potential miscodes (vs 0.4% for thyroid cases reported as females). For liver, a site 3–8 times more common in men than women, 1.1% of the liver cases among females and 0.3% among males were identified as a potential miscoded sex. For these sex-skewed sites, the sex ratio of cases identified as potential miscoded sex is exactly inversely proportional to the sex ratio of the cancers themselves. For colorectal cancer, a site with similar rates for both sexes, the percent of cases identified with a potential miscoded sex was similar for cases reported as men (0.5%) and as women (0.6%).

The utility of the edit was higher for non-Hispanic whites than other race/ethnicity categories (Table 2). As with Alabama, a greater proportion of Hispanics and non-whites were not identified with a potential sex by the edit—meaning the date of birth was missing, the decade of birth was prior to 1890 (72 cases), or the name was not gender-specific or not popular enough to be ranked in the top 1,000 most common names by decade (55,393 cases).

Table 2. Percent of Cases With Potential Sex Not Determined by Edit by Race, Ethnicity in Florida

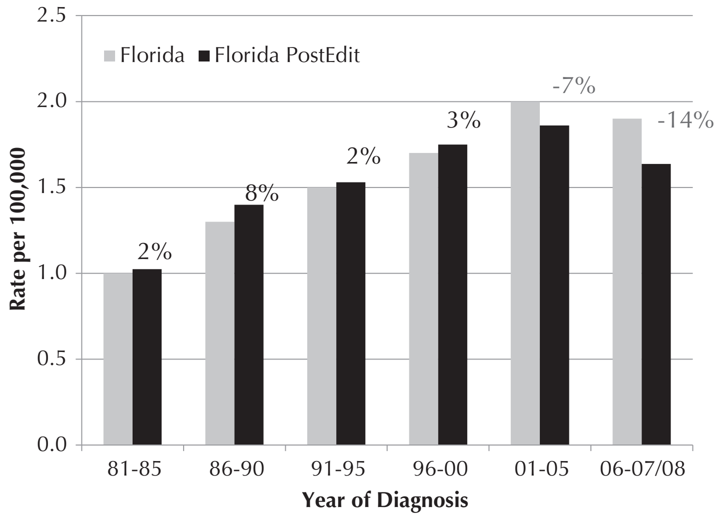
Race	*% Not Determined*
White	31%
Black	34%
American Indian	37%
All Others	55%
Ethnicity	
Hispanic	46%
Non-Hispanic	30%

Discussion

It is clear the sex edit can be used to improve the quality of sex coding in cancer registry data. The extent of misclassification of sex is low as evidenced by the results from the 3 registries. But even a few cases of miscoded sex a year can potentially impact rates of rare cancers or in small-area analyses.

Alaska is one of the least populated states with one of the smallest registries, making it reasonable to conduct manual follow-back. The difference of 19% true misclassification in the early years compared to 31% in the more recent years may reflect the on-going QC efforts of registry staff in Alaska. So even in states like Alaska, the sex edit can be effectively applied proactively to address misclassification of sex. In all states, the sex edit can reduce the extent of manual follow-back, which is a significant barrier to QC efforts in many registries.

Figure 2. Male Breast Cancer Rates: Florida Pre- and Post-Edit



Using the sex edit can reduce the impact of miscoded sex on male breast cancer rates, as we can demonstrate using the Florida data. Specifically, the results of the comparison of sex edit to the QC project indicates the edit is reliable enough to reclassify all reported male breast cancer cases indicated as potential miscodes to female, a total of 1,076 cases. Ideally, we would like to conduct manual follow-back of the Florida data to determine which potential miscodes are truly miscoded. But Florida is a large registry and relies heavily on automated algorithms and edit checks with very little follow-back, unlike Alaska. If we apply the 45% correct miscoded proportion from Alabama for the potentially miscoded females, we would recode 440 of the female breast cancer cases as male. So we would move 1 of 78 previously unknown sex cases to male based on the potential sex identified by the edit. When we compare the results in Figure 2, we see that the rates of male breast cancer actually increase for diagnosis years 1981–2000 and decrease for more recent diagnosis years in Florida. This is because the historical QC project only resolved miscoded breast cases reported as male and did not review cases that were reported as female for any potential miscodes. Focusing on improbable sex for male breast cancers only removes female cases miscoded as male, which falsely suppresses the rate because no misclassified female breast cancer cases are added back into the male category. For more recent years, no manual resolution of male breast cancer patients was conducted so the impact of recoding sex based on the edit was an overall decrease in male breast cancer rates in more current years.

One limitation of the edit is an inability to determine a potential sex for many cases due to the lack of assessment for unusual names, for cases lacking date of birth, and for decades for which the data is not incorporated into the edit. There is a current NAACCR effort to incorporate more recent years of data (2009 year of birth forward) to update the edit. But the edit will continue to be less effective for minorities with names that are less likely to be popular. In addition, first generation males may not be accurately determined or not determined at all. For instance, Andrea, Angel, Carmen, and Jean are common names for females in much of the United States but are male names among Hispanics, Haitians, and Italians. These names are excluded from the edit for patients who are foreign-born. But they had to be removed completely from the edit when applied in Florida based on a preliminary review of the results. Similar adjustments might need to be made that could be informed based on other state's demographic profile.

Also, a registry may be tempted to automatically change the sex of all cases identified as potentially miscoded rather than committing the fiscal resources and personnel needed for follow-back or confirmation from secondary sources. However, the edit was intended to be used to flag cases for further follow-back only. The results from Alabama and Alaska indicate that, overall, the sex edit correctly flags a true miscoded sex as a potentially miscoded sex less than 50% of the time. If the sex edit is implemented at the central registry level, sex must be confirmed through an external source—not automatically updated. However, for cases where sex is unknown, registries that are unable to perform manual review, either due to large size, like Florida, or lack of access to useful outside data, might consider assigning the potential sex identified by the edit. This should be documented so that if additional information is reported to the registry it is assessed appropriately.

Breast cancer, however, is a special circumstance. When applying the sex edit, the percentage of cases flagged as a potential miscoded sex that are truly miscoded is significantly higher than sites with more similar male to female rates. In fact, the edit did not have any cases falsely identified as a miscode. Although we have no "gold standard" to use in a formal calculation, we can consider the edit to have 100% sensitivity but a modest specificity for male breast cancer. It may be efficient for larger registries, for which manual review is impossible on all cases, to automatically recode male breast cancer cases flagged cases as female without confirmation from secondary sources. However, if the registry only resolves breast cases that are potentially miscoded as male but not the reverse (reported female cases identified as potentially male), the registry will be falsely suppressing male breast cancer rates.

Conclusions

Overall, the extent of miscoding of sex appears minimal in cancer registries, less than 1%. But miscoding disproportionately affects sex-skewed sites, like liver, thyroid, and breast. The problem is highlighted in male breast cancer and artificially inflates male breast cancer rates to the point that can cause unwarranted alarm, as occurred in Florida, or might misdirect public health resources. However, subset-specific quality control projects on male breast cancer alone artificially suppresses rates because such projects only remove miscoded males and do not add in miscoded females. The use of the NYSCR sex edit can improve quality of sex codes by significantly reducing the number of cases requiring manual follow-back.

Acknowledgements

We would like to acknowledge the contributions of Jackie Button, Megsys Herna, and Stuart Herna for their work on the original QC project while at the Florida Cancer Data System.

References

1. Klassen AC, et al. Missing stage and grade in Maryland prostate cancer surveillance data, 1992-1997. *Am J Prev Med.* 2006. 30(2 suppl):S77-S87.
2. Ruddy KJ and Winer EP. Male breast cancer: risk factors, biology, diagnosis, treatment, and survivorship. *Ann Oncol.* 2013;24(6):1434-1443.
3. Soloway LE, B.F., Kahn AR. A New Edit for Identifying Potential Gender Misclassification in Central Cancer Registry Databases. Paper presented at: North American Association of Central Cancer Registries Annual Conference; June 2010; Quebec City, Quebec, Canada.
4. Hodgson NC, Button JH, Franceschi D, Moffat FL, Livingstone AS. Male breast cancer: is the incidence increasing? *Ann Surg Oncol.* 2004;11(8):751-755.