The Relative Importance of Race Compared to Health Care and Social Factors in Predicting Prostate Cancer Mortality: A Random Forest Approach

Heidi A. Hanson1,2, Christopher R. Martin1, Brock O’Neil1, Claire L. Leiser2, Eric N. Mayer3, Ken R. Smith2,4, William T. Lowrance1
1. Department of Surgery, University of Utah; 2. Population Sciences, Huntsman Cancer Institute; 3. UCLA Medical School; 4. Department of Family and Consumer Studies, University of Utah

BACKGROUND

- There is a differential in prostate cancer (PCa) mortality between black and white men. Advances in precision medicine have shifted the research focus toward underlying genetic differences. However, non-biologic factors may play a large role in these observed disparities.

INNOVATION

- This study expands upon previous studies by identifying patterns of risk of lethal PCa using the National Cancer Institute’s Surveillance, Epidemiology, and End Results (SEER) registry’s national cancer surveillance data combined with the Health Resources and Services Administration’s Area Health Resource File (AHRF).
- Identifying the most important factors for predicting prostate cancer mortality requires methodologies that take into account multiple variables (and their interactions) simultaneously. Random forest regression methods incorporate all multiway interactions and provide measures of variable importance, which quantify the importance of each variable in predicting survival classes.

OBJECTIVE

To measure the relative importance of race compared to healthcare and social factors on PCa specific mortality.

METHODS

Data: Cancer related data are captured by the 18 registries comprising the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI). County-level health care data were obtained from the Department of Health and Human Services Area Health Resource File (AHRF).

The Center for Disease Control (CDC) Social Vulnerability Index (SVI) was included in the models.

Statistical Methods: Random forests (RF) are an extension of Classification and Regression Tree modeling. In this method, n trees were grown using a bootstrapped sample from the learning sample. The out of bag (oob) observation data was used to compute permutation accuracy variable importance measure (VIMP) by predicting class membership in the oob sample and then permuting the variables and calculating the predictive accuracy with permuted variables.

Random forest models with all variables were also run for each age and stage. These models only included main effects and were used to assess the benefit of using an RF approach. We found that there were substantive differences in interpretation between the RF and Cox PH models.

RESULTS

Distribution of Individual Level Variables by Race/Ethnicity

Distribution of Social Factors by Race/Ethnicity

Distribution of Access to Healthcare by Race/Ethnicity

Random Forest Variable Importance Measures

CONCLUSIONS

Tumor characteristics were the most important factors for predicting PCa mortality for all ages and stages. All other factors combined ranged from 1/4th to 1/10th as important when directly compared to tumor characteristics. While race was a measurably important factor in predicting PCa mortality, the healthcare and social factors (factors also known to be associated with racial disparities) were more important for all but 2 of the 18 groups in our analysis. Precision medicine has largely focused on the biological variables that can be used to individualize health outcomes, but precision medicine that fails to incorporate the framework of socially based health disparities, that includes but is not limited to SES, will likely fail to improve outcomes. Eradicating disparities in PCa survival will take a multi-pronged approach including advances in precision medicine. Disparities will persist unless healthcare access and social equality is achieved among all populations.

Acknowledgements: We thank the Pediatric and Population Research of the Huntsman Cancer Institute and the Office of Minority Disease-Oriented Teams (SUMDOT) for providing the pilot funds, data and valuable computing support. We also thank the National Institutes of Health for funding part of this research (U0S2CA163248). We acknowledge the Huntsman Cancer Institute's National Cancer Institute Cancer Center Support Grant P30 CA042014.