Published in final edited form as:
J Registry Manag. 2008 January 1; 35(4): 145–148.

Evaluating the Completeness of the SEER-Medicare Linked Database for Oral and Pharyngeal Cancer

Jonathan D. Mahnken, PhD,
Department of Biostatistics and Kansas University Cancer Center, University of Kansas Medical Center, MSN 1026, 3901 Rainbow Blvd, Kansas City, KS 66160, Phone: 913.588.2696, Fax: 913.588.0252

John D. Keighley, PhD,
Department of Biostatistics, University of Kansas Medical Center, MSN 1026, 3901 Rainbow Blvd, Kansas City, KS 66160, Phone: 913.588.2792, Fax: 913.588.0252

Christopher G. Cumming, PhD, DMD,
KU Dental Associates, 4720 Rainbow Blvd, Westwood, KS 66205, Phone: 913.588.9200, Fax: 913.588.9203

Douglas A. Girod, MD, and
Department of Otolaryngology and Kansas University Cancer Center, University of Kansas Medical Center, MSN 3010, 3901 Rainbow Blvd, Kansas City, KS 66160, Phone: 913.588.6719, Fax: 913.588.4676

Matthew S. Mayo, PhD, MBA
Department of Biostatistics and Kansas University Cancer Center, University of Kansas Medical Center, MSN 1026, 3901 Rainbow Blvd, Kansas City, KS 66160, Phone: 913.588.4703, Fax: 913.588.0252

Jonathan D. Mahnken: jmahnken@kumc.edu; John D. Keighley: jkeighle@kumc.edu; Douglas A. Girod: dgirod@kumc.edu; Matthew S. Mayo: mmayo@kumc.edu

Abstract

Oral and pharyngeal cancer is a persistent oral health problem. Baseline and trend data to measure progress are lacking. Our long-term goal is to create an algorithm using Medicare claims to identify oral and pharyngeal cancer cases among those ages 65 and older. The goal of this project was to assess the completeness of the SEER-Medicare linked database for identifying incident oral and pharyngeal cancer cases. We compared incidence rates from the “gold-standard” SEER limited-use database to those from the SEER-Medicare linked database using a quasi-likelihood extension of Poisson regression, allowing for over-dispersion. Adjustment for age, sex, race and ethnicity, and interaction terms between these explanatory variables with data source were used to assess the completeness of the SEER-Medicare linked database among these subgroups. Approximately 6.4% of the cases were missing from the SEER-Medicare linked database. The completeness varied by race and ethnicity (p=0.066). Future development of an algorithm to identify oral and pharyngeal cancer cases using Medicare claims alone can potentially identify over 93% of the cases; however, Hispanic, non-Hispanic black, and non-Hispanic other race and ethnicity subgroups will be less likely than non-Hispanic whites to be identified in such future algorithms.

Correspondence to: Jonathan D. Mahnken, jmahnken@kumc.edu.

Keywords

secondary data analysis; health outcomes research; Poisson regression; quasi-likelihood

INTRODUCTION

Approximately 30,000 people are diagnosed with oral and pharyngeal (OP) cancers every year in the U.S., and nearly one-fourth of that number die from the disease in the U.S. annually.[1–3] It is ranked the sixth most common type of cancer in the developed world.3 OP cancer impacts heavily among the elderly.2 Nearly half of incident OP cancers are diagnosed in individuals ages 65 and older, and the median age at diagnosis is 64 years.[4] Thus, as the U.S. population ages, the burden of OP cancer will likely follow the trend of other cancers—worsening over time.

While incidence rates for OP cancers have modestly decreased over the past 20 years,[3] prognosis remains poor.[2] During this same period, the 5-year survival rate has remained between 50% and 60%.[3] The prognosis is even worse for some subgroups; for example, the 5-year survival rate among black or African Americans is only 34%.[2] Other disparities are also seen in the U.S. among rural populations, individuals with disabilities, the homeless, and the frail elderly, among others.[2]

OP cancers disproportionately affect the elderly, so as the elderly population continues to grow the need for more research to reduce the burden of OP cancer among this population becomes more pressing. Unfortunately, the elderly population has been underrepresented in cancer treatment trials.[4–7] Consequently, little is known about potential drug-drug interactions or the impact of the common comorbid conditions with cancer treatment among the elderly,[4, 5, 7, 8] underscoring the need for further research in the elderly population from all races and ethnicities.

The Department of Health and Human Services, in their report of the Surgeon General on oral health in America, noted that national, state, and local surveillance databases regarding oral and craniofacial diseases, health services, and utilization of care are limited or are lacking.[2] They further stated that, “reliable and valid measures of oral health outcomes do not exist and need to be developed, validated and incorporated into practice and programs.”[2] This sentiment is also consistent with the NIDCR Strategic Plan 2003–08.[1]

The Institute of Medicine advocates linkages between population-based registries with administrative databases.[9] Weir et al.[10] discussed the need for such treatment and comorbidity information to supplement information used to direct interventions aimed at improving cancer outcomes, particularly among the underserved. Such linkages will facilitate investigations into the patterns and the quality of care received[10–12] as cases exist where individuals with cancer have not received treatment known to be effective.[13, 14] Results could then be disseminated at the national, state, and local levels.

The long-term, future goals of this research are to develop, validate, and utilize an algorithm based on Medicare claims that can identify incident cases of oral and pharyngeal cancer in the U.S. population ages 65 and older. Use of this algorithm could provide a vital link by enabling the study of health trajectories and facilitating the development of quality control measures that could be monitored efficiently over time. Though data exist (such as the SEER-Medicare linked database described below) that enable such assessments, the ability to draw cases from the entire Medicare population—rather than only those in both Medicare and a SEER Program (described below) registry—could provide a more general population base of study. Such

coverage is especially relevant for OP cancer since its burden is more pronounced in rural parts of the country, while the SEER coverage areas are more urban than the general U.S. population (among other differences).[13,15] This manuscript describes the first step in this process, the evaluation of the completeness of the SEER-Medicare linked database compared to the "gold-standard" SEER limited-use database. The SEER-Medicare linked database is the data source that will later be used to derive and validate an algorithm to identify incident cases from Medicare claims.

METHODS

This project was approved by the University of Kansas Medical Center Human Subjects Committee (HSC #10914).

SEER Limited-use Database

The Surveillance, Epidemiology and End Results (SEER) Program provides information from population based tumor registries in the U.S. SEER collects and provides cancer incidence and survival statistics for approximately 26% of the population.[15] In addition, the SEER Program releases data in electronic format called the SEER limited-use database based on a subset of their registries. SEER data are frequently validated, ensuring the information reported is of high quality. SEER's standard for case ascertainment stands at 98%, and is "considered the standard for quality among cancer registries around the world."15 Information includes demographic and tumor characteristics, as well as follow-up information from the date of diagnosis.

In addition to information on cancer cases, the SEER Program also made available population data sets (herein referred to as the SEER population database). These data are a modification to the annual county population estimates produced by the Bureau of the Census.[15] Estimates are produced, down to the county level, by age group, sex, and race and ethnicity.

SEER-Medicare Linked Database

The Center for Medicare and Medicaid Services has linked SEER tumor registry data with Medicare claims and census information to create the SEER-Medicare linked database. These data contain information on individuals from the SEER Program who also appeared in Medicare enrollment records. Among the SEER cases diagnosed with cancer at age 65 or older, 94% were identified—across cancer sites—in the Medicare enrollment records.

Medicare provides health insurance for about 97% of the population ages 65 and older in the U.S.[16] It provides benefits covering inpatient hospitalizations, skilled nursing facilities, and home health and hospice care. These claims data will be used in our subsequent work, which will derive an algorithm based on these claims to identify oral and pharyngeal cancer cases.

Medicare also collects information about beneficiaries.[16] The SEER-Medicare linked data containing individual demographic as well as tumor characteristics for the cancer cases from SEER are in the Patient Entitlement and Diagnosis Summary File, or PEDSF. The PEDSF includes patients in the SEER-Medicare linked database whose cancer diagnosis took place from 1973 to 2002, though some registries were not included for all of these years.

Case Ascertainment

OP cancers were identified using the Site Recode variables. Values of 20010(10)20100 and 1 (1)10 in the SEER limited-use and SEER-Medicare linked databases, respectively, were studied. These Site Recode values related to ICD-O-2 codes as shown in Table 1. We used these variables to identify the OP cancers to be included in the numerators for the incidence

rates from both the SEER limited-use and PEDSF files. We retained the first primary diagnosis of OP cancer in both data sets of subjects that were at 65 years of age or older at diagnosis, whose diagnosis occurred during the years 1992–2002 in the Connecticut, Hawaii, Iowa, New Mexico, Utah, Detroit, San Francisco-Oakland, Atlanta, Seattle-Puget Sound, San Jose, Los Angeles, or rural Georgia registries. There were a total of 19,593 subjects in the SEER limited-use files and 18,313 in the PEDSF files that met these criteria.

Subgroup Measures

In addition to the overall completeness of the PEDSF, we compared completeness among different demographic subgroups. These subgroups were limited in that only those subgroups indicated in the SEER limited-use files, the SEER population files, and the PEDSF could be assessed. These subgroups were age at diagnosis (65–69, 70–74, 75–79, 80–84, 85 and older; population age group for the SEER population files), sex, and race and ethnicity (Hispanic, non-Hispanic black, non-Hispanic white, and other).

Statistical Analysis

To evaluate the completeness of the SEER-Medicare linked database in measuring OP cancer incidence among the U.S. population ages 65 and older, we compared incidence rates from both the SEER limited-use and PEDSF files. The SEER population files were used as the common denominator for these rates. The years 1992–2002 were used for this analysis. Poisson regression was used to model these rates and compare the incidence rates with an incidence rate ratio (IRR). The response variable was the count of cases, and the natural log of the population count from the SEER population files was used as an offset variable. The model was assessed for over- or under-dispersion, and the scale parameter was estimated using the deviance parameter to model this dispersion parameter.17 Backwards elimination was used to build a model from the set of candidate explanatory variables: data source; age group; sex; race and ethnicity; year of diagnosis; and two-way interactions with data source. A global test of all candidate explanatory variables simultaneously was conducted.18 Next, explanatory factors with large p-values ($p>0.05$ for main effects and $p>0.1$ for interactions) from the F-tests were removed from the model on a step-by-step basis. (The F-test was used instead of the chi-square test because of the inclusion of the dispersion parameter.) The fit of the model was assessed via the scaled Pearson chi-square test for lack of fit. Finally, one-sided, lower 95% confidence intervals incorporating the dispersion parameter19 for the IRRs were generated. One-sided intervals were used because, by nature of the data, only subjects included in the SEER limited-use files could be included in the PEDSF (ie, the PEDSF is a subset of the SEER limited-use files). Therefore, the maximum value for the IRR comparing these two groups was *1.00*, so the type I error for this interval was placed entirely on the lower bound. Prior to conducting the analysis, we identified a rule for declaring equivalence between the incidence rates from these two sources as the lower bound of the one-sided 95% confidence interval for the IRR being greater than or equal to 0.9, indicating that nearly all of the OP cancer cases were present in the PEDSF. IRRs for subgroup variables whose interaction terms with data source remained in the model used this same 0.9 threshold with a Bonferroni correction for the lower 95% confidence limits.

All data management and data analysis were conducted using SAS version 9.1. The Poisson regression analysis was conducted using PROC GENMOD in SAS.

RESULTS

Crude OP cancer incidence rates (/100,000) from the SEER-Medicare linked database and the SEER limited-use files were 39.7 and 42.4, respectively. The crude IRR (one-sided 95% confidence interval) was 0.94 (0.89–*1.00*). Stratified by race and ethnicity groups, these rates

were: 23.5, 34.5, 43.5, and 25.9 for Hispanic, non-Hispanic black, white and other subgroups from SEER-Medicare, respectively; and in the SEER limited-use files these rates were 27.3, 38.4, 45.9, and 30.0, respectively. These findings are presented in Fig. 1.

The global test of no effect among all candidate explanatory variables was rejected ($p<0.0001$). Backwards elimination produced a reduced model that contained main effects for data source (SEER-Medicare linked versus SEER limited-use database), age group, sex, race and ethnicity, and year of diagnosis ($p<0.0001$ in each case), plus a data-source-by-race and ethnicity interaction term ($p=0.066$). All interaction terms removed by backwards elimination had $p>0.87$. Significant over-dispersion was detected from these data, indicating that the standard errors should be adjusted upwards by approximately 17% (ie, the scale parameter was estimated to be 1.17 in the adjusted model). The scaled Pearson statistic failed to detect a lack of fit ($p=0.71$) in this multivariable model that was adjusted for over-dispersion. Adjusted OP incidence rate ratios (SEER-Medicare linked database versus SEER limited-use database) and 95% one-sided confidence intervals (adjusted using the Bonferroni correction) for the Hispanic, non-Hispanic black, white, and other subgroups were 0.86 (0.74–*1.00*), 0.90 (0.79–*1.00*), 0.95 (0.92–*1.00*), and 0.86 (0.76–*1.00*), respectively.

DISCUSSION

Our results indicated that over 90% of the incident OP cancer cases among those 65 years of age and older could be identified through an algorithm using Medicare claims. The overall IRR was above the pre-specified level of 0.9 to signify equivalence, but the lower bound of the one-sided confidence interval fell slightly below this threshold (0.89–*1.00*); thus, we concluded that Medicare claims may provide a means to study OP cancer cases among the elderly, but will miss a significant portion of the cases. We found that this undercount was not evenly distributed by race and ethnicity. The adjusted IRR was highest among the non-Hispanic whites (0.95), and the lower bound of the confidence interval for this IRR was 0.92, which was above our 0.9 equivalence threshold. In other words, approximately 95% of these cases in SEER linked to Medicare, whereas only about 86–90% linked from other racial and ethnic groups that were identified as cases by SEER. The reasons for this variation seem most likely attributable to differences in Medicare coverage among these subgroups. These results also demonstrated the completeness of the SEER Program in identifying cases even among those lacking Medicare coverage—some likely having no insurance coverage at all—in individuals ages 65 and older.

Conclusion

Though some cases of OP cancer among the U.S. population ages 65 and older would be missed, the development of an algorithm to identify these individuals using only Medicare claims could prove useful for future population-based research of OP cancer. Specifically, nearly the entire population ages 65 years and older could be examined—including people from all of the rural areas within the U.S. Further, the Medicare claims could also provide data with treatment history and comorbidity information for this large, broadly generalizable population. This could prove advantageous for future studies of OP cancer, which has more heavily burdened rural communities.

Acknowledgments

The research was supported by Grant Number R03DE016958 from the National Institute of Dental & Craniofacial Research, the Kansas Masonic Foundation (Kansas Masonic Cancer Research Institute Pilot Award, 7/1/2006), and Grant Number R24CA95835 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Dental & Craniofacial Research or the National Institutes of Health. The authors acknowledge the efforts of the Applied Research Program, NCI; the

Office of Research, Development and Information, CMS; Information Management Services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database. The authors also used data from Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Public-Use Data (1973–2003), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2006, based on the November 2005 submission. The authors would also like to thank Dr. Joan Warren and the Reviewers for their helpful and insightful comments, which served to strengthen this work.

References

1. NIDCR Strategic Plan FY 2003–08. National Institutes of Dental and Craniofacial Research; Sep 27. 2004
2. Oral Health in America: A Report of the Surgeon General--Executive Summary. Rockville, MD: U.S. Department of Health and Human Services, National Institutes of Dental and Craniofacial Research, National Institutes of Health; 2000.
3. Reis, LAG.; Eisner, MP.; Kosary, CL.; Hankey, BF.; Miller, BA.; Clegg, L.; Mariotto, A.; Feuer, EJ.; Edwards, BK., editors. SEER Cancer Statistics Review, 1975–2001. Bethesda, MD: National Cancer Institute; 2004.
4. Edwards BK, Howe HL, Ries LAG, Thun MJ, Rosenberg HM, Yancik R, Wingo PA, Ahmedin J, Feigal EG. Annual report to the Nation on the status of cancer, 1973–1999, featuring implications of age and aging on U.S. cancer burden. Cancer 2002;94:2766–2792. [PubMed: 12173348]
5. Balducci L, Extermann M. Management of cancer in the older person: a practical approach. Oncologist 2000;5:224–237. [PubMed: 10884501]
6. Trimble EL, Carter CL, Cain D, Freidlin B, Ungerleider RS, Friedman MA. Representation of older patients in cancer treatment trials. Cancer 1994;74:2208–2214. [PubMed: 8087794]
7. Hutchins LF, Unger JM, Crowley JJ, Coltman CA, Albain KS. Under representation of patients 65 years of age or older in cancer treatment trials. New England Journal of Medicine 1999;341:2061–2067. [PubMed: 10615079]
8. Yancik R, Ganz PA, Varricchio CG, Conley B. Perspectives on comorbidity and cancer in older patients: approaches to expand the knowledge base. Journal of Clinical Oncology 2001;19:1147–1151. [PubMed: 11181680]
9. Institute of Medicine. Enhancing data systems to improve the quality of cancer care. Washington D.C: National Academy Press; 2000.
10. Weir HK, Thun MJ, Hankey BF, Ries LAG, Howe HL, Wingo PA, Jemal A, Ward E, Anderson RN, Edwards BK. Annual report to the Nation on the status of cancer, 1975–2000, featuring the uses of surveillance data for cancer prevention. Journal of the National Cancer Institute 2003;95:1276–1299. [PubMed: 12953083]
11. Du XL, Key CR, Osborne C, Mahnken JD, Goodwin JS. Discrepancy between consensus recommendations and actual community use of adjuvant chemotherapy in women with breast cancer. Annals of Internal Medicine 2003;138:90–98. [PubMed: 12529090]
12. Ayanian JZ, Zaslavsky AM, Fuchs CS, Guadagnoli E, Creech CM, Cress RD, et al. Use of adjuvant chemotherapy and radiation therapy for colorectal cancer in a population-based cohort. Journal of Clinical Oncology 2003;21:1293–1300. [PubMed: 12663717]
13. Jemal A, Clegg LX, Ward E, Ries LAG, Wu X, Jamison PM, Wingo PA, Howe HL, Anderson RN, Edwards BK. Annual report to the Nation on the status of cancer, 1975–2001, with a special feature regarding survival. Cancer 2004;101:3–27. [PubMed: 15221985]
14. National Cancer Policy Board IoMaCoLS, National Research Council. Ensuring Quality Cancer Care. Washington D.C: National Academy Press; 1999.
15. SEER Program. [Accessed September 14, 2004]. http://seer.cancer.gov
16. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States Elderly Population. Medical Care 2002;40:IV-3–IV-18.
17. McCullagh P. Quasi-likelihood functions. The Annals of Statistics 1983;11:59–67.
18. Harrell, FE. Regression modeling strategies: with applicatiosn to linear models, logistic regression, and survival analysis. New York, NY: Springer-Verlag New York, Inc; 2001.

19. Agresti, A. Categorical Data Analysis. 2. Hoboken, NJ: John Wiley & Sons, Inc; 2002.

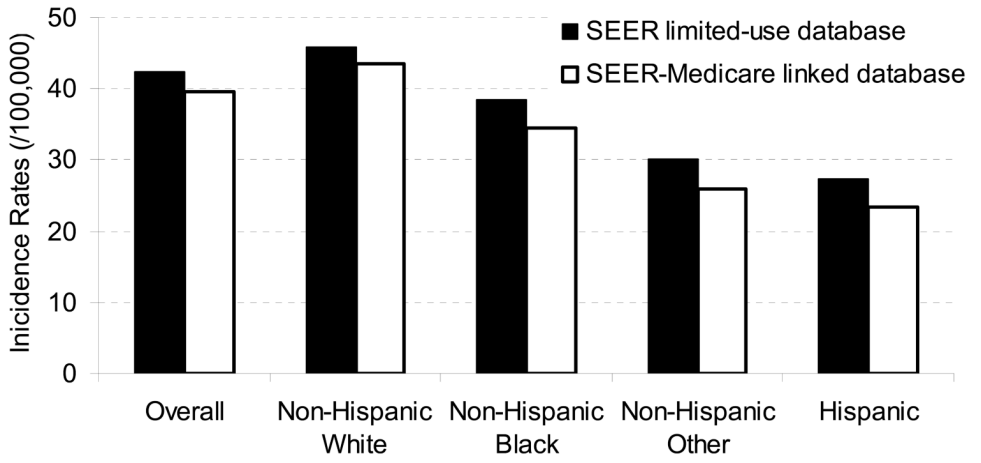

Figure 1. Comparison of crude incidence rates from SEER-Medicare linked database with SEER limited-use database*

*Completeness of the SEER-Medicare linked database varied by race in the Poisson regression model (p=0.066)

Table 1

Relationship between ICD-O-2 values and Site Recode variables from SEER limited-use database (SEER) and SEER-Medicare linked database (PEDSF)

		Site Recode	
Site	**ICD-O-2 Codes***	**SEER**	**PEDSF**
Buccal cavity and pharynx			
Lip	C000:C009	20010	1
Tongue	C019:C029	20020	2
Salivary gland	C079:C089	20030	3
Floor of mouth	C040:C049	20040	4
Gum and other mouth	C030:C039,C050:C059, C060:C069	20050	5
Nasopharynx	C110:C119	20060	6
Tonsil	C090:C099	20070	7
Oropharynx	C100:C109	20080	8
Hypopharynx	C129,C130:C139	20090	9
Other buccal cavity and pharynx	C140,C142:C148	20100	10

*Excluding types 9590:9989