Development and Validation of a Clinically Usable Prediction Model for Other-Cause Mortality in Men with Prostate Cancer using Two Prospective National Cohorts

**Abstract**

**Background**: Most men diagnosed with localized prostate cancer die of a cause other than prostate cancer. NCCN guidelines for prostate cancer depend on life expectancy in addition to cancer stage; optimal treatment selection depends on risk of other cause (OC) mortality. However, few tools exist to predict OC mortality risk in this population. We aimed to develop and validate a clinically usable prediction model for OC mortality in men with prostate cancer using two prospective national cohorts.

**Methods**:Model training was performed using the National Health and Nutrition Examination Survey (NHANES), a cross-sectional survey. We included men ages > 40 from 1999-2010, with mortality follow-up through 2014. We built a Cox proportional hazards model with all-cause mortality as the outcome. The model was validated in the Prostate, Lung, Colon, and Ovarian (PLCO) Cancer Screening Trial with OC mortality as the outcome.

**Findings**: The NHANES training data included 2,420 men with 459 deaths during follow-up. The PLCO validation data included 8,220 men over age 55 diagnosed with prostate cancer, of whom 2,415 died of other causes. In the validation cohort, our model had a time-dependent AUC of 0.75 and 0.78 at 10 and 15 years. The final model included 8 predictors: age, education level, marital status, diabetes, hypertension, stroke, BMI, and smoking status.

**Interpretation**: We have developed and validated an OC mortality model that can be used in clinic for men with prostate cancer. Our OC mortality prediction model uses 8 easily obtained predictors and shows comparable or superior performance to more complex OC mortality prediction models already in existence.

**Funding**: National Science Foundation (DGE-1256260)

Abstract (max 275): 259

Main text (max 3000): 3862

Figures + tables (max 6): 6

**Introduction**

In 2019, there were 174,650 new cases of prostate cancer in the U.S. However, most men diagnosed with prostate cancer will not die of their cancer, instead dying of old age or other comorbidities [1]. Because competing risks of other-cause (OC) mortality are substantial, they may render prostate cancer treatment or treatment intensification futile. As a result, prostate cancer treatment guidelines recommend that clinicians take the patient's risk of OC mortality and life expectancy into account when making treatment decisions.

Little guidance on how to estimate OC life expectancy in prostate cancer patients has been provided. The National Comprehensive Cancer Network (NCCN) recommends that clinicians use Social Security Administration (SSA) actuarial tables to assess patient life expectancy [2]. The SSA tables have several weaknesses. First, research has shown that prostate cancer patients with localized disease usually live 4-6 years longer than SSA predictions would suggest, while the SSA predictions overestimate the life expectancy of patients with distant disease [3]. Second, the SSA tables do not adjust for patient comorbidities, although the life expectancy of a 70-year-old man may vary by more than 11 years from his SSA predicted expectancy depending on his comorbidity burden [4]. To address these problems, the NCCN advised that clinicians combine SSA estimates with their clinical assessment of the patient's comorbidity burden.

Unfortunately, research suggests that clinical intuition for estimating comorbidity burden is poor [3]. In two studies that studied physician performance for estimating prostate cancer patient life expectancy, both found that physicians correctly predicted whether a patient would be alive in ten years only 68% of the time, suggesting that as many as 1 in 3 patients could receive inappropriate treatment due to misestimation of their life expectancy [5, 6]. There was substantial heterogeneity between physicians, with predictive performance across individual physicians ranging from 52% to 78%, another weakness of relying on physician opinion [6]. These incorrect estimates of patient life expectancy may have major ramifications on treatment plan: Daskivich et al. report a high rate of overtreatment in patients with low-risk prostate cancer and high comorbidity burden, which may have negative effects on patient overall mortality and quality of life [7].

In response to these challenges, researchers have worked to build and validate OC mortality prediction tools in men with prostate cancer. Based on a literature review, we identified four OC mortality prediction models specifically for men with prostate cancer that provided survival predictions and incorporated comorbidity information in some capacity [8, 9, 10, 11]. However, we had concerns about all four models. Hoffman et al. (2015) built a model with C-index of 0.73 using demographic information, treatment information, Gleason score, and self-reported health status [8]. We questioned if model performance could be further improved by incorporating more comorbidity information beyond self-reported health status, and the inclusion of treatment information as a predictor makes it difficult to use the tool for treatment decision-making because of the confounding between treatment assignment, OC mortality risk, and prostate cancer specific (PC) mortality risk. Daskivich et al. (2015) built a model that age-adjusted the Charlson Comorbidity Index (CCI) and calibrated it to men with prostate cancer, obtaining a C-index of 0.77 in the veteran population [9]. Daskivich’s model requires clinicians to input 21 comorbidities to calculate the CCI, in addition to age, race, primary treatment, date of diagnosis, and AJCC stage, which seemed burdensome, and we wanted to incorporate comorbidities without the constraint of a pre-formulated comorbidity score.

Kent et al. (2016) combined a comorbidity-life expectancy tool built in a British private insurance claims database with age-shifted SSA tables; their model had a C-index of 0.73 [10]. We had concerns about 1) Kent et al.’s decision to sum the odds ratios from multiple univariable models rather than fitting a multivariable model, 2) their three-year SSA shift approach, and 3) whether a British private insurance population is representative of the general U.S. patient population. In addition, the model requires 19 specialized predictors, many of which may not be known by the average patient, or even appear in his medical record, thus necessitating additional diagnostic workup in order to use the model. Finally, Riviere et al. (2019) built an OC mortality prediction model in SEER-Medicare claims data, reporting a of 0.6. Their final model involved 143 covariates, which we thought would be burdensome in a clinical setting [11]. We were also concerned by their use of SEER-Medicare claims data, which may produce models that do not generalize well to non-claims settings. For example, health conditions like obesity and smoking are often underdiagnosed in claims data, because they do not generate an insurance claim unless very severe. This can lead to unexpected biases and generalizability issues.

We seek here to develop a clinically usable, validated prediction model for OC mortality in US men with prostate cancer to help personalize treatment decision making. We use twelve years of data from the National Health and Nutrition Examination Survey (NHANES) to build our prediction model and validate it in the prostate cohort of the Prostate, Lung, Colon, and Ovarian Cancer Screening Trial (PLCO). Our final model is parsimonious, does not rely on treatment information, and is built and validated in a diverse sample that reflects most U.S. patients with localized disease assessed in a clinical setting. We have made it available for public use in app format at https://elizabethchase.shinyapps.io/ocm\_app/.

**Methods**

*Data Acquisition*

Our training data come from NHANES, a cross-sectional survey conducted every two years in the U.S. since 1999. NHANES is nationally representative of the U.S. non-institutionalized civilian population and contains information on patient demographics and health. More information on NHANES' sample construction and data collection protocols can be found at the survey's website [13]. Mortality follow-up on all NHANES participants is provided through a linkage with the National Death Index, with mortality follow-up through Dec. 31, 2014. Critically, NHANES is not limited to prostate cancer patients, but instead represents the entire U.S. population.

We used NHANES data from 1999-2010 and restricted to men older than 40 free of non-prostate cancer with complete data for all covariates used in model-building: age, race, educational attainment, marital status, veteran status, insurance status; diagnosis of anemia, angina, arthritis, asthma, chronic bronchitis, coronary heart disease, congestive heart failure, diabetes, emphysema, high cholesterol, hypertension, kidney issues, liver disease, mental health concern, or myocardial infarction; and alcohol use, regular access to healthcare, whether the patient had been hospitalized in the past year, BMI, and smoking status. More information on variable definition is provided in the supplementary materials; the training data can be reconstructed using the code provided at https://github.com/elizabethchase/PCOtherCause.

*Model Building*

We considered three classes of models in initial model building: Cox proportional hazards models [14], survival random forests [15], and parametric cubic spline hazard models [16]. In all cases we used overall mortality as our outcome (as opposed to OC-specific mortality) because most patients in our population did not have prostate cancer, and the rate of PC mortality in the general population is low (19.1 per 100,000 annually [17]). Because PC mortality comprises only a small portion of overall mortality in men with localized prostate cancer, we thought that OC mortality could be approximated by overall mortality. We did sensitivity analyses, explained further below, to further support this choice. We used the NHANES survey weights when performing variable selection for the Cox models, but otherwise did not weight our analysis. Including the survey weights complicated model performance assessment and is beyond the scope of this paper. All analyses were performed in R 3.6.2 [18].

Initial checks suggested that model performance would not suffer substantially from restricting to variables that also appeared in our validation data, so we focused on age, race, educational attainment, marital status, arthritis, chronic bronchitis, coronary heart disease, diabetes, emphysema, hypertension, BMI, and smoking status. We considered clinically relevant interactions and flexible functional forms for the effect of age. Using a combination of clinical knowledge, statistical significance, and assessment of cross-validated C-index, we built separate models using each of our three modeling strategies in the NHANES data, ultimately settling on a Cox proportional hazards model as our leading candidate model.

We conducted sensitivity analyses to assess whether our OC mortality prediction model developed in a non-prostate cancer population would translate to a prostate cancer patient population. In addition to considering prostate cancer diagnosis, years since diagnosis, and relevant interactions with these covariates as predictors in all models, we examined relationships between the linear predictor of our final models and prostate cancer diagnosis and years since diagnosis (all in NHANES data). These analyses all suggested that having prostate cancer had little effect on predictions for overall mortality. Despite its minimal impact on model performance, we included prostate cancer as a predictor in all models in order to adjust for its effect.

After completing model-building but before advancing to external validation, we re-estimated model effects and the baseline hazard in the training data using a larger sample of patients with complete data for the final group of predictors. Modeling code is provided at https://github.com/elizabethchase/PCOtherCause.

*Model Validation*

We validated our model in the Prostate, Lung, Colon, and Ovarian Cancer Screening Trial (PLCO) prostate cohort. PLCO enrolled 155,000 participants between November 1993 and July 2001, with mortality follow-up through Dec. 31, 2009. In order to enroll in the trial, participants had to be ages 55-74, with no history of prostate, lung, colon, or ovarian cancer, and with prostate, lungs, colon, and ovaries intact. Additionally, participants could not be receiving treatment for any cancer, and they could not have received screening for prostate, lung, colon, or ovarian cancer in the past year. Patients who met trial inclusion criteria completed a medical history questionnaire at baseline and were randomized to receive screenings for prostate, lung, and colon cancers if male. If a screening indicated a potential cancer, patients received diagnostic workup, which was included in the PLCO data along with their primary treatment information. For more information on the PLCO protocol and data, please see the trial's website [19].

For validation, we restricted the PLCO sample to men who had been diagnosed with prostate cancer and had complete data for the covariates in our candidate model. To obtain cause-specific performance estimates, we censored men who died of prostate cancer at their time of death. Cause-specific performance was our primary interest, because we wanted to build a cause-specific model, but we also considered performance in the competing risks setting, by setting the censoring time of men who died of prostate cancer at 100,000 months and using that as our endpoint for performance calculation. Using our candidate model and the baseline hazard estimated from NHANES, we obtained survival predictions for all men in PLCO and compared these predictions to their true mortality outcomes. We calculated the inverse probability of censoring weighted (IPCW) time-dependent AUC to assess accuracy and obtained bootstrapped confidence intervals for it. We also calculated the IPCW C-index. We calculated the median survival (truncated at 15 years) as an estimate of life expectancy and made calibration plots to assess the calibration of these life expectancy predictions to observed patient life expectancy. Additionally, we obtain OC mortality predictions from the 2001 SSA actuarial life tables, the National Vital Statistics System’s (NVSS) life expectancy estimates, and a Cox model using only age as a predictor. We compare the IPCW time-dependent AUC, bootstrapped confidence intervals, and life expectancy calibration of these models to our final candidate model in order to better understand the importance of comorbidity/demographic information for OC mortality predictions. Validation code is provided on GitHub; the validation data are not publicly accessible.

**Results**

*Model Building and Assessment*

After restricting the NHANES data to patients who met our inclusion criteria (Supplementary Figure 1), we were left with a training sample of 2,420 men, of whom 459 died over a median of 109.0 months (9.1 years) of follow-up. Of these deaths, 111 (24.1%) were from malignant neoplasms. Characteristics of the sample are given in Table 1. Mean age was 59.4 years. 127 patients (5.2%) had been diagnosed with prostate cancer at the time of survey collection; of these patients, 18 had died of malignant neoplasms and 35 had died of other causes by the end of follow-up. Almost two-thirds of the sample were current or former smokers, and more than 75% of the sample was overweight or obese. The PLCO external validation data consisted of 8,220 patients with complete data for all predictors (Supplementary Figure 2). PLCO patients were markedly different from NHANES patients (Table 1). PLCO patients were older, more likely to be white, more educated, more likely to be married, and generally healthier than NHANES patients. The only characteristic on which NHANES and PLCO patients were not significantly different was previous heart attack or diagnosis of coronary heart disease; roughly 12% of patients had this diagnosis in both samples.

Our final Cox model included eight predictors (age, diabetes, education, hypertension, marital status, smoking status, previous stroke, and BMI) and interactions between age and diabetes, age and education, age and hypertension, and age and previous stroke. After selecting this model, we refit the model in an expanded NHANES training sample of 7,369 men. Model effect estimates for the final Cox OC mortality model are given in Figure 1. Increased age, diabetes, hypertension, current smoking, previous stroke, and non-normal BMI all had harmful effects on OC mortality, while increased education and being married were protective. All model main effects were significant. When we consider the proportion of explainable log-likelihood contributed by each predictor, we find that age is the most important predictor, explaining 76.3% of the likelihood, followed by smoking status (5.4%) and marital status (4.0%), then stroke, education, diabetes, hypertension, and BMI. The least important predictor is prostate cancer (0.1%). We also present the survey-weighted Cox OC mortality model in Supplementary Figure 3, for reference.

In external validation, our Cox model had a cause-specific IPCW AUC of 0.75 at 10 years and 0.78 at 15 years (Figure 1). The IPCW C-index was 0.70. For comparison, the SSA life tables, NVSS life expectancy estimates, and age-only Cox model all produced IPCW AUCs of 0.71 and 0.75 at 10 and 15 years, respectively. The IPCW AUCs of our Cox model and the age-only methods were significantly different at 10 years, suggesting that including comorbidity information does improve OC mortality predictions compared to relying only on age. Estimates of the IPCW AUC for competing risks are given in Supplementary Table 1; performance was largely similar to performance in the cause-specific setting. Model performance was also similar across prostate cancer treatment groups (Supplementary Table 2).

Calibration of the final Cox model with baseline hazard estimated in NHANES (NHANES-Calibrated Cox), the final Cox model with baseline hazard re-estimated in PLCO (PLCO-Calibrated Cox), and the SSA actuarial table predictions is given in Figure 2. We see that the linear predictor of our model performs very well; calibration of the PLCO-Calibrated Cox model is excellent. However, we see evidence that the NHANES population is poorly calibrated to PLCO. NHANES patients are at increased risk of OC mortality compared to the PLCO population; despite having the same linear predictor, the NHANES-Calibrated Cox model shows pessimism for predicting patient life expectancy in PLCO by about 1.5 years. This pessimism is even more pronounced in the SSA life tables, which generally underestimate life expectancy in PLCO by about 4 years.

*Other Cause Mortality Risk Drivers*

Table 2 presents a cross-tabulation of predicted OC life expectancy with patient age among PLCO patients in our sample. We see that although age is the primary driver—no men in PLCO ages 55-64 had an OC life expectancy of less than 5 years—there is still substantial variation in life expectancy within age bracket, particularly in patients ages 65-84. In patients ages 75-84, 39.8% have life expectancy less than 10 years, while 60.2% have life expectancy greater than 10 years, and even among younger patients ages 65-74, 3.8% of patients have life expectancy less than 10 years. In the oldest age bracket (85+), incorporating comorbidity information may still be useful for dealing with high risk prostate cancers, in which the NCCN recommends considering 5 years of life expectancy as a cut-point for more aggressive treatment, as roughly 20% of men ages 85+ have life expectancy less than 5 years.

More patient characteristics, split by predicted OC life expectancy, can be found in Supplementary Table 2. We see that patients with reduced OC life expectancy are more likely to be black and other race. Graduating from college is highly protective for OC mortality, as is being married. Current smokers and former smokers are overrepresented in the high OC mortality risk group, as are underweight and obese patients. Despite not being considered in modeling, prostate cancer characteristics were also linked with OC mortality predictions. Patients with more aggressive cancers were overrepresented in the high OC mortality risk group, while patients with less aggressive cancers generally had reduced OC mortality risk. Patients in the high OC mortality risk group were also more likely to die of their prostate cancer. Sample sizes for aggressive cancers are quite low, though, given the nature of the PLCO trial, so conclusions about these patients must be made with care. OC mortality predictions can be further explored using the app given at https://elizabethchase.shinyapps.io/ocm\_app/.

In fact, when we use the linear predictor from our final Cox model as a predictor for PC mortality in PLCO, it still has a time-dependent cause-specific AUC of 0.72 at 15 years, despite not including any prostate cancer specific factors other than age.

*OC Mortality Risk and Treatment Assignment*

Figure 4 shows the proportion of patients in PLCO who received a particular treatment grouped by prostate cancer risk and OC life expectancy. We see evidence that both OC mortality and prostate cancer risk were taken into account when selecting treatment,as patients with reduced life expectancy are overall less likely to receive aggressive treatment. In the active surveillance, radiation + hormone, and prostatectomy categories, the proportion of patients receiving treatment is almost perfectly ordered in concordance with life expectancy, with low life expectancy patients most likely to receive active surveillance and least likely to receive prostatectomy or radiation + hormone. We see that intermediate OC mortality risk patients are most likely to receive radiation alone, particularly those with lower risk prostate cancer.

**Discussion**

We present a parsimonious and accurate prediction model for OC mortality in US prostate cancer patients, built and tested in two diverse, prospective national cohorts. Our model shows excellent discrimination and calibration despite being validated in a population that is very different from its training population. It requires only eight easily obtained predictors and can be accessed at https://elizabethchase.shinyapps.io/ocm\_app/. It joins a growing list of OC mortality prediction models for men with prostate cancer; however, our model stands out for its simplicity, usability and impressive predictive performance in diverse patient populations. We make three key contributions to the discussion of OC mortality in prostate cancer:

1. We estimate OC life expectancy to allow clinicians to follow NCCN guidelines.
2. When combined with cause-specific estimates of PC mortality (like the recent STAR-CAP model of Dess et al. [20]), we can provide personalized estimates of cumulative incidence for both PC and OC mortality, potentially under different treatment paradigms.
3. Because of our model’s simplicity and accessibility, it is a resource for future research on OC mortality in prostate cancer patients.

When building the model, we considered a long list of potential predictors and several modeling strategies. The statistical methods used to produce the model are rigorous and do not rely on intermediate risk scores. We present both life expectancy and risk predictions over time, rather than just point estimates. These features make it attractive for future statistical work, and by providing all training data and code used to produce the model, our model can be combined with other prediction tools and treatment effect estimates to build new models. For example, following the approach of Pfeiffer and Gail [21], we can combine our cause-specific OC mortality hazard estimates with cause-specific PC mortality hazard estimates to estimate the cumulative incidence of both OC and PC mortality.

We have also made it easy for future researchers to reproduce our analysis and perform similar analyses for new disease sites and applications. This is particularly important because the approach we use here—building an OC mortality model for cancer patients in a non-cancer patient study—may be generalizable to other cancer patient populations where the cancer-specific mortality is low, and using cohorts likes NHANES vastly increases the number and quality of the OC mortality predictors that can be considered as candidates. Further research is needed, but this idea may be useful to researchers working on breast cancer, head and neck cancer, and other disease sites where a substantial proportion of patients die of OC mortality.

We also provide evidence that incorporating comorbidity information is beneficial to predicting OC mortality, although these gains were modest on the AUC metric. Our OC mortality model significantly outperformed the SSA life tables, NVSS life expectancy estimates, and an age-only Cox model. However, in terms of absolute performance the age-only predictors do still perform fairly well and provide a good benchmark for future models.

Our study is subject to several limitations. Although it has its strengths, building a cancer OC mortality prediction model in a non-cancer patient population could introduce bias. We addressed this limitation through sensitivity analyses of the effect of prostate cancer within our model, and we validated our model in a prostate cancer population, so we believe the risk of bias is low. However, this approach limits our ability to obtain predictions of OC cumulative incidence without refitting the model within PLCO or combining our model with other prostate cancer specific mortality prediction tools. Some may argue that our use of a cause-specific endpoint instead of a more traditional competing risks endpoint is a weakness, necessitated by NHANES not being a prostate cancer patient population. However, we argue that the cause-specific endpoint is more reflective of clinical use (i.e. when the NCCN refers to OC mortality life expectancy, they refer to OC mortality *in the absence* of prostate cancer), and we wanted a cause-specific model in order to provide integrated cumulative incidence predictions in future work. Despite being a cause-specific model, our model still appears to perform well on the competing risks metrics presented in the Supplement. As with all cause-specific research, ascertainment of cause of death is a challenge. We sidestep this issue somewhat during model-building by using an overall mortality endpoint, but in PLCO we validate for an OC mortality endpoint. PLCO assessed cause of death using annual patient surveys and a linkage with the National Death Index. Uncertain deaths were assessed by an adjudication board. Despite this, some PLCO participants may have been incorrectly assigned to OC or PC mortality endpoints.

*Conclusions*

We have built and validated an accurate and succinct model for OC mortality in US prostate cancer patients that is ready for clinical use. Our model will improve quality of care when patient life expectancy is a factor in treatment decision-making. The generalizability of our approach and transparency of our documentation will aid future research on cancer OC mortality across disease sites.

**References**

1. *American Cancer Society*. 2019. Key Statistics for Prostate Cancer. https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html.
2. National Comprehensive Cancer Network. 2020. NCCN Guidelines Version 2.2020: Prostate Cancer.
3. JD Sammon, F Abdollah, A D'Amico, M Gettman, A Haese, N Suardi, A Vickers, QD Trinh. 2015. Predicting life expectancy in men diagnosed with prostate cancer. *European Association of Urology* 68: 756-65.
4. C Jeldres. 2012. Life expectancy estimation in prostate cancer patients. *Canadian Urological Association Journal* 6.5: 374-5.
5. KMYB Leung, WM Hopman, J Kawakami. 2012. Challenging the 10-year rule: the accuracy of patient life expectancy predictions by physicians in relation to prostate cancer management. *Canadian Urological Association Journal* 6.5: 367-73.
6. J Walz, A Gallina, P Perrotte, C Jeldres, QD Trinh, GC Hutterer, M Traumann, A Ramirez, SF Shariat, M McCormack, JP Perreault, F Bénard, L Valiquette, F Saad, PI Karakiewicz. 2007. Clinicians are poor raters of life-expectancy before radical prostatectomy or definitive radiotherapy for localized prostate cancer. *BJU International* 100.6.
7. TJ Daskivich, K Chamie, L Kwan, J Labo, R Palvolgyi, A Dash, S Greenfield, MS Litwin. 2010. Overtreatment of men with low-risk prostate cancer and significant comorbidity. *Cancer* 117.10: 2058-66.
8. RM Hoffman, T Koyama, PC Albertsen, MJ Barry, TJ Daskivich, M Goodman, AS Hamilton, JL Stanford, AM Stroup, AL Potosky, DF Penson. 2015. Self-reported health status predicts other-cause mortality in men with localized prostate cancer: results from the prostate cancer outcomes study. *Journal of General Internal Medicine* 30.7: 924-34.
9. TJ Daskivich, L Kwan, A Dash, C Saigal, MS Litwin. 2015. An age adjusted comorbidity index to predict long-term, other cause mortality in men with prostate cancer. *Journal of Urology* 194: 73-78.
10. M Kent, DF Penson, PC Albertsen, M Goodman, AS Hamilton, JL Stanford, AM Stroup, B Ehdaie, PT Scardino, AJ Vickers. 2016. Successful external validation of a model to predict other cause mortality in localized prostate cancer. *BMC Medicine* 14.25.
11. P Riviere, C Tokeshi, J Hou, V Nalawande, R Sarkar, AJ Paravat, M Schiaffino, B Rose, R Xu, JD Murphy. 2019. Claims-Based Approach to Predict Cause-Specific Survival in Men With Prostate Cancer. *JCO Clinical Cancer Informatics* 3: 1-7.
12. A Tan, YF Kuo, JS Goodwin. 2013. Predicting life expectancy for community-dwelling older adults from Medicare claims data. *American Journal of Epidemiology* 178.6: 974-83.
13. National Center for Health Statistics, Centers for Disease Control and Prevention. 2019. National Health and Nutrition Examination Survey Data, 1999-2010. https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx
14. DR Cox. 1972. Regression models and life tables. *Journal of the Royal Statistical Society Series B (Methodological)* 34.2: 187-220.
15. H Ishwaran, UB Kogalur, EH Blackstone, MS Lauer. 2008. Random survival forests. *Annals of Applied Statistics* 2.3: 841-60.
16. P Royston, MKB Parmar. 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 21: 2175-97.
17. U.S. Cancer Statistics Working Group. U.S. Cancer Statistics Data Visualizations Tool, based on 2019 submission data (1999-2017): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; [www.cdc.gov/cancer/dataviz](https://www.cdc.gov/cancer/dataviz), released in June 2020.
18. R Core Team, R Foundation for Statistical Computing. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. https://www.R-project.org/
19. Division of Cancer Prevention, National Cancer Institute. 2019. Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO). https://prevention.cancer.gov/ major-programs/prostate-lung-colorectal-and-ovarian-cancer-screening-trial
20. Dess RT, Suresh K, Zelefsky MJ, Freedland SJ, Mahal BA, Cooperberg MR, Davis BJ, Horwitz EM, Terris MK, Amling CL, Aronson WJ, Kane CJ, Jackson WC, Hearn JWD, DeVille C, DeWeese TL, Greco S, McNutt TR, Song DY, Sun Y, Mehra R, Kaffenberger SD, Morgan TM, Nguyen PL, Feng FY, Sharma V, Tran PT, Stish BJ, Pisansky TM, Zaorsky NG, Moraes FY, Berlin A, Finelli A, Fossati N, Gandaglia G, Briganti A, Carroll PR, Karnes RJ, Kattan MW, Schipper MJ and Spratt DE. 2020. Development and Validation of a Clinical Prognostic Stage Group System for Non-Metastatic Prostate Cancer: Disease Specific Mortality Results from the International Staging Collaboration for Cancer of the Prostate (STAR-CAP). JAMA Oncology.
21. Pfeiffer RM and MH Gail. *Absolute Risk: Methods and Applications in Clinical Management and Public Health.* London: Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 2017.

**Supplementary Materials**

*Variable Definitions*

We constructed some of our variables from other NHANES variables. In particular:

1. We defined a patient as having hypertension if they reported a previous diagnosis of hypertension or if their blood pressure reading was hypertensive (we followed standard NHANES guidelines on constructing the blood pressure reading).
2. We defined a patient as having diabetes if they reported a previous diagnosis of diabetes or if their blood glucose met clinical criteria for diabetes.
3. We defined a patient as having high cholesterol if they reported high cholesterol, or if their blood work met clinical criteria for high cholesterol.
4. We defined patients as current smokers if they had smoked 100 cigarettes in their lifetime and reported smoking cigarettes every day or some days at present; we defined patients as former smokers if they had smoked 100 cigarettes in their lifetime and reported smoking not at all at present; we defined patients as never smokers if they had not smoked 100 cigarettes in their lifetime.
5. We defined being overweight as having BMI greater than or equal to 25 and less than 40 and being obese as having BMI greater than or equal to 40.
6. We collapsed marital status into three categories: married if the patient was currently married or living with a partner, separated if the patient was widowed, divorced, or separated from their partner, and single if the patient had never married.
7. We collapsed race into non-Hispanic white, non-Hispanic black, and all other races.

For all other variables, we used unmodified results from the NHANES questionnaires about demographics, health insurance, health care access, mental health, and medical history. For more information on variable definition and construction, please see the repository posted at: https://github.com/elizabethchase/PCOtherCause.