

# **Renal Cancer Patients with Unknown Ethnicity in Cancer Registry Data: Comparisons to Patients with Known Ethnicity**

Jie Lin, PhD, MPH<sup>1\*</sup>, Elizabeth Butts, MPH<sup>2</sup>, Paul D. Rockswold, MD<sup>2</sup>, Craig D. Shriver, MD<sup>1,3</sup>,

Kangmin Zhu, MD, PhD<sup>1,3\*</sup>

<sup>1</sup>John P. Murtha Cancer Center, Walter Reed National Military Medical Center, Bethesda, MD

<sup>2</sup>Health Analysis, Navy and Marine Corps Public Health Center, Portsmouth, VA

<sup>3</sup>Uniformed Services University, Bethesda, MD

**Disclaimers:** The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Army, Department of Defense, nor the U.S. Government. Nothing in the presentation implies any Federal/DOD endorsement.

**Running title:** renal cancer patients with unknown ethnicity

**Key Words:** unknown ethnicity, cancer registry, renal cell carcinoma, survival

**Manuscript Category:** Research Article

**Word Count:** 3,175

**Number of tables:** 2

**Number of Figures:** 1

**\*Corresponding Authors:** Kangmin Zhu, MD, PhD or Jie Lin, PhD, MPH; Division of Military Epidemiology and Population Sciences, John P. Murtha Cancer Center, Walter Reed National Military Medical Center, 11300 Rockville Pike, Suite 1120, Rockville, MD 20852; 301-816-4787 (Dr. Zhu), 301-816-4787 (Dr. Lin), 301- 881-7197 (fax), [kzhu@murthacancercenter.org](mailto:kzhu@murthacancercenter.org) (Dr. Zhu) or [jlin@murthacancercenter.org](mailto:jlin@murthacancercenter.org) (Dr. Lin).

## Abstract

**Background:** Information on ethnicity is important for health disparity research and health service planning. However, information on ethnicity is often incomplete in large routine databases such as cancer registries. This study aimed to compare survival status and other characteristics between cancer patients with and without information on Hispanic ethnicity in cancer registry data.

**Methods:** The study included 2,426 patients with clear cell renal cell carcinoma (RCC) diagnosed between 1988 and 2004 and identified from the U.S. Department of Defense (DoD)'s Automated Central Tumor Registry (ACTUR) database. There were 1,353 non-Hispanics, 134 Hispanics and 939 patients with unknown ethnicity. Patients were followed through death, date of last contact, or censored on December 31, 2007.

**Results:** Patients with unknown ethnicity exhibited significantly shorter survival than non-Hispanic or Hispanic patients (Log Rank  $P < 0.0001$ ). Further analysis showed that compared to patients with known ethnicity, patients with unknown ethnicity were more likely to have advanced tumor stage at diagnosis, more likely to have missing information on tumor grade and size, and less likely to receive surgery. After adjustment for demographic, tumor and treatment variables, patients with unknown ethnicity still exhibited significantly higher mortality than non-Hispanic patients (Hazard ratio (HR) = 1.69; 95% confidence interval (CI) = 1.48 - 1.92), while Hispanic patients were not different from non-Hispanic patients (HR=0.95; 95% CI=0.71 - 1.28). The shorter survival in the unknown ethnicity group was consistently observed in subgroups by age, race, stage, grade and surgical treatment, suggesting factors other than those investigated in the current study may play a role in the survival differences between patients with and without information on Hispanic ethnicity.

**Implications:** This study suggests that clear cell RCC patients with unknown ethnicity in ACTUR may be different from those with known ethnicity in survival, tumor features, cancer

treatment, completeness of tumor data, and other factors. The results warrant future studies about missing mechanisms. The extent of unknown Hispanic ethnicity, which was also associated with missing in other variables, suggests the need to improve the collection of data on ethnicity.

## Introduction

Race and ethnicity data are essential for health disparity research that provides evidence for policy-making and service planning in reducing racial and ethnic disparity. Large routine health datasets such as cancer registry data, medical claims data, often contain information on race and ethnicity, providing resources for research on racial and ethnic disparities. However, missing race and/or ethnicity is common in such datasets, which limits the use of such datasets.<sup>1,2</sup> Furthermore, the mechanisms of missing is not adequately addressed and investigated by researchers.<sup>3,4</sup> Therefore, the effects of missing ethnicity on research results are unclear.

Understanding of missing mechanisms, such as why data is missing in some subjects and whether the missing is random or not, is critical to the interpretation of results and development of imputing algorithms.<sup>2,4-9</sup> However, few previous studies investigated why data on race and/or ethnicity are missing for some subjects but not others. A recent study, which did not distinguish race and ethnicity, found that breast cancer screening rate was significantly lower among women with missing race/ethnicity as compared to women with known race/ethnicity, including minorities<sup>10</sup>. Further analysis revealed that women with missing race/ethnicity data had fewer routine office visits and were less likely to have an identified primary care physician as compared to women with known race/ethnicity<sup>10</sup>. These observations suggest the importance of understanding missing mechanisms and better characterizing subjects with missing race/ethnicity in large routine datasets.

Conceptually, race and ethnicity are different. While race usually refers to biologic inheritance and phenotypic traits such as skin color and facial features, ethnicity is more cultural-related and is characterized by distinctive social and cultural tradition as well as biologic heritage<sup>11-13</sup>. The U.S. Census Bureau has distinguished race and ethnicity since 1990<sup>14-16</sup>. Two ethnicities, namely, Hispanics and Non-Hispanics are required to report by the Office of Management and Budget (OMB). Since ethnicity reflects social and cultural features that may

be modifiable to improve health-related behaviors<sup>17,18</sup>, it is interesting to assess whether individuals with unknown ethnic information may differ from those with known ethnicity in cancer outcomes and related factors. This helps facilitate research in cancer disparity and identify specific groups for intervention. To date, there has been little research comparing individuals with and without data on ethnicity among cancer patients.

As part of an ongoing research on survival among patients with renal cell carcinoma (RCC) identified from the Department of Defense's (DoD) Automated Tumor Registry (ACTUR), we observed a high percentage of patients with unknown Hispanic ethnicity and a significantly worse survival for the group compared to those with known ethnicity. The current study aimed to characterize patients with unknown Hispanic ethnicity in survival as compared to those with known ethnicity and assess factors related to differences in survival including demographic variables, tumor characteristics and receipt of treatments.

## **Methods**

### Sources of data

Sources of data were described previously<sup>19</sup>. Briefly, data on patients diagnosed with RCC between 1988 and 2004 were obtained from the ACTUR, a clinical tracking system for all cancer patients who were diagnosed and/or received cancer treatment at military treatment facilities. The facilities are required to report cancer data to ACTUR on all DoD beneficiaries including active-duty members, retirees and their dependents. The ACTUR data are reviewed by specialized registrars to verify diagnosis. ACTUR meets all requirements of the North American Association of Central Cancer Registry (NAACCR) standards<sup>20</sup>. The ACTUR data contain information on age at diagnosis, sex, race, ethnicity, marital status, active duty status, military service branch, tumor stage, tumor grade, tumor size, cancer treatment (e.g. surgery, chemotherapy, radiation, etc.), tumor recurrence, vital status, and date of last contact or death.

Regarding ethnicity, ACTUR has a Spanish/Hispanic origin data field that has code for non-Hispanics and codes for different Hispanic origins (Mexican, Puerto Rican, Cuban, South or Central America (except Brazil), other specified Spanish/Hispanic origin (includes European), other non-specified Spanish/Hispanic origins, and Spanish surname only). Ethnicity was coded as “unknown” if ethnicity information was not found in medical records, death certificates or other sources that define Hispanic origin. In our data analysis, subjects with different Hispanic origins were grouped into “Hispanic”. As a result, there were three study groups: non-Hispanic, Hispanic and unknown ethnicity. This study was based on the non-identifiable ACTUR data approved for our research by the institutional review boards of the former U.S. Military Cancer Institute, Walter Reed National Military Medical Center, and the former Armed Forces Institute of Pathology.

#### Study Subjects

Study subjects were patients who were diagnosed with histologically confirmed primary renal cell carcinoma (RCC) with clear cell histology between January 1, 1988 and December 31, 2004. Clear cell type constitutes over 85% of all RCC.<sup>21,22</sup> The diagnosis was defined by the tumor site code (C64.9) and morphology codes (8310-8312) of the International Classification of Diseases for Oncology, third edition (ICD-O-3)<sup>23</sup>. Patients with a history of cancers other than RCC were excluded from the study to minimize the potential effect of multiple cancers on survival. A total of 2,426 patients were included in the final analysis.

Patients were followed through death, date of last contact, or censored on December 31, 2007. The observed survival time was calculated as the difference between date of diagnosis and date of death for patients who died during the study period. For patients who did not die during the study period, survival time was censored at the date of last contact or December 31, 2007. The study outcome was all-cause mortality. As the first step of data analysis, Kaplan-Meier survival curves and Log-rank test were used to assess whether survival differed between

unknown ethnicity group, non-Hispanic and Hispanic groups. We then compared the groups in demographic variables, tumor characteristics and receipt of treatments to investigate whether these factors were associated with unknown ethnicity status. Finally, we assessed whether the differences in survival between the unknown ethnicity and known ethnicity groups remained in the Cox regression model adjusting for age, gender, race, active duty status, marital status, tumor stage, grade, size, recurrence, receipt of surgery, chemotherapy, radiation therapy, and recurrence. We further conducted Cox regression analysis stratified by demographic characteristics, tumor stage, tumor grade, and treatments (surgery, chemotherapy, and radiation therapy). Hazard ratios (HRs) and their 95% confidence intervals (95% CIs) associated with ethnicity were calculated. Statistical analyses were conducted using the SAS software version 9.3.0.

## **Results**

There were 1,353 non-Hispanics, 134 Hispanics and 939 with unknown ethnicity, representing 55.8%, 5.5% and 38.7% of the study subjects, respectively. During follow up, 529, 50, and 609 non-Hispanic, Hispanic, and unknown ethnicity patients died, respectively. The Kaplan-Meier analysis showed significant worse survival in patients with unknown ethnicity than non-Hispanic or Hispanic groups (Log Rank  $P < 0.0001$ ) (Figure 1). The worse survival of patients with unknown ethnicity was consistently observed in subgroups stratified by age, sex, race and tumor stage (Results not shown).

Table 1 shows the distributions of demographic, tumor and treatment characteristics by ethnicity status. The distributions of age and sex were similar between subjects of unknown ethnicity and non-Hispanic subjects, while the Hispanic group tended to be younger at diagnosis. The unknown ethnicity group was more likely to be White (86.9%) than the non-Hispanic group (81.7%). Patients with unknown ethnicity were more likely to be diagnosed at distant stage (23.4%), as compared to non-Hispanic (13.0%) and Hispanic subjects (16.4%).

They also had the highest percentage of unknown status in terms of tumor grade, tumor size, marital status, and service branch. Regarding treatments, 19.7% of patients with unknown ethnicity did not receive surgery of any type, while 10.1% of non-Hispanic patients and 11.9% of Hispanic patients, respectively, received no surgery. However, the percentage of receiving chemotherapy was higher in patients with unknown ethnicity (4.9%) than non-Hispanic patients (2.3%) and Hispanic patients (3.0%). Similarly, higher percentage of receiving radiation therapy was observed in patients with unknown ethnicity (10.0%) than non-Hispanic (5.3%) and Hispanic patients (5.2%). Also, patients with unknown ethnicity were more likely to have no cancer recurrence (80.6%) than non-Hispanic (72.0%) and Hispanic (76.9%) patients.

After adjustment for demographic, tumor and treatment variables, patients with unknown ethnicity exhibited a higher hazard ratio compared to non-Hispanic subjects (HR = 1.69, 95%CI, 1.48-1.92), while Hispanic patients were similar to non-Hispanics in HR (HR=0.95; 95% CI=0.71 to 1.28)(Table 2). The higher HR for the patients with unknown ethnicity than non-Hispanics was consistently present regardless of race, age, sex, tumor stage, tumor grade, receipt of surgery, chemotherapy or radiation therapy in stratified analyses (Table 2). Similar to the overall analysis, no survival disadvantage was observed among patients of Hispanic origin compared to non-Hispanic subjects.

## **Discussion**

This study showed that renal cancer patients with unknown Hispanic ethnicity experienced significantly worse survival as compared to both non-Hispanic and Hispanic patients in the ACTUR data. Further analyses showed that they were more likely to be diagnosed at a more advanced stage, were less likely to receive surgery, but more likely to receive chemotherapy and/or radiation therapy, than patients with known ethnicity. In addition, patients with unknown ethnicity were more likely to have missing information on tumor grade,

tumor size, marital status, and service branch. Finally, the poor survival in this group was consistently observed regardless age, sex, race, tumor stage, tumor grade or receipt of surgery.

Few previous studies compared characteristics between patients with and without information on Hispanic ethnicity. Two studies on survival in breast cancer patients, which did not distinguish race and ethnicity, reported survival difference between the unknown racial group and the known racial groups.<sup>24,25</sup> However, worse survival of the unknown racial group was observed in one study<sup>24</sup> but the other study found the opposite<sup>25</sup>. We know of no studies comparing patients with unknown Hispanic ethnicity and those with known ethnicity in survival and related factors among cancer patients.

To identify factors that may be related to the poor survival in renal cancer patients with unknown ethnicity, we compared demographic, tumor and treatment characteristics of this group with the non-Hispanic and Hispanic groups. Results showed that they were more likely to be diagnosed at advanced stage, receive chemotherapy and/or radiation therapy, but less likely to receive surgery than those with known ethnicity. The lower rate of receiving surgery and higher rate of receiving chemotherapy and radiation therapy in patients with unknown ethnicity might result from their later tumor stage, which might lead to their poorer survival. However, further multivariate analysis showed that the poor survival still existed after adjustment for tumor stage and other variables, and the survival disadvantage of the group with unknown ethnicity was consistently observed in subgroups of patients stratified by tumor stage, tumor grade, cancer treatments, and demographic characteristics. Thus, factors other than those investigated in the current study may also be associated with the poorer survival among renal cancer patients with unknown ethnicity.

Among factors not examined, negative health promotion behaviors and attitudes towards screening and treatments in patients with missing race/ethnicity data were indicated.<sup>10</sup> For example, compared to women with known race/ethnicity, women with missing race/ethnicity were less likely to seek primary care and visit doctors less frequently<sup>10</sup>, suggesting the

possibility that patients with missing race/ethnicity are less likely to have positive health promotion behaviors.<sup>10</sup>

Our findings that patients with unknown ethnicity were more likely to have missing information on tumor grade, tumor size, marital status, and other variables suggest that the missing information on ethnicity is not random<sup>9</sup>. The non-random missing might result from some underlying differences between this group and other groups. For example, a higher percentage of missing on tumor grade in patients with unknown ethnicity may imply less adequate or unavailable pathologic specimens for assessing tumor grade than those with known ethnicity.

While it is currently difficult to comprehend underlying specific mechanisms behind the coded unknowns, understanding the sources of information on Hispanic ethnicity may be helpful to infer possible mechanisms. The information on Hispanic ethnicity in the ACTUR is obtained from administrative and medical records which are often based on self-reported Hispanic ethnicity. Hispanic ethnicity is also defined by registrars using information on Hispanic origin stated on death certificates, birthplace, country of origin, life history and/or language spoken, patient's last name or maiden name found on a list of Hispanic names, and a combination of methods<sup>20</sup>, which is convoluted or cumbersome. When evidence from these sources is not available or not readily accessible, registrars code ethnicity as unknown. Thus, missing ethnicity may primarily reflect unwillingness to report ethnicity or unavailable information on language, life history, or other factors that help determine ethnicity. Unwillingness to report ethnicity may be related to socioeconomic status and health behavior. Unavailable information on language or life history may imply less utilization of health care service, particularly considering the higher percentages of unknown tumor grade and size in the group.

While our study was based on the military health care system and thus the results may not be directly generalizable to other populations, our results have important implications for understanding missing mechanisms and reducing the missing ethnicity information in the

ACTUR data. First, the poor survival in renal cancer patients with unknown ethnicity suggests that they may represent a susceptible subgroup that needs special attention in cancer surveillance and survivorship programs. Second, a better understanding of missing mechanisms such as why missing occurred in some subjects but not others may suggest potential perspectives to be considered in the development of imputation algorithms. Third, the extent of unknown Hispanic ethnicity in the ACTUR, which was also associated with missing in other variables, suggests the need to improve the collection of data on ethnicity.

## **Acknowledgements**

This project was supported by John P. Murtha Cancer Center, Walter Reed National Military Medical Center via the Uniformed Services University of the Health Sciences under the auspices of the Henry M. Jackson Foundation for the Advancement of Military Medicine. The authors thank the Joint Pathology Center (formerly Armed Forces Institute of Pathology) for providing the data.

## References

1. Izquierdo JN, Schoenbach VJ. The potential and limitations of data from population-based state cancer registries. *Am J Public Health*. May 2000;90(5):695-698.
2. Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res*. Oct 2008;43(5 Pt 1):1722-1736.
3. Long JA, Bamba MI, Ling B, Shea JA. Missing race/ethnicity data in Veterans Health Administration based disparities research: a systematic review. *J Health Care Poor Underserved*. Feb 2006;17(1):128-140.
4. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
5. Fiscella K, Fremont AM. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Serv Res*. Aug 2006;41(4 Pt 1):1482-1500.
6. Derose SF, Contreras R, Coleman KJ, Koebnick C, Jacobsen SJ. Race and ethnicity data quality and imputation using U.S. Census data in an integrated health system: the Kaiser Permanente Southern California experience. *Med Care Res Rev*. Jun 2013;70(3):330-345.
7. Ryan R, Vernon S, Lawrence G, Wilson S. Use of name recognition software, census data and multiple imputation to predict missing data on ethnicity: application to cancer registry records. *BMC Med Inform Decis Mak*. 2012;12:3.
8. He Y, Yucel R, Zaslavsky AM. Misreporting, Missing Data, and Multiple Imputation: Improving Accuracy of Cancer Registry Databases. *Chance (N Y)*. Sep 2008;21(3):55-58.

9. Egleston BL, Wong YN. Sensitivity analysis to investigate the impact of a missing covariate on survival analyses using cancer registry data. *Stat Med*. May 1 2009;28(10):1498-1511.
10. Kempe KL, Larson RS, Shetterley S, Wilkinson A. Breast cancer screening in an insured population: whom are we missing? *Perm J*. Winter 2013;17(1):38-44.
11. Last J. *A dictionary of epidemiology*. New York, NY: Oxford University Press; 1995.
12. Lin SS, Kelsey JL. Use of race and ethnicity in epidemiologic research: concepts, methodological issues, and suggestions for research. *Epidemiol Rev*. 2000;22(2):187-202.
13. Ahdieh L, Hahn RA. Use of the terms 'race', 'ethnicity', and 'national origins': a review of articles in the American Journal of Public Health, 1980-1989. *Ethn Health*. Mar 1996;1(1):95-98.
14. Grieco EM, Cassidy RC. Overview of race and Hispanic origin. *Census 2000 Brief* 2000; <http://www.census.gov/prod/2001pubs/c2kbr01-1.pdf>. Accessed March 5, 2014.
15. Humes KR, Jones NA, Ramirez RR. Overview of race and Hispanic origin. *2010 Census Briefs* 2010; <http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf>. Accessed March 5, 2014.
16. Ennis SR, Rios-Vargas M, Albert NG. The Hispanic population: 2010. *2010 Census Briefs* 2010; <http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf>. Accessed March 5, 2014.
17. Kagawa-Singer M, Dadia AV, Yu MC, Surbone A. Cancer, culture, and health disparities: time to chart a new course? *CA Cancer J Clin*. Jan-Feb 2010;60(1):12-39.
18. Cannon AJ. Delivering culturally competent care in clinical practice: a call to action. *J Natl Med Assoc*. Jan-Feb 2012;104(1-2):104-107.

19. Zheng L, Enewold L, Zahm SH, et al. Lung cancer survival among black and white patients in an equal access health system. *Cancer Epidemiol Biomarkers Prev.* Oct 2012;21(10):1841-1847.
20. Tryon J. User's Guide For ACTUR Cancer Registry Software System Abstracting Module2007.
21. Chow WH, Shuch B, Linehan WM, Devesa SS. Racial disparity in renal cell carcinoma patient survival according to demographic and clinical characteristics. *Cancer.* Jan 15 2013;119(2):388-394.
22. Patard JJ, Leray E, Rioux-Leclercq N, et al. Prognostic value of histologic subtypes in renal cell carcinoma: a multicenter experience. *J Clin Oncol.* Apr 20 2005;23(12):2763-2771.
23. Fritz A, Percy C, Lack A, et al. International Classification of Diseases for Oncology. 3rd ed. Geneva: World Health Organization; 2000.
24. Downing A, West RM, Gilthorpe MS, Lawrence G, Forman D. Using routinely collected health data to investigate the association between ethnicity and breast cancer incidence and survival: what is the impact of missing data and multiple ethnicities? *Ethn Health.* Jun 2011;16(3):201-212.
25. Jack RH, Davies EA, Moller H. Breast cancer incidence, stage, treatment and survival in ethnic groups in South East England. *Br J Cancer.* Feb 10 2009;100(3):545-550.

## Figure Legend

**Figure1.** Kaplan-Meier survival curves by ethnicity in clear cell RCC patients diagnosed from 1988 to 2007 in the U.S. Department of Defense Cancer Registry