Claims-Based Approach to Predict Cause-Specific Survival in Men With Prostate Cancer

Paul Riviere¹; Christopher Tokeshi²; Jiayi Hou, PhD¹; Vinit Nalawade, MS¹; Reith Sarkar¹; Anthony J. Paravati, MD¹; Melody Schiaffino, PhD, MPH³; Brent Rose, MD¹; Ronghui Xu, PhD¹; and James D. Murphy, MD¹

PURPOSE Treatment decisions about localized prostate cancer depend on accurate estimation of the patient's life expectancy. Current cancer and noncancer survival models use a limited number of predefined variables, which could restrict their predictive capability. We explored a technique to create more comprehensive survival prediction models using insurance claims data from a large administrative data set. These data contain substantial information about medical diagnoses and procedures, and thus may provide a broader reflection of each patient's health.

METHODS We identified 57,011 Medicare beneficiaries with localized prostate cancer diagnosed between 2004 and 2009. We constructed separate cancer survival and noncancer survival prediction models using a training data set and assessed performance on a test data set. Potential model inputs included clinical and demographic covariates, and 8,971 distinct insurance claim codes describing comorbid diseases, procedures, surgeries, and diagnostic tests. We used a least absolute shrinkage and selection operator technique to identify predictive variables in the final survival models. Each model's predictive capacity was compared with existing survival models with a metric of explained randomness (p^2) ranging from 0 to 1, with 1 indicating an ideal prediction.

RESULTS Our noncancer survival model included 143 covariates and had improved survival prediction ($p^2 = 0.60$) compared with the Charlson comorbidity index ($p^2 = 0.26$) and Elixhauser comorbidity index ($p^2 = 0.26$). Our cancer-specific survival model included nine covariates, and had similar survival predictions ($p^2 = 0.71$) to the Memorial Sloan Kettering prediction model ($p^2 = 0.68$).

CONCLUSION Survival prediction models using high-dimensional variable selection techniques applied to claims data show promise, particularly with noncancer survival prediction. After further validation, these analyses could inform clinical decisions for men with prostate cancer.

Clin Cancer Inform. © 2019 by American Society of Clinical Oncology

INTRODUCTION

Survival prediction plays a central role in prostate cancer. The question of whether to offer screening, although controversial, depends on a patient's life expectancy. With treatment, current clinical guidelines for the management of prostate cancer hinge on the estimation of two factors: aggressiveness of a patient's tumor and the patient's overall life expectancy.^{1,2} The exercise of estimating life expectancy in prostate cancer attempts to identify patients with a life expectancy long enough to benefit from active treatment.

Life expectancy with prostate cancer depends on both the risk of cancer-specific mortality and the risk of noncancer mortality. Among men with localized prostate cancer, the risk of dying of noncancer causes often substantially outweighs the risk of dying of prostate cancer.³ Despite this critical need for noncancer survival prediction, we lack accurate clinical prediction

tools. For example, the current National Comprehensive Cancer Network clinical guidelines for prostate cancer rely on a rudimentary technique of using actuarial tables from the Social Security Administration and inflate or deflate this estimate on the basis of a subjective assessment of the patient's health.¹

The current era of electronic health records (EHRs) offers a unique opportunity for survival prediction modeling. EHR data contain an indirect story of a patient's medical history, and accessing and processing these data may allow for an estimation of a patient's risk of death. In this study, we explored the predictive utility of using insurance claims to estimate survival. Older claims-based approaches exist,^{4,5} though these use only a small fraction of preselected health conditions to estimate risk. Here, we tested a high-dimensional approach to survival prediction that uses all available claims data to create a prediction model. The purpose of this study was to measure the

ASSOCIATED CONTENT

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on

December 7, 2018

and published at

[ascopubs.org/journal/](https://ascopubs.org/journal/cci)

cci on March 4, 2019;

DOI <https://doi.org/10.1200/CCI.18.00111>

1200/CCI.18.00111

CONTEXT

Key Objective

Can billing claims data be used to generate models predicting cause-specific survival in the prostate cancer setting?

Knowledge Generated

We demonstrated that in the noncancer mortality setting, a high-dimensional shrinkage approach can risk stratify patients better than comparator comorbidity indices (Charlson and Elixhauser). Additionally, this approach can identify patients at high risk of dying from comorbid conditions who had been labeled as low-risk by these indices, offering better evaluation of patients with noncanonical comorbidities. In contrast, applying this approach to prostate-cancer specific mortality resulted in similar variable selection to existing models, and virtually identical model performance.

Relevance

Claims data may offer an unbiased representation of patients' health status, and the use of shrinkage techniques in predictive modeling may improve comorbidity risk stratification. In prostate cancer, specifically, improved characterization of cancer versus noncancer predicted mortality could lead to improved clinical decision making.

accuracy of this survival prediction approach in a large cohort of Medicare beneficiaries with prostate cancer and to compare this new approach against standard existing prediction models.

METHODS

Data Source

We identified Medicare beneficiaries with prostate cancer from within the SEER cancer registry. The National Cancer Institute manages the SEER program, which pools data from individual cancer registries from across the United States. SEER covers 28% of the population and provides a diverse cohort of patients that approximates the demographics of the United States. Medicare provides federally funded health insurance for individuals older than 65 years. Medicare data include information on inpatient hospitalizations, outpatient visits, diagnoses, procedures, and surgeries. The SEER-Medicare linkage provides Medicare claims for all beneficiaries within SEER. As a result, this population-based data set provides the opportunity to study longitudinal patterns of care from before cancer diagnosis through death.⁶

Study Population

This study included men within the SEER-Medicare database who were at least 66 years of age with histologically confirmed nonmetastatic prostate cancer diagnosed between January 2004 and December 2009. We sought to estimate survival at the time of diagnosis and construct a survival prediction model based on Medicare claims during the year before cancer diagnosis. Therefore, patients were required to have complete follow-up data including continuous Medicare Part A and B coverage in the year before cancer diagnosis. Medicare Part C includes managed care organizations that do not routinely submit detailed claims to Medicare; therefore, patients with Part C were excluded. Overall, 57,011 patients met the inclusion criteria and were included in this study.

Survival Model Covariates

We created high-dimensional survival prediction models that used all available demographic, cancer-related, and claims-based variables to create a patient-specific risk score to separately estimate the risk of cancer-related and noncancer-related survival. Demographic variables included age, race, marital status, geographic region, and population density. Cancer-related variables included prostate-specific antigen (PSA) level, American Joint Committee on Cancer T stage, Gleason score, and year of diagnosis.

The specific claims codes evaluated in the prediction model included International Classification of Diseases (ninth revision) diagnosis and procedure codes and Healthcare Common Procedure Coding System codes. These codes represent a common language used across different EHR systems; therefore, building predictive models based on these codes allows for translation across different platforms. Survival prediction was measured from time of diagnosis, and the survival prediction model included claims during the year before diagnosis. Initial query of the data revealed 22,297 unique claim codes. We included only claims that appeared in more than 10 patients, leaving 8,971 unique claims as potential predictors in the survival model. The presence of a claim in the year before diagnosis was treated as a binary predictor. Repeated claims were not included in the model.

Survival Models and Risk Score

Details of model construction have been published elsewhere.⁷ Briefly, we used a high-dimensional variable selection technique to build a cause-specific hazard model to predict noncancer survival and cancer-specific survival. Cancer-specific death was defined as death from any malignancy, and noncancer death was death from any noncancer cause. We randomly divided our data into a training data set ($n = 28,505$) to construct our survival models and a test data set ($n = 28,506$) to measure performance.

We first conducted a univariate screen of the demographic, cancer-related, and 8,971 claim variables with Cox regression models to determine possible associations with noncancer survival and cancer-specific survival. Second, we used a least absolute shrinkage and selection operator (LASSO) with a Cox regression technique to identify predictive variables in the final survival models. The LASSO technique reduces the high-dimensional covariate pool down to an optimal set of predictive variables using cross-validation to optimize model fitting. The predictability of each model was assessed with ρ^2 , which represents the proportion of the explained randomness.⁸ These ρ^2 values range from 0 to 1, with a value of 1 representing a perfect prediction. Finally, we constructed a risk score for individual patients that consisted of a linear predictor of the β coefficients from each of the survival models. We categorized each risk score according to quintiles.

Although it is well understood that both risks contribute to incidence rate of each causal of survival, our previous work⁷ showed that for these data, the cancer-specific risk played a leading role in predicting cancer-specific survival, and the noncancer-specific risk played a leading role in predicting noncancer-specific survival. Therefore, to formulate a clinically easy-to-use risk score, we used cumulative incidence plots to depict the categories of each cause-specific survival risk score. The predictor variables to build the individual risk scores arose from analysis on the training data set. The performance of these risk scores using cumulative incidence plots was assessed on the test data set.

Comparator Prediction Models

We compared noncancer survival predictions against the Charlson comorbidity index (CCI)⁹ and Elixhauser comorbidity index (ECI). The CCI represents a composite index of 19 diseases, which together generate an individual score for patients. This index represents a standard proxy used to estimate the risk of noncancer mortality and is used extensively in health services research and in clinical medicine. The analysis here used the Deyo adaptation of the CCI.^{10,11} The ECI is a composite score of 30 comorbidity measures¹²; it was developed in a noncancer cohort, though it predicts survival among patients with cancer.¹³ With cancer-specific survival, we compared our survival predictions to the approach put forth by Memorial Sloan Kettering Cancer Center (MSKCC), which represents a validated and widely used approach to estimating the risk of cancer mortality.¹⁴ The MSKCC Prostate Cancer nomogram depends on tumor characteristics including clinical T stage, Gleason score, and PSA level.

RESULTS

Noncancer-Specific Mortality

The initial screen of 8,991 variables identified 20 predictive demographic or clinical variables, and 2,188 predictive claim codes. The second stage of variable selection identified 143 predictor variables in the final predictive

model for noncancer survival. The predictor variables in the final model included four (2.8%) demographic variables, two (1.4%) clinical or cancer-related variables, and 137 (95.8%) claims-related variables (Table 1). The majority of claim-based variables were comorbid diagnoses (61%), though these data also included procedures (15%) and diagnostic tests (9%). A complete list of all predictor codes is provided in the Data Supplement. Overall, 19 of the predictive variables in the final model were associated with a reduced risk of noncancer death (negative β coefficient), and the remaining 124 variables were associated with an increased risk of noncancer death (positive β coefficient).

Figure 1A demonstrates the cumulative incidence of noncancer death for patients according to their risk-score quintile. The 5-year cumulative incidence of noncancer death was 1.9% for patients in the lowest risk group, compared with 31.5% for patients in the highest risk group. Our risk model demonstrated mixed degrees of overlap when compared with the CCI and ECI (Fig 2). The majority of patients in our low-risk quintile also had a CCI score of zero. However, among those in our highest risk group, 44% had a CCI score of zero. Among patients with a CCI score of zero, we found that our risk score continued to provide reasonable survival discrimination (Fig 3A). A similar pattern emerged when considering overlap between our risk model and the ECI (Fig 3B). In all LASSO risk-score quintiles, low-risk ECI scores (ie, scores ≤ 0) described most patients.

When assessing model accuracy for predicting noncancer survival, our claims-based risk model had a ρ^2 value of 0.60, which was substantially higher than the CCI, which had a ρ^2 value of 0.26, and the ECI, which had a ρ^2 value of 0.26.

Cancer-Specific Mortality

For cancer-specific survival, the initial screen identified 1,079 predictive variables, and the second stage of variable selection yielded nine predictor variables in the final

TABLE 1. Summary of Variables in Prediction Models

Classification of Variables	Variables in Prediction Model, No. (%)	
	Noncancer Survival	Cancer Survival
Demographic	4 (3)	1 (11)
Cancer related	2 (1)	2 (22)
Claims related		
Diagnosis of comorbidity	87 (61)	5 (56)
Diagnosis of operation	21 (15)	—
Diagnostic test	13 (9)	—
Other	16 (11)	1 (11)
Total	143	9

NOTE. Dash indicates no variables of this type were selected in the final model.

FIG 1. Plots representing (A) the cumulative incidence of noncancer-related death and (B) cancer-related death among patients with prostate cancer, stratified by risk-score (RS) quintile.

survival prediction model. The predictor variables in the final model included classic clinical and cancer-related predictors including age, Gleason score, and PSA level, as well as six claims-based variables (Table 1). The predictive variables in the final model are given in the Data Supplement. Five of the nine cancer-mortality-specific variables were also present in the noncancer model: age, Gleason score, PSA level, retention of urine, and use of ambulance services. The cumulative incidence of cancer-related death was lower than that of noncancer death (Fig 1). The 5-year cumulative incidence of cancer-related death was 0.5% for patients in the lowest risk group, compared with 8.2% for patients in the highest risk group. With respect to model accuracy for predicting cancer-specific survival, the p^2 value for our claims-based

prediction model was 0.71, which was virtually identical to the MSKCC-derived prediction, with a p^2 of 0.68.

Sensitivity Analysis

Our survival model construction approach did not restrict variables entered into the LASSO variable selection technique. Our final models ultimately included a limited number of claims-based variables related to prostate cancer work-up and diagnosis. For sensitivity analysis, we removed these prostate cancer-related claims variables, recalculated each risk score, and remeasured model prediction accuracy. Overall, when removing these prostate cancer work-up and diagnostic variables from the analysis, our model accuracy remained similar, with a p^2 of 0.59 for noncancer survival and 0.68 for cancer-specific survival,

FIG 2. Plots demonstrating the cross-distributions of the claims-based noncancer survival prediction algorithm and (A) the Charlson comorbidity index and (B) the Elixhauser comorbidity index.

FIG 3. Cumulative incidence of noncancer mortality across noncancer least absolute shrinkage and selection operator (LASSO) model tiers in patients with a (A) Charlson comorbidity index score of 0 ($n = 24,960$) and (B) Elixhauser comorbidity index < 0 ($n = 8,793$). The LASSO model characterizes a broad range of mortality in ostensibly low-risk patients from both models.

compared with 0.60 and 0.71 achieved by the respective full models.

DISCUSSION

Many treatment decisions in prostate cancer depend on the estimated life expectancy of the patient. Questions surrounding who should be screened¹⁵ as well as deciding who should receive active treatment depend on understanding which patients will live long enough to benefit from treatment. Prostate cancer represents a unique malignancy, given the older age of onset coupled with the often slow-growing cancer. This combination of factors creates a paradigm in which more men will die of competing causes of death unrelated to their prostate cancer. As such, the risk-benefit ratio of different interventions in prostate cancer depends substantially on a patient's risk of noncancer mortality.

This study presents an approach to creating survival prediction models. With noncancer survival prediction, our claims-based model outperformed conventional prediction algorithms including the CCI and ECI scores. Conventional survival-prediction algorithms typically rely on a discrete number of comorbid diagnoses, and our claims-based model demonstrates that the approach of using a more exhaustive data set that includes procedures, surgery, and diagnostic tests can produce more precise noncancer survival predictions. In addition, our model found that a small fraction of claims were protective. For example, a lipid panel test measures a patient's cholesterol levels, and the presence of this potentially preventive test was associated with improved survival. The use of LASSO for predictor selection offers the possibility of an agnostic approach to identifying a less-biased set of predictor variables, though, importantly, this approach does not imply

causality between variables and survival. Although the claims-based approach appears promising for predicting noncancer mortality, in the cancer-mortality model, the variable selection overlapped largely with the classic MSKCC nomogram. Given the added complexity of using claims in the prediction model, there would be little impetus to use this new approach in a clinical or research setting when evaluating cancer-specific mortality risks.

This claims-based predictive model has potential practical advantages over other prediction models. First, using a claims-based approach offers the possibility of integration into existing EHRs. Although different EHR systems record health care data in different formats, all EHRs share a common language with respect to billing claims. A practical application of this survival prediction approach could draw from claims within a patient's EHR and produce automated patient-specific survival estimations in a clinical setting. Physicians could use such risk-prediction tools to inform individualized conversations with patients about their relative risks of death from comorbidities versus cancer. Considering the risks of cancer-specific mortality and noncancer mortality can help better inform a provider's recommendation or a patient's decision with respect to treatment of prostate cancer. Beyond the clinical setting, our noncancer survival algorithms could have applications in cancer research. Our noncancer prediction algorithm represents a more sensitive indicator of patient comorbidity, and comorbidity in general represents an important confounder to consider in observational research. More-sensitive indicators of comorbidity stand to reduce bias in observational research. Finally, novel clinical trial designs are beginning to incorporate competing morbidity into patient selection ([ClinicalTrials.gov](https://clinicaltrials.gov) identifier:

NCT03258554). Incorporating improved indicators of noncancer mortality could improve efficacy and efficiency of clinical research.¹⁶

This analytic approach to survival prediction has limitations worth considering. First, one must consider that death represents a somewhat random stochastic event, and that precise patient-level prediction will always include some degree of error. Second, this analysis focused on a population of elderly patients with prostate cancer, and therefore this algorithm may not generalize to other populations including younger patients or those with other malignancies. Additional research on different cohorts of patients will be needed to determine how these prediction approaches perform outside this context. Third, cause-specific survival represents the primary end point of this

study, though accurate attribution of cause of death among patients with cancer can pose a challenge.¹⁷ Misclassification in cause of death could potentially bias our results. Finally, insurance claims themselves have the capacity for inaccuracy. While Although, in general, providers and insurers have incentives for accurate billing, misclassification of claims data could potentially bias our analysis, though this would likely work to decrease our model's predictive capacity toward the null.

Despite these limitations, the claims-based prediction model presented here substantially outperformed existing measures of noncancer survival. Predictive algorithms such as the models presented here have the potential to influence clinical decision-making and optimize care among patients with cancer.

AFFILIATIONS

¹University of California, San Diego, La Jolla, CA

²University of Hawaii, Honolulu, HI

³Institute for Behavioral and Community Health, San Diego State University, San Diego, CA

CORRESPONDING AUTHOR

James D. Murphy, MD, MS, 9452 Medical Center Dr, La Jolla, CA 92037; e-mail: j2murphy@ucsd.edu.

SUPPORT

Partially supported by the National Institutes of Health (Grant TL1TR001443 [P.R., R.S.]) and an ASCO Young Investigator Award (A.J.P.).

AUTHOR CONTRIBUTIONS

Conception and design: Paul Riviere, Anthony J. Paravati, Brent Rose, Ronghui Xu, James D. Murphy

Collection and assembly of data: Vinit Nalawade

Data analysis and interpretation: All authors

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated.

Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Paul Riviere

Employment: Peptide Logic

Anthony J. Paravati

Stock and Other Ownership Interests: XBI Biotech ETF

No other potential conflicts of interest were reported.

REFERENCES

1. National Comprehensive Cancer Network: NCCN Clinical Practice Guidelines in Oncology Prostate Cancer, Version 4.2018, 2018. https://www.nccn.org/professionals/physician_gls/pdf/prostate.pdf
2. Thompson I, Thrasher JB, Aus G, et al: Guideline for the management of clinically localized prostate cancer: 2007 Update. *J Urol* 177:2106-2131, 2007
3. Hamdy FC, Donovan JL, Lane JA, et al: 10-Year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *N Engl J Med* 375:1415-1424, 2016
4. Carmona R, Zakeri K, Green G, et al: Improved method to stratify elderly patients with cancer at risk for competing events. *J Clin Oncol* 34:1270-1277, 2016
5. Warren JL, Klabunde CN, Schrag D, et al: Overview of the SEER-Medicare data: Content, research applications, and generalizability to the United States elderly population. *Med Care* 40:IV-3-IV-18, 2002 (8, Suppl)
6. Daskivich TJ, Wood LN, Skarecky D, et al: Limitations of the National Comprehensive Cancer Network® (NCCN®) guidelines for prediction of limited life expectancy in men with prostate cancer. *J Urol* 197:356-362, 2017
7. Hou J, Paravati A, Hou J, et al: High-dimensional variable selection and prediction under competing risks with application to SEER-Medicare linked data. *Stat Med* 37:3486-3502, 2018
8. O'Quigley J, Xu R, Stare J: Explained randomness in proportional hazards models. *Stat Med* 24:479-489, 2005
9. Charlson ME, Pompei P, Ales KL, et al: A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J Chronic Dis* 40:373-383, 1987
10. Klabunde CN, Potosky AL, Legler JM, et al: Development of a comorbidity index using physician claims data. *J Clin Epidemiol* 53:1258-1267, 2000
11. Deyo RA, Cherkin DC, Ciol MA: Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 45:613-619, 1992
12. Elixhauser A, Steiner C, Harris DR, et al: Comorbidity measures for use with administrative data. *Med Care* 36:8-27, 1998
13. Lieffers JR, Baracos VE, Winget M, et al: A comparison of Charlson and Elixhauser comorbidity measures to predict colorectal cancer survival using administrative health data. *Cancer* 117:1957-1965, 2011

14. Kent M, Penson DF, Albertsen PC, et al: Successful external validation of a model to predict other cause mortality in localized prostate cancer. *BMC Med* 14:25, 2016
15. Grossman DC, Curry SJ, Owens DK, et al: Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *JAMA* 319:1901-1913, 2018
16. Rose BS, Jeong JH, Nath SK, et al: Population-based study of competing mortality in head and neck cancer. *J Clin Oncol* 29:3503-3509, 2011
17. Hinchliffe SR, Abrams KR, Lambert PC: The impact of under and over-recording of cancer on death certificates in a competing risks analysis: A simulation study. *Cancer Epidemiol* 37:11-19, 2013