Published in final edited form as:
J Cancer Educ. 2012 December ; 27(4): 664–669. doi:10.1007/s13187-012-0383-7.

Understanding the Limits of Large Datasets

Catherine Sanders, OMS IV, **Sidney L Saltzstein, MD, MPH**, **Duy H Nguyen, BS**, **Helen Shi Stafford, MS IV**, **Matthew Schultzel, MS IV**, and **Georgia Robins Sadler, BSN, MBA, PhD**
Rebecca and John Moores UCSD Cancer Center, University of California San Diego, La Jolla, California (CS, SLS, DN, HS, MS, GRS); Department of Surgery/Division of General Surgery, University of California San Diego, La Jolla, California (GRS); Departments of Pathology and Family and Preventive Medicine, University of California San Diego, La Jolla, California (SLS); Ohio State University College of Medicine, Columbus, Ohio (CS); University of California San Diego School of Medicine, La Jolla, California (HS); Touro College of Osteopathic Medicine – Mare Island, California (MS); and The Sam and Rose Stein Institute for Research on Aging, University of California, San Diego (CS, HS, MS).

Abstract

BACKGROUND—This study investigated missing data in a large cancer dataset, to alert educators to the implications and limitations of missing data.

METHODS—The authors examined the California Cancer Registry for missing data by eight common cancer sites, seven sociodemographic and clinical variables, and the top three reporting sources.

RESULTS—The *gender* variable had no missing data, followed by *age* (0.1% missing), *ethnicity* (2.2%), *stage* (7.0%), *differentiation* (36.3%), and *birthplace* (42.5%). Hospitals'/clinics' reports had the lowest percentages of missing data.

CONCLUSIONS—Educators should anticipate the limitations of missing data in large datasets to prevent methodological flaws and misinterpretations of research findings.

Keywords

Cancer Registry; Reporting; Missing Data; Clinical Variables; Tumor Variables; Data Limitations

Corresponding Author Sidney L Saltzstein, MD, MPH, Moores UCSD Cancer Center, 3855 Health Sciences Drive MC 0850, La Jolla, CA 92093-0850, (858) 822-7611, (858) 534-7628 FAX, ssaltzstein@ucsd.edu.
Authors' Institutions and Contact Information:
Catherine Sanders, MS IV, The Ohio State University College of Medicine, Columbus, Ohio, Moores UCSD Cancer Center, 3855 Health Sciences Drive MC 0850, La Jolla, CA 92093-0850, (858) 822-7611, (858) 534-7628 FAX, Catherine.Sanders@osumc.edu
Sidney L Saltzstein, MD, MPH, Professor of Clinical Oncology and Pathology, Emeritus, Adjunct Professor of Family and Preventive Medicine, Moores UCSD Cancer Center, 3855 Health Sciences Drive MC 0850, La Jolla, CA 92093-0850, (858) 822-7611, (858) 534-7628 FAX, ssaltzstein@ucsd.edu
Duy H Nguyen, BS, University of California, San Diego, Moores UCSD Cancer Center, 3855 Health Sciences Drive MC 0850, La Jolla, CA 92093-0850, (858) 822-7611, (858) 534-7628 FAX, dhn004@ucsd.edu
Helen Shi Stafford, MS IV, UCSD School of Medicine, Moores UCSD Cancer Center, 3855 Health Sciences Drive MC 0850, La Jolla, CA 92093-0850, (858) 822-7611, (858) 534-7628 FAX, heshi@ucsd.edu
Matthew Schultzel, OMS IV, Touro College of Osteopathic Medicine – Mare Island, California, Moores UCSD Cancer Center, 3855 Health Sciences Drive MC 0850, La Jolla, CA 92093-0850, (858) 822-7611, (858) 534-7628 FAX, m1schult@gmail.com
Georgia Robins Sadler, MBA, PhD, Clinical Professor of Surgery, Associate Director for Community Outreach, Moores UCSD Cancer Center, 3855 Health Sciences Drive MC 0850, La Jolla, CA 92093-0850, (858) 822-7611, (858) 534-7628 FAX, gsadler@ucsd.edu

INTRODUCTION

Educators, researchers, health care providers, policy makers, philanthropists, and others use large-scale databases to: ground their research foci and methodological approaches; develop public health interventions designed to make the best use of limited resources; inform legislative and governmental policy agendas; focus philanthropic strategies; and make patient care decisions. The authors have been using the California Cancer Registry (CCR) to demonstrate to medical students, residents and faculty how large datasets can be used to explore complex medical questions.[1–8] A key part of this training is to underscore the problems presented by missing data within a dataset. This article focuses the readers' attention on large datasets because many people erroneously feel that these are without fault and that their size compensates for missing data. This misunderstanding may be less likely with small, local databases where the deficiencies are more easily discerned, but where the problems associated with missing data are magnified.

This paper explores the CCR to help educators understand some of the key limitations of such databases.[9] California law mandates that, as of 1988, all health care facilities, physicians, and laboratories operating in California must report their patients' cancer diagnoses to the CCR; there are fines and other penalties for not doing so. Only basal and squamous cancers of the skin and a few tumors of borderline malignance are exempted from this mandate. The CCR is certified by the North American Association of Central Cancer Registries (NAACCR) and currently, the entire State's data are included in the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results Program (SEER) [10] and the NAACCR's combined incidence reports. The variables (fields) in the Registry are defined by these two organizations, and all are required to be gathered by the data abstractors.[11–13]

The CCR, a program of the California Department of Health Services, employs medical records abstractors, mostly tumor registrars who are certified, or eligible to be certified by the National Cancer Registrars Association,[14] to abstract the data from patients' charts. At the central registry, further data collection and quality assurance measures are undertaken, including the elimination of duplicate entries, clearance with death certificates, and mutual referral of patients who actually reside in other states.

The CCR is a statewide, total population-based registry that gathers information from the entire state population (36 million in 2004).[9] Data from 1988 through 2004 (2,393,084 individual cases), with all patient, provider, and institutional identification removed, are available to researchers. Such a large body of scientific data is particularly useful for clinical and epidemiologic researchers who may be interested in monitoring the changes in cancer incidence and mortality, surveillance of cancer prevention and control strategies, [9, 15] and other epidemiologic facets of cancer. As with all research data, the full value of a database is dependent on the accuracy and completeness of the data collection. However, creating a complete data set for each patient is a challenge because it depends upon health care providers' remembering to collect and record key data, patients' willingness to relinquish key data points, and the abstractors' ability to find and discern these data quickly in the patients' often-complex medical records.

No previous studies could be found which had quantified the percentages of missing data from the in-depth database users' perspective. For this study, missing data were defined as CCR codes "unknown," "unspecified," or "absent."

The hypothesis was that there were considerable variations in the percentages of missing sociodemographic and clinical data in the CCR among the eight most common cancer sites, as well as among reporting sources. Given that more accurate conclusions can be drawn from a complete database, this paper explores the extent of missing data in the CCR database from 1988 through 2004, the dates of the most complete data set available, as an example of limitations in the use of large central registry data. However, the purpose of quantifying missing data in this report is not to suggest means for more completeness of data, but rather to point out to users of the database, such as educators, clinicians, and researchers, the limitations of epidemiologic data in large, population-based datasets. These missing data are a type of reporting bias in that they are not in the hospital chart, the doctor's office record, or the laboratory report. Educators, like clinicians and researchers, must be aware of how large numbers of cases with missing data, which are not necessarily representative of the dataset as a whole, can skew their results and conclusions. Equally importantly, users of these findings need to understand the possibility of skewed results and incorporate such limitations when they consider potential applications of their studies.

This paper uses the CCR only as an example of a large data set and studies only a few of the variables that are deemed to be essential variables for most research projects. Other variables or other datasets could well show corresponding deficiencies. These deficiencies are identified in the CCR only to point out that they may well exist in any well-run dataset.

MATERIALS AND METHODS

For the purposes of this study, the eight most common cancer sites in the CCR for the years 1988 to 2004 were selected: breast, prostate, lung, colorectal, gynecological, lymphoma, melanoma, and leukemia. Together, these 1,781,869 cancer cases accounted for three-quarters of the cancer cases in the CCR (Table 1).

Four commonly used sociodemographic variables (*gender, age*, *ethnicity*, and *birthplace*) and three commonly used clinical variables (*diagnostic confirmation*, *stage*, and *differentiation*) of the CCR were studied. These are critical in the conduct of cancer disparities research and are required variables when abstracting cases[13] "*Diagnostic confirmation*" was defined as the basis for confirming a cancer diagnosis: tissue examination, bone marrow examination, clinical findings, radiologic interpretation, and other clinical bases. Percentages of missing data were recorded for each variable in relation to each cancer site. Ethnic groups included non-Hispanic White (73.7%), Black (6.2%), Hispanic (11.6%), and Asian/Pacific Islander (6.2%). Too few American Indians, Aleutians, or Eskimos were identified (0.1%) to allow analysis by this ethnic category. While the entire Asian/Pacific Islander group as a whole was used, the 24 sub-groups within this category were not, again because of small numbers.[11]

Table 1 shows the frequency of reporting for each of the six sources that report cases to the CCR. The top three reporting sources (hospitals/clinics, non-hospital independent laboratories, and physicians' offices), which together contributed 98.8% of cases to the CCR, were further analyzed to determine the percentages of data missing for each variable among the eight cancer sites. Data from nursing homes, autopsies only, and death certificates only cumulatively contributed only 1.2% of cases and were not further analyzed. When subsequent data are submitted on a patient, they are treated in a predetermined hierarchical order, with hospitals' and clinics'data being given precedence, followed by doctors' offices and laboratories.

RESULTS

Sociodemographic Data

For the three sources of data and the four sociodemographic characteristics, gender was the only socioeconomic characteristic for which there were no missing data and, therefore, was not included in tabulations. For age, only 0.1% was missing for the eight cancer sites (Table 2). *Ethnicity* followed, with only 2.2% missing. *Birthplace* was the most likely variable to be missing for all eight cancer sites, separately or combined (42.5%). For seven of the eight cancer sites, percentages of missing *birthplace* data ranged between 30.0% and 50.0%; melanoma was the exception with a percentage of missing data of 62.2%.

When the completeness of data for the six remaining variables was examined by *age at diagnosis*, there was a frequent, albeit occasionally inconsistent, trend for increased incompleteness with increased *age*. Further, as *age* increased, fewer reports came from hospitals/clinics, and more came from independent laboratories and physicians' offices. All ethnic groups examined showed a similar trend between increasing *age* and the distribution of cases among reporting sources.

Clinical Data

Of the clinical data examined, *diagnostic confirmation* had the least data missing, ranging from 0.3% to 4.4% for separate cancer sites and 1.5% for all eight cancers combined (Table 2). Examining the data missing by the *stage of disease* showed considerable variation among the eight cancer sites, with prostate (29.1%) showing a two-fold increase in missing *stage* data over the next highest cancer's (lung) missing data (13.4%) and more than a 13-fold increase over the missing *stage* data for leukemia (2.2%).

Differentiation was found to have the highest rate of missing data, with 36.3% of the data missing for all eight cancer sites combined. The percentages of missing *differentiation* data varied greatly among the eight cancer sites, with 99% of the melanoma *differentiation* data missing. In contrast, prostate cancer had the lowest percentage of *differentiation* data missing (8.2%).

Missing Data by Reporting Sources for all Eight Cancers Combined

The top three reporting sources were examined to evaluate the degree of missing data for each of the variables among the eight cancer sites being studied. Hospitals/clinics had the least missing data for each of the seven variables for all eight cancer sites combined (Table 3). Specifically, hospitals/clinics had very low percentages of missing data for *age*, *ethnicity*, and *diagnosis confirmation*, but the percentages of missing data increased for *birthplace* (40.8%) and *differentiation* (33.5%).

Physicians' offices had greater amounts of missing data for all seven variables compared to hospitals/clinics. Particularly noteworthy was the 71.7% missing data for *birthplace* and the 67.8% missing data for *differentiation* (Table 4).

Hospital/clinics outperformed the two aforementioned reporting sources in completion of data. Comparing all three reporting sources, physicians' offices had the most missing data for *differentiation* (67.8%) and *diagnostic confirmation* (5.4%) Independent laboratories reports had the highest percentages of missing data for *age (6.1%), ethnicity (48.4%), birthplace (83.7%)*, and *stage (26.9%)* for all eight cancer sites combined (Table 5).

Given that *birthplace* was the most frequently missing variable, the authors examined the Registry in search of a better source for these data. When data for living patients and deceased patients were compared, significantly more deceased patients (68.3%) than living patients (49.0%) had birthplace data present.

DISCUSSION

As hypothesized, this report has shown that considerable variation exists in the percentages of missing sociodemographic and clinical data in the CCR among reporting sources and eight cancer sites. Our findings are consistent with the problems with missing SEER data discussed previously by Mettlin et al. in an analysis of cancer patients diagnosed in 1992[16]. By using the CCR as an example of a large dataset, the quantification of missing CCR data as reported is only intended to caution users of large datasets about the possible limitations within such extensive bodies of scientific data. Moreover, this study, while not intending to offer solutions to this problem of missing data in large cancer registries, is suggesting some interpretations to why the data are missing in certain registry variables (such as differentiation, stage, ethnicity, and birthplace) as well as the significance of the missing data.

As reported in this study, there was tremendous variation in the percentages of missing *differentiation (grade)* data, ranging from a low of 8.2% in prostate to a high of 99% in melanoma. Part of this difference may be due to the widespread use of the only long-standing and well-accepted scoring system – the Gleason Score – for the *grade* of prostate cancer.[17] While there are some differentiation grading systems that are gaining acceptance for breast and colon cancer, there is essentially none for other cancer sites (melanoma for example).

Notably, *stage* has no significance in leukemia since leukemia is, by general agreement, always in the disseminated *stage.*[12, 18] Staging for leukemia was, in fact, not described in either the American Joint Committee on Cancer Staging Manual [19] or Furie et al.'s Clinical Hematology and Oncology textbook [20] suggesting that *stage* may not be a valid field to include in leukemia. Death clearance cases for leukemia are required to be staged as “unknown” by SEER and CCR and this may account for leukemia cases having 2.2% staged as “unspecified.”

Previous studies reported that percentages of missing birthplace data are notably high in certain ethnic groups [21–23]. The results of the current study support the findings of earlier studies by confirming that *birthplace* data did, in fact, have the highest percentages of missing data of all the examined variables for the eight most common cancer sites. There are various barriers to data completeness for variables such as *birthplace*. Some hospitals selectively inquire and/or record *birthplace* data of patients who were most likely foreign-born [21–23], whereas other hospitals did not collect *birthplace* data at all [24]. Some physicians may feel that asking such questions as *birthplace* and *ethnicity* may be perceived by patients as an intrusion into their privacy. Others may not see the value in collecting such data.[25, 26]

The actual reporting source of cancer information to the CCR should also raise the users’ caution, since discrepancies were observed when cancer data were examined by reporting source. As shown in the results, incomplete reporting is more prevalent in laboratories than in physicians’ offices or hospitals/clinics. These data suggest that laboratories receive from the physicians’ offices only the vital data needed for laboratory purposes. This is compounded by some laboratories being outside of California where procedures may be different.

A possible minor source of data omission is the quality of case abstraction by the registrar. However, the great majority of these individuals are well-trained, certified by the National Cancer Registrars Association (NCRA), and diligent in their search through patients’ charts for the needed data. Certified Tumor Registrars (CTR) strive to collect complete cancer information, provided it is documented in the patients’ medical records. CTR are required to participate in continuing education programs. Quality assurance programs at the regional and central registries monitor the CTR’s skills and performance.

We found that the very few cases located by death certificates had very complete *birthplace* and *ethnicity* data. Most likely this is due to the legal requirement to submit complete data on death certificates, suggesting that if the reporting process requires the information, the information is more likely to be provided.

CONCLUSIONS

This study, done from a cancer educator’s viewpoint, emphasizes the need for users of large datasets to be aware of the possible limitations of registry data. Such awareness may, in turn, prevent methodological flaws and misinterpretations of research findings. The database’s usefulness and its worth to users, reporting sources, and the public, are contingent on the completeness of the data submitted.

Acknowledgments

ACKNOWLEDGEMENTS AND FUNDING CONSIDERATIONS

This study was supported by funding from: the National Cancer Institute’s grants R25CA65745 and Cancer Center Core grant #5 P30 CA023100-22, the National Institutes of Health’s Division of National Center on Minority

Health and Health Disparities EXPORT grant P60MD00220; and the Minority Institution/Cancer Center Partnership Program grants #U56 CA92079 and #U56 CA92081. The student authors were funded by the National Institute of Aging grant T35AG026757 and the Undergraduate Summer Research Program at the University of California San Diego.

The authors wish to thank Lilia O'Connor and Nancy Schlag of the California Cancer Registry for their review of, and suggestions for, this article.

DISCLAIMER

The collection of cancer incidence data used in this study was supported by the California Department of Health Services as part of the statewide cancer reporting program mandated by the California Health and Safety Code Section 103885, the National Cancer Institute's Surveillance, Epidemiology and End Result Program, and the Centers for Disease Control and Prevention National Program of Cancer Registries. The ideas and opinions expressed herein are those of the author and endorsement by the State of California, Department of Health Services, the National Cancer Institute and the Centers for Disease Control and Prevention is not intended nor should be inferred.

REFERENCES

1. Blair SL, Sadler GR, Bristol R, Summers C, Tahir Z, Saltzstein SL. Early cancer detection among rural and urban Californians. BMC Public Health. 2006; 6:194. [PubMed: 16869975]

2. Grabowski J, Saltzstein SL, Sadler GR, Blair SL. Intracystic papillary carcinoma: a review of-917 cases. Cancer. 2008; 113(5):916–920. [PubMed: 18661510]

3. Grabowski J, Saltzstein SL, Sadler GR, Blair SL. Squamous cell carcinoma of the breast: a review of 177 cases. (Submitted for publication).

4. Grabowski J, Saltzstein SL, Sadler GR, Tahir Z, Blair S. A comparison of merkel cell carcinoma and melanoma: results from the California Cancer Registry. Clin Med: Oncol. 2008; 2:327–333. [PubMed: 21892294]

5. Saltzstein, SL.; Behling, CA. Cancer in the Chronologically Gifted. Irvine, CA: Cancer Surveillance Program of Orange Country/San Diego and Imperial Organization for Cancer Control; 2003 Jul.

6. Schultzel M, Saltzstein SL, Downs TM, Shimasaki S, Sanders C, Sadler GR. Late age (85 years or older) peak incidence of bladder cancer. J Urol. 2008 Apr; 179(4):1302–1305. discussion 1305–1306. [PubMed: 18289593]

7. Stafford HS, Saltzstein SL, Shimasaki S, Sanders C, Downs TM, Sadler GR. Racial/ethnic and gender disparities in renal cell carcinoma incidence and survival. J Urol. 2008; 179(5):1704–1708. [PubMed: 18343443]

8. Summers C, Saltzstein SL, Blair SL, Tsukamoto TT, Sadler GR. Racial/ethnic differences in early detection of breast cancer: A study of 250,985 cases from the California Cancer Registry. (Submitted for publication).

9. California Cancer Registry (CCR). [Accessed July 12, 2007] About the California Cancer Registry. 2007. http://www.ccrcal.org/abouttheccr.html.

10. Division of Cancer Control and Population Sciences, National Cancer Institute. [Accessed July 9, 2007] SEER (Surveillance Epidemiology and End Results): Providing information on cancer statistics to reduce the burden of this disease on the U.S. population. http://seer.cancer.gov.

11. [Accessed April 30, 2009] Cancer Reporting in California, Volume III: Data Management System Standards (2009 Data Changes Included). 2008 Dec. http://www.ccrcal.org/VOL3/Title_Page.htm.

12. Young, JLJ.; Roffers, SD.; Ries, LAG.; Fritz, AG.; Hurlbut, AA., editors. SEER Summary Staging Manual - 2000: Codes and Coding Instructions. Bethesda, MD: National Cancer Institute; 2001. NIH Pub. No. 01-4969;

13. Havener, LA.; Thornton, ML., editors. Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Thirteenth Edition, Version 11.3. Springfield, IL: North American Association of Central Cancer Registries; 2008 Apr.

14. Personal communication: Email from Roshala, W. to Saltzstein, S.L. re: Certification of Tumor Registrars. 2009 May 4.

15. National Center for Chronic Disease Prevention and Health Promotion. Centers for Disease Control and Prevention. National Program of Cancer Registries (NPCR). 2001. http://apps.nccd.cdc.gov/uscs/.
16. Mettlin CJ, Menck HR, Winchester DP, Murphy GP. A comparison of breast, colorectal, lung, and prostate cancers reported to the National Cancer Data Base and the Surveillance, Epidemiology, and End Results Program. Cancer. 1997 May 15; 79(10):2052–2061. [PubMed: 9149035]
17. Rosai, J. Ackerman's Surgical Pathology. 8th ed. St. Louis: Mosby-Year Book, Inc; 1996.
18. California Cancer Reporting System Standards. Volume I: Cancer Reporting in California, Abstracting and Coding Procedures for Hospitals (Eighth Edition). Revised May 2008. http://www.ccrcal.org/Vol_1_May_2008_html/0-Front_Matter/Cover_Pagehtm.
19. American Joint Committee on Cancer. Manual for Staging of Cancer. Third ed. Philadelphia: J.B. Lippincott; 1988.
20. Furie, B.; Cassileth, PA.; Atkins, MB.; Mayer, RJ. Clinical Hematology and Oncology. Presentation, Diagnosis, and Treatment. Philadelphia: Churchill Livingstone; 2003.
21. Gomez SL, Glaser SL. Quality of cancer registry birthplace data for Hispanics living in the United States. Cancer Causes Control. 2005 Aug; 16(6):713–723. [PubMed: 16049810]
22. Gomez SL, Glaser SL, Kelsey JL, Lee MM. Bias in completeness of birthplace data for Asian groups in a population-based cancer registry (United States). Cancer Causes Control. 2004 Apr; 15(3):243–253. [PubMed: 15090719]
23. Lin SS, O'Malley CD, Lui SW. Factors associated with missing birthplace information in a population-based cancer registry. Ethn Dis. 2001 Fall;11(4):598–605. [PubMed: 11763284]
24. Gomez SL, Le GM, West DW, Satariano WA, O'Connor L. Hospital policy and practice regarding the collection of data on race, ethnicity, and birthplace. Am J Public Health. 2003 Oct; 93(10): 1685–1688. [PubMed: 14534222]
25. Konowitz PM, Petrossian GA, Rose DN. The underreporting of disease and physicians' knowledge of reporting requirements. Public Health Rep. 1984 Jan-Feb;99(1):31–35. [PubMed: 6422492]
26. Seixas NS, Rosenman KD. Voluntary reporting system for occupational disease: pilot project, evaluation. Public Health Rep. 1986 May-Jun;101(3):278–282. [PubMed: 3086920]

Table 1

Frequency of data reporting to the CCR according to medical sources

	Frequency of reporting (%)						
	Reporting Sources						
	Number Of Cases	Hospitals	Laboratories*	Physicians' Offices	Nursing Homes	Autopsies Only	Death Certificates Only
Cancer							
Breast	393,172	97.8	0.5	1.1	0.1	0.1	0.5
Prostate	329,374	88.9	2.2	7.7	0.1	0.3	0.9
Lung	292,509	95.3	0.2	2.2	0.2	0.5	1.6
Colorectal	251,757	97.0	0.5	1.4	0.1	0.1	0.8
Gynecological	204,321	93.1	2.2	4.2	0.1	0.1	0.4
Lymphoma	125,862	94.9	1.0	2.8	0.1	0.5	0.8
Melanoma	124,831	66.4	7.1	26.4	0.1	0.1	0.1
Leukemia	60,043	92.2	1.1	4.2	0.2	0.2	2.1
All 8 cancers	1,781,869	92.6	1.4	4.8	0.1	0.2	0.8

[*] non-hospital independent laboratories

Table 2

Percentages of missing data in the CCR according to sociodemographic and clinical variables

	Sociodemographic Variables			Clinical Variables		
	Age **% missing**	**Ethnicity** **% missing**	**Birthplace** **% missing**	**Diagnostic Confirmation** **% missing**	**Stage** **% missing**	**Differentiation** **% missing**
Cancer						
Breast	0.1	0.9	43.0	0.7	2.9	25.8
Prostate	0.1	3.6	47.3	1.7	29.1	8.2
Lung	0.0	0.3	30.5	3.0	13.4	45.5
Colorectal	0.0	0.8	40.8	1.1	5.9	19.2
Gynecological	0.4	3.8	46.4	0.7	4.0	47.8
Lymphoma	0.1	1.4	37.4	2.0	10.5	54.0
Melanoma	0.5	8.9	62.2	0.3	4.3	99.0
Leukemia	0.1	1.4	34.2	4.4	2.2	79.4
All 8 cancers	0.1	2.2	42.5	1.5	7.0	36.3

[*] The gender variable was included in examination, but not tabulation

Table 3

Percentages of missing data in the CCR from hospitals/clinics

	Percentage Data Missing (%)					
	Reporting Variables					
	Birthplace	**Age**	**Ethnicity**	**Diagnosis Confirmation**	**Stage**	**Differentiation**
Cancer						
Breast	42.8	0.0	0.5	0.1	1.8	24.8
Prostate	44.5	0.0	1.0	0.6	25.2	6.8
Lung/Bronchus	31.6	0.0	0.2	0.6	10.8	43.3
Colorectal	41.0	0.0	0.4	0.1	4.3	17.3
Gynecological	43.9	0.0	1.3	0.2	3.0	44.5
Lymphoma	37.8	0.1	1.4	0.0	9.0	52.0
Melanoma	51.0	0.0	1.9	0.1	3.7	98.7
Leukemia	34.1	0.0	0.5	1.8	0.2	78.6
All 8 cancers	40.8	0.0	0.7	0.4	5.2	33.5

Table 4

Percentages of missing data in the CCR from physicians’ offices

	Percentage Data Missing (%)					
	Reporting Variables					
	Birthplace	Age	Ethnicity	Diagnosis Confirmation	Stage	Differentiation
Cancer						
Breast	58.8	0.8	18.4	8.3	34.3	62.3
Prostate	76.0	0.3	22.3	2.5	53.0	10.9
Lung/Bronchus	7.3	0.1	1.1	33.7	71.1	92.6
Colorectal	46.2	0.2	12.3	13.6	42.2	76.8
Gynecological	83.8	2.7	32.6	3.2	12.4	92.6
Lymphoma	57.5	0.2	13.9	9.2	34.0	68.9
Melanoma	84.9	0.4	16.0	0.3	3.3	99.5
Leukemia	43.5	0.3	10.6	12.0	0.7	84.2
All 8 cancers	71.7	0.6	18.1	5.4	18.0	67.8

Table 5

Percentages of missing data in the CCR from non-hospital independent laboratories

	Percentage Data Missing (%)					
	Reporting Variables					
	Birthplace	Age	Ethnicity	Diagnosis Confirmation	Stage	Differentiation
Cancer						
Breast	79.9	4.2	44.3	1.7	37.8	52.7
Prostate	82.7	5.0	48.3	1.1	74.9	12.3
Lung/Bronchus	40.1	1.8	17.4	18.8	68.1	67.7
Colorectal	73.8	4.2	43.8	3.2	38.2	59.4
Gynecological	92.2	11.1	59.5	0.5	12.1	90.7
Lymphoma	78.3	4.9	48.5	3.0	54.1	67.6
Melanoma	87.6	5.8	47.1	0.1	8.9	99.6
Leukemia	73.8	6.1	45.2	4.1	0.4	81.4
All 8 cancers	83.7	6.1	48.4	1.5	26.9	65.4