original reports abstract

Validation of a Mortality Composite Score in the Real-World Setting: Overcoming Source-Specific Disparities and Biases

Michelle H. Lerman, BS[1]; Benjamin Holmes, MS[1]; Daniel St Hilaire, BA[1]; Mary Tran, MS[1]; Matthew Rioth, MD[1]; Vinod Subramanian, BE, MBA[1]; Alissa M. Winzeler, PhD[1]; and Thomas Brown, MD, MBA[1]

PURPOSE This study tested whether a composite mortality score could overcome gaps and potential biases in individual real-world mortality data sources. Complete and accurate mortality data are necessary to calculate important outcomes in oncology, including overall survival. However, in the United States, there is not a single complete and broadly applicable mortality data source. It is further likely that available data sources are biased in their coverage of sex, race, age, and socioeconomic status (SES).

METHODS Six individual real-world data sources were combined to develop a high-quality composite mortality score. The composite score was benchmarked against the gold standard for mortality data, the National Death Index. Subgroup analyses were then conducted to evaluate the completeness and accuracy by sex, race, age, and SES.

RESULTS The composite mortality score achieved a sensitivity of 94.9% and specificity of 92.8% compared with the National Death Index, with concordance within 1 day of 98.6%. Although some individual data sources show significant coverage gaps related to sex, race, age, and SES, the composite score maintains high sensitivity (84.6%-96.1%) and specificity (77.9%-99.2%) across subgroups.

CONCLUSION A composite score leveraging multiple scalable sources for mortality in the real-world setting maintained strong sensitivity, specificity, and concordance, including across sex, race, age, and SES subgroups.

JCO Clin Cancer Inform 5:401-413. © 2021 by American Society of Clinical Oncology

INTRODUCTION

Real-world data (RWD) collected from routine patient care are valuable for expediting and enhancing outcomes research in oncology. It provides an opportunity to characterize cancer care and outcomes among a broader set of patients, including groups often underrepresented in traditional prospective clinical trials and population studies.[1-3] RWD applications are expanding, particularly in oncology, with regard to research, clinical care, regulatory, and commercial applications. However, capturing accurate and complete RWD necessary to power meaningful research is challenging given the fragmented nature of healthcare delivery and associated data collection in the United States (US).

Complete and accurate mortality data are necessary to calculate important outcomes in oncology, including overall survival. Incomplete mortality data can cause inaccurate estimation of survival and result in erroneous conclusions in comparative studies.[4] In the United States, there is no single mortality data source that is both complete and broadly applicable.

Available data sources for mortality each have limitations. The Centers for Disease Control and Prevention's National Death Index (NDI) captures all US death certificates and is considered the gold standard, but access to and use of NDI data is limited. The National Cancer Institute's SEER database also captures death certificates but has limited geographic coverage.[5] Online obituary aggregators and the Social Security Death Index and/or Death Master File (SSDI) capture broad but incomplete data. Documentation in electronic health records (EHRs) is also often incomplete, particularly for deaths occurring outside of healthcare facilities. Finally, many cancer registries often rely on SSDI, EHRs, or manual curation (eg, by searching online obituaries) and do not include all patients.[6]

Mortality data sources in the real world likely have biases that influence which patients are represented. It is well-documented that obituaries underrepresent women.[7,8] Additional studies have shown differences in obituary descriptions by race for those individuals who receive obituaries.[9] Beyond these known biases, disparities in coverage of mortality data across race

ASSOCIATED CONTENT

Appendix

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

Accepted on February 8, 2021 and published at ascopubs.org/journal/cci on April 8, 2021: DOI https://doi.org/10.1200/CCI.20.00143

CONTEXT

Key Objective

What are the gaps in completeness and accuracy among real-world sources for mortality data, and can a composite mortality score overcome these gaps?

Knowledge Generated

Compared with the gold-standard source for vital status and dates of death, real-world data sources were generally accurate but had significant gaps in data coverage. Development of a mortality composite score that combines data across sources addressed these gaps and achieved a sensitivity of 94.9% compared with the gold standard.

Relevance

Mortality data from electronic health records and registries at community health systems tend to be complete but have notable gaps and biases. Other real-world data sources, including digitized obituaries, do not capture all deaths and may underrepresent patients on the basis of age, race, sex, and socioeconomic factors. However, combining across sources to develop a composite mortality score addresses biases associated with individual sources and provides more complete capture of patient vital status and dates of death.

and/or ethnicity, age, or socioeconomic status (SES) have not been well-documented.

The hypothesis for this work was that a composite mortality score could overcome gaps and biases in real-world mortality data sources. Potential areas of bias by age, sex, race, and SES were evaluated. Previous research has examined approaches to improving the accuracy of real-world mortality data by combining data from multiple sources;[10,11] however, to our knowledge, this is the first study to evaluate whether a composite score overcomes biases pertaining to age, sex, race, and SES associated with individual sources.

METHODS

Data Sources and Study Population

A sample of patients diagnosed with cancer between 2011 and 2017 was selected from the Syapse Learning Health Network (LHN), a proprietary database of patients from US health systems data sources including EHRs, enterprise data warehouses, laboratories, tumor registries, digitized obituaries, and other clinical sources. International Classification of Disease diagnosis codes and tumor-specific data tables were used to identify patients with cancer.

Vital status was based on the presence of death data obtained from six sources: (1) hospital EHR data feeds, (2) hospital tumor registries, (3) digitized obituaries, (4) SSDI, (5) SEER, and (6) manual curation. EHR data feeds and hospital tumor registry sources were obtained via direct integrations with health systems. Digitized obituaries and SSDI were sourced through publicly available US mortality data sources and linked using probabilistic patient matching.[12] Patient-level identifiable data from SEER was obtained via direct bidirectional data sharing with health systems. Manual curation was conducted by Certified Tumor Registrars with access to enterprise-wide EHRs in partner health systems.

If the patient was listed as deceased from any source, that patient was marked as deceased in the composite score. Date of death was determined as the first death date found when searching these sources in a waterfall method. The waterfall method examined data sources in ascending order to resolve discrepancies: hospital tumor registries, SEER, EHRs, SSDI, digitized obituaries, and manual curation. Data across sources were linked at the patient level (Fig 1).

At the time of this study, NDI reported dates of death occurring until the end of 2017. To ensure an appropriate comparison, other mortality data sources were limited to the same time window (eg, 2011-2017).

NDI was selected as the comparator given its status as the gold standard for mortality in the United States.[13-16] Patient identifiers (first and last name, sex, race, dates of birth and death, age at death, and location of death by state) for these patients were submitted for matching with NDI. On the basis of NDI User's Guide,[6] matches with a score above 27 were considered positive matches, with an estimated accuracy of $>99\%$. Patients with NDI matching scores above this threshold were therefore marked as deceased according to the NDI.

Statistical Analysis

Extraction of patients and division of patients into subsets for analysis was performed using RStudio v. 3.6.1, along with the Open Database Connection (obdc) v. 1.2.2, and dplyr v. 1.0.2 libraries for query creation, eeptools v. 1.2.0 for age calculation, and data.table v. 1.13.2 for representing data structures.

Calculation of sensitivity and specificity was performed with Python 3.8, using pandas v 1.1.4 and numpy v 1.17.4 for dataframe representation and transformation and datetime.datetime for representing dates. CIs were calculated using Python 3.8 using the math and scipy v. 1.4.1

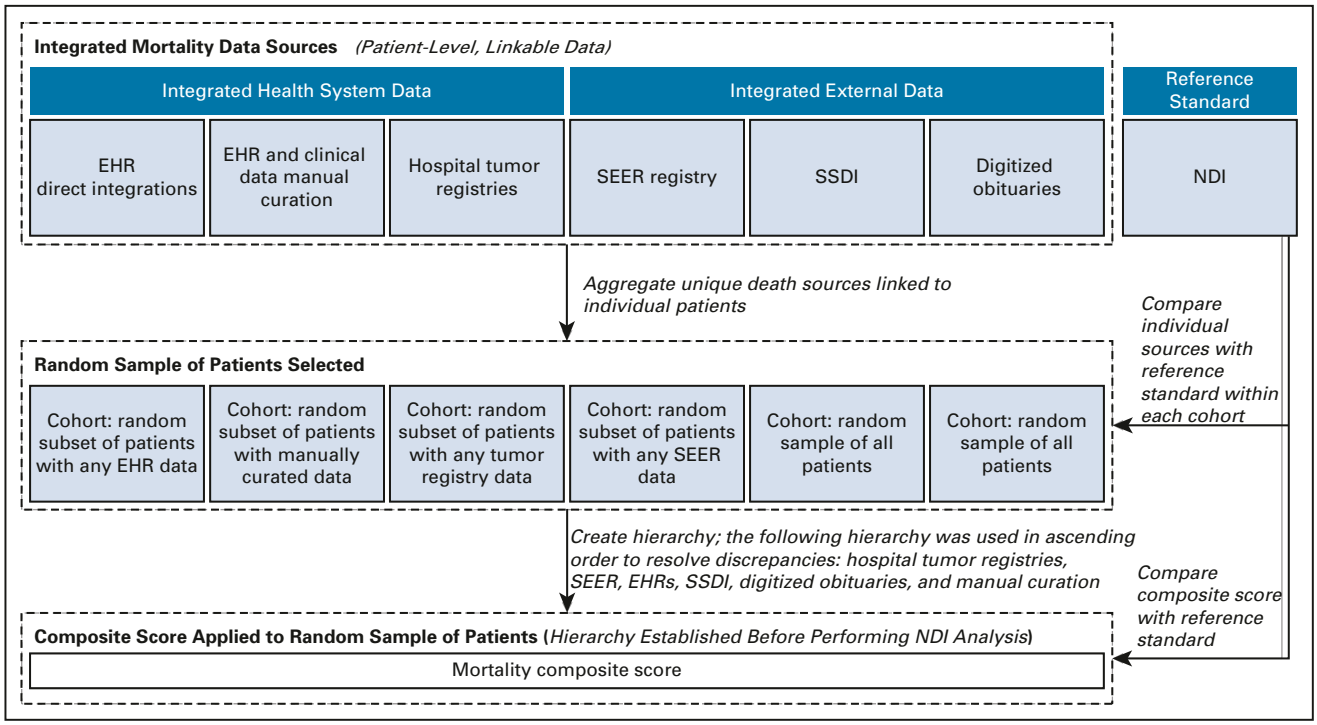

FIG 1. Composite score development flowchart. EHR, electronic health record; NDI, National Death Index; SSDI, Social Security Death Index.

libraries. Chi-square tests were conducted with the scipy.stats chi2_contingency library.

Sensitivity, specificity, positive predictive value, negative predictive value, and date concordance were evaluated against the NDI (Table 1). Assessments were conducted for each data source individually versus NDI and for the composite mortality score versus NDI. This composite mortality score was also assessed with and without data from manual curation.

Patients were considered deceased according to NDI if they matched an NDI record with a matching score above 27; otherwise, the patients were considered alive according to NDI. For individual data sources, analyses were performed using the subset of patients for whom each individual source was relevant (eg, only patients with tumor registry data available) or all patients (eg, SSDI and digitized obituaries), as appropriate.

Sensitivity and specificity were calculated for subgroups defined by extent of cancer spread (metastatic and nonmetastatic), race (White, Black, Asian or Pacific islander, American Indian or Alaskan Native, and Unknown), age (younger than 30, 30-59 years, 60 and older), sex, and SES on the basis of residential zip code (household income $<$ \$30,000 in US dollars [USD], \$30,000-\$59,999 USD, \$60,000-\$99,999 USD, $\geq$ \$100,000 USD). If multiple zip codes were recorded for a patient, then the last location on file was used to establish the median income category.

CIs for sensitivity, specificity, and concordance were computed using Wilson's method, chosen to give accurate CIs for mismatched and small sample sizes. Chi-squared statistics with $P \leq .05$ indicate statistical significance.

Additional details on the code used to generate the analysis are available in the Appendix.

Ethical Considerations

The research conducted by Syapse is exempt from Institutional Review Board review because this is a retrospective research project using de-identified patient data. Syapse received this exemption from Advarra, an external and independent Institutional Review Board.

RESULTS

The final study population included 90,993 unique patients after removing duplicates, of whom 17,614 patients had metastatic cancer (Table 2). Approximately 47.6% of the population was female (13.5% unknown sex), 80.8% were 60 years and older, and 61.1% were White. Patients were concentrated in the low to middle household income range (\$30,000-\$59,999 USD: 54.7%; \$60,000-\$99,999 USD: 40.5%).

Overall, the sensitivity for vital status across individual data sources ranged from 14.9% (SSDI) to 98.6% (SEER) (Table 3). Specificity ranged from 79.0% (manual curation) to 99.4% (SSDI). For true positive dates of death, concordance (date agreement) within 1 day of NDI date of

TABLE 1. Confusion Matrix Comparing RWD Sources with the NDI

	NDI Data		
	Deceased	**Alive**	
Real-world mortality data sources			
Deceased	True positives (A)	False positives (B)	PPV = A/(A + B)
Alive	False negatives (C)	True negatives (D)	NPV = D/(C + D)
	Sensitivity = A/(A + C)	Specificity = D/(B + D)	

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

death ranged from 94.7% (hospital tumor registries) to 99.9% (SEER) across sources.

The composite score achieved a sensitivity of 94.9% with a specificity of 92.8% and concordance within 1 day of 98.6% (Table 3). Excluding dates of death obtained via manual curation, the composite score achieved sensitivity of 93.4%, specificity of 93.4%, and concordance within 1 day of 97.3%. The composite score achieved similarly high sensitivity (95.9%) for patients with metastatic cancer (Table 4), with 86.4% specificity and 98.6% concordance within 1 day.

There were a number of statistically significant differences in sensitivity for subgroup analyses from individual data sources (Table 5 and Appendix Table A1). Mortality data in obituaries, SSDI, and EHRs were more complete for patients age 60 years and older compared with their younger counterparts (all $P < .05$). Black patients and Asian patients were also less likely to have data captured in obituaries and SEER compared with White patients (all $P < .05$). Manual curation and registries were less likely to capture Black patients compared with White patients, whereas SSDI and EHRs were less likely to capture Asian patients compared with White patients (all $P < .05$). SSDI was more likely to capture patients residing in areas with median household income of $30,000-$59,000 USD compared with patients in the $60,000-$99,999 USD category ($P < .05$). By contrast, EHRs and tumor registries were less likely to capture patients in the $30,000-$59,000 USD category compared with patients in the $60,000-$99,999 USD category ($P < .05$). Differences in sensitivity were not observed in the composite mortality score for any subgroup, either including or not including manual curation (all $P > .05$). By combining data across sources, the composite score was able to overcome these biases and maintain high sensitivity (84.6%-96.1%), specificity (77.9%-99.2%), and date concordance within one day (95.7%-100.0%) across subgroups.

TABLE 2. Patient Demographics

Characteristics	No. Patients	% Total
All patients	90,993	
Sex		
Male	35,392	38.9
Female	43,295	47.6
Unknown	12,306	13.5
Race		
White	55,624	61.1
Black or African American	3,437	3.8
Asian or Pacific Islander	1,726	1.9
American Indian or Alaskan Native	296	0.3
Unknown	29,910	32.9
Age		
Younger than 30	508	0.6
30-59	16,952	18.6
60 and older	73,533	80.8
Metastatic status		
Metastatic	17,614	19.4
Nonmetastatic	73,379	80.6
Socioeconomic status		
Less than $30,000 USD	2,374	2.6
$30,000-$59,999 USD	49,766	54.7
$60,000-$99,999 USD	36,813	40.5
$100,000 and higher USD	2,040	2.2

Abbreviation: USD, US dollars.

DISCUSSION

Individual data sources for death have intrinsic biases in specific populations that limit their utility in determining mortality for a group of unselected patients. A composite mortality score developed from these sources overcomes these biases, in particular regarding race and sex.

This study demonstrated that a composite mortality score was capable of achieving sensitivity of 94.9% and specificity of 92.8% compared with the NDI, with concordance within 1 day of 98.6%.

Excluding manual curation from the mortality composite score, sensitivity and concordance decreased by 1.6 percentage points or less across these measures (sensitivity 93.4%, specificity 93.4%, and concordance within 1 day 97.3%), indicating that a high-quality mortality composite score can be built from scalable sources alone.

The concordance for all individual data sources versus NDI showed high degrees of accuracy (94.7%-99.9% concordance within 1 day). Thus, if a date of death was present in any source, it could be reasonably trusted to be accurate.

TABLE 3. Mortality Completeness and Accuracy Versus NDI by Source for All Patients

	Data Source *v* NDI							
All Patients	**SEER**	**Digitized Obituaries**	**SSDI**	**Manual Curation**	**EHRs**	**Hospital Tumor Registries**	**Composite Score With All Sources**	**Composite Score Without Manual Curation**
Overall number	14,812	90,993	90,993	11,011	86,180	24,369	90,993	90,993
Sensitivity, %	98.6	47.8	14.9	93.5	76.9	87.2	94.9	93.4
95% CI	98.2 to 99.0	47.1 to 48.6	14.4 to 15.4	92.8 to 94.2	76.3 to 77.5	86.4 to 87.9	94.6 to 95.2	93.0 to 93.7
Specificity, %	97.7	99.3	99.4	79.0	94.6	92.9	92.8	93.4
95% CI	97.4 to 98.0	99.2 to 99.3	99.3 to 99.5	78.2 to 79.8	94.4 to 94.8	92.5 to 93.2	92.6 to 93.0	93.3 to 93.6
PPV, %	93.3	94.2	85.8	65.7	77.3	84.4	76.3	77.6
95% CI	92.4 to 94.0	93.7 to 94.7	84.6 to 87.0	64.6 to 66.9	76.7 to 78.0	83.6 to 85.2	75.7 to 76.8	77.1 to 78.2
NPV, %	99.6	88.6	82.7	96.6	94.5	94.2	98.7	98.3
95% CI	99.4 to 99.7	88.4 to 88.9	82.5 to 83.0	96.2 to 97.0	94.3 to 94.6	93.9 to 94.6	98.6 to 98.8	98.2 to 98.4
True positives (included in concordance)	3,526	8,533	2,654	3,100	12,821	6,515	16,930	16,655
Date concordance (% within 0 day)	99.8	98.7	97.1	97.9	90.0	92.8	96.7	92.5
95% CI	99.7 to 100.0	98.5 to 99.0	96.4 to 97.7	97.4 to 98.4	89.5 to 90.5	92.2 to 93.4	96.4 to 96.9	92.1 to 92.9
Date concordance (% within 1 day)	99.9	99.4	97.7	99.0	97.9	94.7	98.6	97.3
95% CI	99.8 to 100.0	99.2 to 99.5	97.1 to 98.2	98.6 to 99.3	97.7 to 98.2	94.2 to 95.3	98.4 to 98.7	97.1 to 97.6
Date concordance (% within 15 days)	100.0	99.7	98.3	99.7	99.0	97.4	99.2	98.7
95% CI	99.9 to 100.0	99.6 to 99.8	97.8 to 98.8	99.6 to 99.9	98.8 to 99.2	97.0 to 97.8	99.1 to 99.4	98.6 to 98.9

Abbreviations: EHR, electronic health record; NDI, National Death Index; NPV, negative predictive value; PPV, positive predictive value; SSDI, Social Security Death Index.

TABLE 4. Mortality Completeness and Accuracy Versus NDI by Source for Metastatic Patients

	Data Source *v* NDI							
Metastatic Patients	**SEER**	**Digitized Obituaries**	**SSDI**	**Manual Curation**	**EHRs**	**Hospital Tumor Registries**	**Composite Score With All Sources**	**Composite Score Without Manual Curation**
Overall number	2,545	17,614	17,614	2,715	16,527	3,976	17,614	17,614
Sensitivity, %	99.4	50.0	13.4	97.5	81.9	88.8	95.9	95.0
95% CI	98.8 to 99.7	48.8 to 51.2	12.6 to 14.2	96.5 to 98.3	81.0 to 82.8	87.4 to 90.1	95.4 to 96.3	94.5 to 95.5
Specificity, %	97.4	99.3	99.5	77.1	88.4	86.4	86.4	87.1
95% CI	96.3 to 98.2	99.1 to 99.4	99.3 to 99.6	75.0 to 79.1	87.8 to 89.0	84.8 to 87.8	85.7 to 87.0	86.5 to 87.8
PPV, %	97.6	98.0	94.6	76.5	83.7	87.8	82.9	83.6
95% CI	96.7 to 98.3	97.5 to 98.4	93.0 to 95.8	74.3 to 78.6	82.8 to 84.5	86.4 to 89.2	82.1 to 83.7	82.7 to 84.3
NPV, %	99.3	74.3	62.6	97.6	87.1	87.4	96.9	96.2
95% CI	98.7 to 99.7	73.6 to 75.0	61.8 to 63.3	96.6 to 98.3	86.4 to 87.8	85.9 to 88.9	96.5 to 97.2	95.8 to 96.6
True positives (included in concordance)	1,307	3,587	960	1,148	5,688	1,855	6,881	6,818
Date concordance (% within 0 day)	99.8	98.9	97.8	95.6	87.6	92.6	95.9	89.9
95% CI	99.5 to 100.0	98.5 to 99.2	96.9 to 98.7	94.5 to 96.8	86.8 to 88.5	91.4 to 93.8	95.4 to 96.4	89.2 to 90.6
Date concordance (% within 1 day)	99.8	99.5	98.2	97.8	98.4	94.3	98.6	97.5
95% CI	99.5 to 100.0	99.2 to 99.7	97.4 to 99.1	97.0 to 98.7	98.1 to 98.7	93.3 to 95.4	98.3 to 98.9	97.1 to 97.9
Date concordance (% within 15 days)	99.9	99.8	98.4	99.5	99.2	97.6	99.3	98.9
95% CI	99.8 to 100.0	99.6 to 99.9	97.7 to 99.2	99.1 to 99.9	99.0 to 99.4	96.9 to 98.3	99.1 to 99.5	98.6 to 99.1

Abbreviations: EHR, electronic health record; NDI, National Death Index; NPV, negative predictive value; PPV, positive predictive value; SSDI, Social Security Death Index.

TABLE 5. Significance Matrix—*P* Values From Chi-Square Comparisons of Sensitivity Between Cohort Subsegments

	Data Source *v* NDI							
Patient Subsegments	**SEER**	**Digitized Obituaries**	**SSDI**	**Manual Curation**	**EHRs**	**Hospital Tumor Registries**	**Composite Score With All Sources**	**Composite Score Without Manual Curation**
Race								
White *v* Black (*P* values)	< .001	< .00001	.4	< .05	.2	< .05	.2	.7
White *v* Asian (*P* values)	< .00001	< .00001	< .05	.2	< .00001	.5	.8	.1
Sex								
Male *v* female (*P* values)	.3	< .00001	.9	.9	.9	.3	.9	.2
Age								
30-59 *v* 60 and older (*P* values)	.7	< .05	< .00001	.2	< .00001	.1	.1	.1
SES								
$30,999-$59,999 *v* $60,000-$99,999 USD (*P* values)	.6	.6	< .00001	.7	< .00001	< .001	.1	.2

Abbreviations: EHR, electronic health record; NDI, National Death Index; SES, socioeconomic status; SSDI, Social Security Death Index; USD, US dollars.

Sensitivity varied by data source, ranging from 14.9% to 98.6%. Additionally, not all sources are relevant to all patient populations, further influencing data completeness by source. Hospital tumor registries are specific to each health system, and include only patients newly diagnosed within that health system, leaving gaps in coverage. SEER mortality data had the highest sensitivity (98.6%), specificity (97.7%), and concordance (99.9% within 1 day) of any source; however, SEER is only relevant for patients treated within its coverage areas (currently ~35% of the US population),[5] and only 16.3% of the randomly selected patient population matched SEER data available at the time of the study. No single source achieved high levels of sensitivity with relevance across the entire real-world patient population. When combined, however, the composite score achieves trustworthy levels of completeness (sensitivity of 94.9%) and accuracy (concordance within 1 day of 98.6%).

When evaluating individual data sources, there were a number of significant differences in sensitivity between different patient subgroups. These potential biases are important because disparities in data completeness and accuracy can cause researchers to draw incorrect conclusions in real-world studies.[4] Differential accuracy of mortality data by demographic or other patient characteristics, as observed for individual data sources in the current study, can invalidate comparative analyses.

First, men were more likely than women to have their deaths captured in digitized obituaries. This is consistent with prior analyses of digitized obituaries that showed that women were awarded significantly fewer obituaries compared with men, potentially because of survivor bias and/or the value placed on women's life achievements. [8] Obituaries were the only mortality source for which a sex bias was statistically significant, and biases were not observed for the composite score.

Second, every individual source exhibited statistically significant differences in sensitivity of vital status by race. Black patients were less likely to have dates of death captured than White patients by SEER, obituaries, manual curation, and tumor registries, whereas Asian patients were less likely than White patients to have dates of death captured by SEER, obituaries, SSDI, and EHRs. The greatest racial bias within these sources was observed in obituary data, which achieved 50.90% sensitivity for White patients but only 31.66% for Black patients and 26.87% for Asian patients. This suggests that Black and Asian patients are either less likely than White patients to have their deaths memorialized in an obituary or that their obituaries are less likely to be captured by digital obituary scrapers. To our knowledge, this appears to be the first documentation of racial bias in digitized obituary data within the United States. These racial disparities in mortality documentation can potentially lead to healthcare disparities within RWD and real-world evidence. First, observational research studies relying on the completeness of these data may misrepresent outcomes in these populations. Second, searchable online obituaries are sometimes used as a resource when manually abstracting registries; thus, biases within digitized obituaries may be propagated through other, seemingly unrelated sources. The composite mortality score was able to overcome these data source biases.

Comparing across age, patients 60 years and older were significantly more likely to have death data captured in digitized obituaries, SSDI, and EHRs than patients between 30 and 59 years. The population age 30-59 years also had more missing identifiers than the population age 60 years and older (data not shown), which may have affected

matching rates with NDI in some cases. It is possible that the availability of Medicare among the older population improves the availability of patient identifiers. As with sex and race, significant biases by age were not observed in the mortality composite score.

Finally, some intriguing trends were observed in segmenting patients by the median income level within their zip code that warrant further investigation. Given the small size of the lowest and highest income populations, trend variability across sources, and the use of an income proxy rather than income, it is difficult to draw clear conclusions.

This study has a number of limitations. Most importantly, NDI has limitations in both patient linkage and recency. Patients are matched to NDI on the basis of available patient identifiers including name, date of birth, sex, race, state of birth, and state of residence. If a match to NDI is found, the corresponding date of death and the confidence level in the match are returned. If a match is not found, the patient is assumed to be alive. If there are errors matching patients, these patients would be erroneously labeled as alive according to NDI. This could result in errant false positives (deceased in source but alive in NDI) and reduced specificity. When limiting analysis to only patients with complete patient identifiers, however, the resulting composite score showed only slightly improved specificity (from 92.8% to 94.9%), sensitivity (from 94.9% to 95.8%), and concordance within 1 day (from 98.5% to 98.9%), indicating that matching errors, while a potential factor, likely did not have a significant impact on the results.

Additionally, NDI mortality data are delayed relative to typical real-world data sets. Data in the LHN real-world database are updated regularly, with many automated real-time or daily feeds. At the time of the analysis, data available from NDI covered deaths occurring only until the end of 2017. Thus, this study was unable to determine the sensitivity, specificity, and concordance of patient deaths occurring between January 2018 and November 2019, when the study was initiated.

The mortality composite score tested in this study was developed a priori, prior to measuring each source in comparison with NDI. If a patient had dates of death from multiple sources, the mortality composite score used the following hierarchy, in ascending order, to resolve discrepancies: hospital tumor registries, SEER, EHRs, SSDI, digitized obituaries, and manual curation. It is likely that a reordering of this composite score ranking methodology could further improve the concordance, given that hospital tumor registries and EHRs were ranked first and third but had among the lowest NDI concordance.

Results are also dependent on availability of multiple high-quality data sources. Patients within the database are treated within large, integrated community health systems with enterprise-wide EHR capture across the system (eg, outpatient oncology clinic, inpatient hospital, and hospice), allowing for greater capture of mortality data. Indeed, the level of sensitivity for EHR data alone within this study (76.9%) exceeds a comparable published report of outpatient oncology EHR data only (65.97%).[10] It further exceeds coverage by the Department of Veterans Affairs Medical SAS Inpatient Datasets (sensitivity of 12.0% compared with NDI).[17] In addition, this study relies on data from both hospital tumor registries and SEER. Because of reporting requirements, these sources are typically maintained by and available to hospitals or health systems, but less so to independent community practices, giving health systems an advantage in creating a strong composite mortality score.

In summary, this study shows that although there were several significant sensitivity biases within individual mortality sources, a composite score was able to successfully overcome these biases. This indicates that merging multiple high-quality but incomplete sources together is able to overcome biases in individual sources and can generate a trustworthy mortality score for an entire real-world patient population.

AFFILIATION

[1]Syapse, San Francisco, CA

CORRESPONDING AUTHOR

Thomas Brown, MD, MBA, Syapse, 303 2nd St, North Tower, Suite 550, San Francisco, CA 94107; Twitter: @Syapse; e-mail: tom.brown@syapse.com.

DISCLAIMER

Syapse is a privately held company. All research was performed in-house by Syapse employees using company resources. There were no sources of external funding used to support this research, including grants, contracts, or philanthropy.

EQUAL CONTRIBUTION

M.H.L., B.H., and D.S.H. contributed equally to this work.

DATA SHARING STATEMENT

Syapse obtains the legal rights to use the de-identified patient data in its possession to conduct research from its business partners. These rights restrict Syapse's ability to publicly share or provide access to the de-identified data.

AUTHOR CONTRIBUTIONS

Conception and design: Michelle H. Lerman, Daniel St Hilaire, Mary Tran, Vinod Subramanian, Alissa M. Winzeler, Thomas Brown
Administrative support: Thomas Brown
Collection and assembly of data: Benjamin Holmes, Daniel St Hilaire, Thomas Brown
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 unless otherwise noted. 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/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Michelle H. Lerman
Employment: Syapse
Stock and Other Ownership Interests: Syapse

Benjamin Holmes
Employment: Syapse
Stock and Other Ownership Interests: Syapse
Travel, Accommodations, Expenses: Syapse

Daniel St Hilaire
Employment: Syapse
Stock and Other Ownership Interests: Syapse

Mary Tran
Employment: Syapse
Stock and Other Ownership Interests: Syapse

Matthew Rioth
Employment: Syapse
Stock and Other Ownership Interests: Syapse

Vinod Subramanian
Employment: Syapse
Leadership: Syapse
Stock and Other Ownership Interests: Syapse
Travel, Accommodations, Expenses: Syapse

Alissa M. Winzeler
Employment: Syapse
Leadership: Syapse
Stock and Other Ownership Interests: Syapse
Travel, Accommodations, Expenses: Syapse

Thomas Brown
Employment: GenomiCare Biotechnology, Syapse
Leadership: Syapse
Stock and Other Ownership Interests: GenomiCare Biotechnology, Syapse
Honoraria: Novartis
Consulting or Advisory Role: Jiahui Health, GenomiCare Biotechnology, Lug Healthcare Technology, Syapse
Speakers' Bureau: Syapse, Novartis
Travel, Accommodations, Expenses: Syapse, Jiahui Health, GenomiCare, Lug Healthcare Technology, Novartis

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors thank Sally Wadedaniel, Chenan Zhang, Yanina Natanzon, Joshua Loving, Andrew Schrag, Jay Ronquillo, Jonathan Hirsch, Alakananda Iyengar, Raghu Warrier, Sheetal Walters, Lydie Prevost, Arnon Moscona, and Antoinette Cummins.

REFERENCES

1. Loree JM, Anand S, Dasari A, et al: Disparity of race reporting and representation in clinical trials leading to cancer drug approvals from 2008 to 2018. JAMA Oncol 5:e191870, 2019
2. Nazha B, Mishra M, Pentz R, et al: Enrollment of racial minorities in clinical trials: Old problem assumes new urgency in the age of immunotherapy. Am Soc Clin Oncol Educ Book 39:3-10, 2019
3. Sharrocks K, Spicer J, Camidge DR, et al: The impact of socioeconomic status on access to cancer clinical trials. Br J Cancer 111:1684-1687, 2014
4. Bretscher MT, Sanglier T: Quantifying the Impact of Mortality Underreporting on Analyses of Overall Survival. Presented at the 34th International Conference for Pharcoepidemiology and Risk Management, Prague, Czech Republic, August 24, 2018
5. National Cancer Institute: About the SEER Registries. Bethesda, MD, NCI
6. National Center for Health Statistics: National Death Index User's Guide. Hyattsville, MD, National Center for Health Statistics, 2013
7. Maybury KK: Invisible lives: Women, men and obituaries. Omega 32:27-37, 1996
8. Ogletree SM, Pena D, Figueroa P: A double standard in death? Gender differences in obituaries. Omega 51:337-342, 2005
9. Marks A, Piggee T: Obituary analysis and describing a life lived: The impact of race, gender, age, and economic status. Omega 31:37-57, 1999
10. Curtis MD, Griffith SD, Tucker M, et al: Development and validation of a high-quality composite real-world mortality endpoint. Health Serv Res 53:4460-4476, 2018
11. Turchin A, Shubina M, Murphy SN: I am not dead yet: Identification of false-positive matches to death master file. AMIA Annu Symp Proc 2010:807-811, 2010
12. Datavant: Further Evidence That COVID-19 Disproportionately Impacts African-American, Hispanic, and Low-Income Populations. San Francisco, CA, Datavant
13. Cowper DC, Kubal JD, Maynard C, et al: A primer and comparative review of major US mortality databases. Ann Epidemiol 12:462-468, 2002
14. Fillenbaum GG, Burchett BM, Blazer DG: Identifying a national death index match. Am J Epidemiol 170:515-518, 2009
15. Rich-Edwards JW, Corsano KA, Stampfer MJ: Test of the National Death Index and Equifax Nationwide Death Search. Am J Epidemiol 140:1016-1019, 1994
16. Wentworth DN, Neaton JD, Rasmussen WL: An evaluation of the social security administration master beneficiary record file and the National Death Index in the ascertainment of vital status. Am J Public Health 73:1270-1274, 1983
17. Sohn MW, Arnold N, Maynard C, et al: Accuracy and completeness of mortality data in the department of veterans affairs. Popul Health Metr 4:2, 2006

APPENDIX
CODE USED TO GENERATE ANALYSES

Libraries

Libraries and versions used:

True positives, false positives, true negatives, and false negatives were derived by comparing death status from Syapse patients and the corresponding death status from NDI.

Python 3.8 was used for coding.

Libraries used:

#######

For general dataframe loading

#######

pandas - 1.1.4

numpy - 1.17.4

datetime.datetime - standard library

######

For confidence interval

######

math.sqrt - standard library (no version)

scipy.special.ndtri - 1.4.1

Confusion Matrix

```
import pandas as pd

import numpy as np

# For regex

# Input dataframe contains columns with comparison values of our various sources vs NDI.

# These can take on four values: TP, FP, TN, FN for True and False Positives and Negatives

# There are also columns with the difference in days of dates of death - these can be from 0 and up - columns are ints.

inputDataFrame = pd.read_csv("~/patientFileForDoDWhitePatients.csv", low_memory=False)

standardDataFrame = pd.read_csv("~/patientFileForDoDAllPatients.csv")

# There are also dataframes containing only a subset of patients

# These include dfs with patients split up by:

# - Tumor type (Any Lung Cancer, NSCLC, SCLC, AML, All CRC, Advanced CRC, All Breast cancer, Metastatic Breast Cancer, Unknown)

# - Race (White, Black or African American, Asian / Pacific Islander, American Indian / Alaskan Native, Unknown)

# - Age (0-30, 30-60, 60 and up, Unknown)

# - Sex (Male, Female, Unknown)

# - SES (as determined by patient zip code: 0-30K, 30-60K, 60-100K, 100-185K, >185K, Unknown)

# - Metastatic Status (Metastatic, Non-Metastatic)

# We want to go column by column comparing the positive/negative status and the date agreement for our various sources.

positivesValues = ['1. SEER vs. NDI', '2. Datavant vs. NDI', '2a. Obit vs. NDI', '2b. SSA vs. NDI', '4. Manual Abstraction vs. NDI', '5. EMR vs. NDI', '6. Hospital Registries vs. NDI', '7. Rolled up Syapse View vs. NDI', 'Rolled Up Syapse View Without MA Vs NDI']

datesValues = ['1. SEER DoD Agreement', '2. Datavant DoD Agreement', '2a. Obit DoD Agreement', '2b. SSA DoD Agreement', '4. Manual Abstraction DoD Agreement', '5. EMR DoD Agreement', '6. Hospital Registries DoD Agreement', '7. Rolled up Syapse View DoD Agreement', 'Rolled Up Syapse View Without MA Date Diff']

for x in range(0, len(positivesValues)):

# Total number is the overall # of patients

# Class number is the # of patients in this 'class', AKA the group under review

totalNumber = len(standardDataFrame.index)

classNumber = len(inputDataFrame.index)

# Our input data frame will be the one we get values from

columnPos = positivesValues[x]

columnDates = datesValues[x]

print(columnPos)

TP = len(inputDataFrame[inputDataFrame[columnPos] == 'TP'])

TN = len(inputDataFrame[inputDataFrame[columnPos] == 'TN'])

FP = len(inputDataFrame[inputDataFrame[columnPos] == 'FP'])

FN = len(inputDataFrame[inputDataFrame[columnPos] == 'FN'])

print('overall # -', (TP + TN + FP + FN))

if totalNumber > 0 and classNumber > 0:

print('percent coverage - all patients', ((TP + TN + FP + FN) / totalNumber))

print('percent coverage - this subgroup', ((TP + TN + FP + FN) / classNumber))

print('tp - ', str(TP))

print('fp - ', str(FP))

print('tn - ', str(TN))

print('fn - ', str(FN))

# This is to handle situations with very low patient numbers:

# We'll to avoid dividing by 0, we'll set 0 numbers to just be very small - this will still show up as 0s

if TP == 0 and FP == 0:

FP = 0.0000000000000000000000000001

TP = 0.0000000000000000000000000001

if TN == 0 and FN == 0:

FN = 0.0000000000000000000000000001

TN = 0.0000000000000000000000000001

if TP == 0 and FN == 0:

TP = 0.0000000000000000000000000001

FN = 0.0000000000000000000000000001

if TP == 0:

TP = 0.0000000000000000000000000001

print('Sensitivity - ' + str(TP / (TP + FN)))

print('Specificity - ' + str(TN / (FP + TN)))

print('PPV - ' + str(TP / (TP + FP)))

print('NPV - ' + str(TN / (TN + FN)))

# Find the number off by zero days - we're also only interested in ones where the patient is in fact deceased

# FP will mean NDI doesn't have a DoD

# FN will mean Syapse doesn't have a DoD

# TN will mean neither has a DoD

numWithExact = len(inputDataFrame[(inputDataFrame[columnDates] == 0) & (inputDataFrame[columnPos] == 'TP')]) + len(
```



```
inputDataFrame[(inputDataFrame[columnDates].isna()) & (inputDataFrame[columnPos] == 'TP')])

print('% within 1 day - ' + str((numWithExact + len(inputDataFrame[(inputDataFrame[columnDates] == 1) & (inputDataFrame[columnPos] == 'TP')])) / TP))

print('# within 0 day - ' + str(numWithExact))

print('# within 1 day - ' + str(len(inputDataFrame[(inputDataFrame[columnDates] == 1) & (inputDataFrame[columnPos] == 'TP')])))

print('% within 1 day - ' + str(len(inputDataFrame[inputDataFrame[columnDates] == 1]) / len(inputDataFrame[inputDataFrame[columnDates].notna()])))

print('# within 7 days - ' + str(len(inputDataFrame[(inputDataFrame[columnDates] <= 7) & (inputDataFrame[columnDates] > 1) & (inputDataFrame[columnPos] == 'TP')])))

print('% within 7 days - ' + str(len(inputDataFrame[(inputDataFrame[columnDates] <= 7) & (inputDataFrame[columnDates] > 1)]) / len(inputDataFrame[inputDataFrame[columnDates].notna()])))

print('# within 15 days - ' + str(len(inputDataFrame[(inputDataFrame[columnDates] <= 15) & (inputDataFrame[columnDates] > 7) & (inputDataFrame[columnPos] == 'TP')])))

print('% within 15 days - ' + str(len(inputDataFrame[(inputDataFrame[columnDates] <= 15) & (inputDataFrame[columnDates] > 7)]) / len(inputDataFrame[inputDataFrame[columnDates].notna()])))

print('# within 30 days - ' + str(len(inputDataFrame[(inputDataFrame[columnDates] <= 30) & (inputDataFrame[columnDates] > 15) & (inputDataFrame[columnPos] == 'TP')])))

print('% within 30 days - ' + str(len(inputDataFrame[(inputDataFrame[columnDates] <= 30) & (inputDataFrame[columnDates] > 15)]) / len(inputDataFrame[inputDataFrame[columnDates].notna()])))

print('# past 30 days - ' + str(len(inputDataFrame[(inputDataFrame[columnDates] > 30) & (inputDataFrame[columnPos] == 'TP')])))

print('% past 30 days - ' + str(len(inputDataFrame[inputDataFrame[columnDates] > 30]) / len(inputDataFrame[inputDataFrame[columnDates].notna()])))

# input()

print('')

print('')

print('################################')

print('PPV - ' + str(TP/(TP+FP)))

print('NPV - ' + str(TN/(TN+FN)))
```

Confidence Intervals

```
from __future__ import print_function, division

from math import sqrt

from scipy.special import ndtri

# Find the confidence interval given a proportion

# This uses the notation described in

# "Proportions and their differences 2nd Ed."

# by, D. G. Altman et al.

def confidence_interval_prop(a, b, c):

X = 2a + c^2

Y = z * sqrt(c^2 + 4a * (1 - (a/b)))

Z = 2 * (b + c^2)

return (str(round(((X - Y) / Z) * 100,2)) + '%', str(round(((X + Y) / Z) *100,2)) + '%')

# Sensitivity and specificity confidence intervals derived using Wilson's method

def sens_spef_conf_interval(TP, FP, FN, TN, alpha=0.95):

z = -ndtri((1.0 - alpha) / 2)

# Compute sensitivity using wilson's method

sens_point_est = TP / (TP + FN)

sens_conf_int = confidence_interval_prop(TP, TP + FN, z)

# Compute specificity using same method

spec_point_est = TN / (TN + FP)

spec_conf_int = confidence_interval_prop(TN, TN + FP, z)

# Compute PPV and NPV

PPV_point_estimate = TP / (TP + FP)

NPV_point_estimate = TN / (TN + FN)

PPV_confidence_interval = confidence_interval_prop(TP, TP + FP, z)

NPV_confidence_interval = confidence_interval_prop(TN, TN + FN, z)

return sens_point_est, spec_point_est, sens_conf_int, spec_conf_int, PPV_point_estimate, PPV_confidence_interval, NPV_point_estimate, NPV_confidence_interval

# Takes the #s of patients who have dates of death off by:

# 0 for same

# 1 for 'next'

# More than 1, less than or equal to 7 for 'week'

# More than 7, less than or equal to 14 for 'twoweek'

# More than 14, less than or equal to 30 for 'month'

#

def days_matching_with_confidence_interval(same, next, week, twoweek, month, total):

zero_day_point = same / total

one_day_point = (same + next) / total

week_point = (same + next + week) / total

two_week_point = (same + next + week + twoweek) / total

month_point = (same + next + week + twoweek + month) / total

zero_day_conf = sqrt(((zero_day_point * (1-zero_day_point)) / total)) * 1.96

one_day_conf = sqrt(((one_day_point * (1-one_day_point)) / total)) * 1.96

week_conf = sqrt(((week_point * (1-week_point)) / total)) * 1.96

two_week_conf = sqrt(((two_week_point * (1-two_week_point)) / total)) * 1.96

month_conf = sqrt(((month_point * (1-month_point)) / total)) * 1.96

zero_day_conf = (str(round((zero_day_point - zero_day_conf) * 100, 2)) + '%', str(round((zero_day_point + zero_day_conf) * 100, 2)) + '%')

one_day_conf = (str(round((one_day_point - one_day_conf) * 100, 2)) + '%', str(round((one_day_point + one_day_conf) * 100, 2)) + '%')

week_conf = (str(round((week_point - week_conf) * 100, 2)) + '%', str(round((week_point + week_conf) * 100, 2)) + '%')

two_week_conf = (str(round((two_week_point - two_week_conf) * 100, 2)) + '%', str(round((two_week_point + two_week_conf) * 100, 2)) + '%')

month_conf = (str(round((month_point - month_conf) * 100, 2)) + '%', str(round((month_point + month_conf) * 100, 2)) + '%')

zero_day_point = str(round(zero_day_point * 100, 2)) + '%'

one_day_point = str(round(one_day_point * 100, 2)) + '%'

week_point = str(round(week_point * 100, 2)) + '%'

two_week_point = str(round(two_week_point * 100, 2)) + '%'
```

```
month_point = str(round(month_point * 100, 2)) + '%'

return zero_day_point, zero_day_conf, one_day_point, one_day_conf, week_point, week_conf, two_week_point, two_week_conf, month_point, month_conf

# These counts are derived From the TP/FP/TN/FN of a particular group, as found in the DoD_Stats.py program

counts = [236, 88, 1691, 25]

# These days are the # that have

# [Exact match, Off by 1, More than 1 <= 7 off, more than 7 <= 14 off, more than 14 <= 30 off, > 30 off]

days = [214, 16, 3, 0, 0, 3]

TP = counts[0]

FP = counts[1]

TN = counts[2]

FN = counts[3]

zeroDay = days[0]

oneDay = days[1]

sevenDay = days[2]

fifteenday = days[3]

thirtyDay = days[4]

past = days[5]

a = 0.95

sens_point_est, spec_point_est, sens_conf_int, spec_conf_int, PPV, PPV_confidence, NPV, NPV_confidence \

= sens_spef_conf_interval(TP, FP, FN, TN, alpha=a)

print("Sensitivity: %f, Specificity: %f" % (sens_point_est*100, spec_point_est*100))

print("sensitivity:", '(' + ', '.join(sens_conf_int) + ')')

print("specificity:", '(' + ', '.join(spec_conf_int) + ')')

print("PPV: %f, NPV: %f" % (PPV*100, NPV*100))

print("PPV:", '(' + ', '.join(PPV_confidence) + ')')

print("NPV:", '(' + ', '.join(NPV_confidence) + ')')

print("")

zdp, zdc, odp, odc, wp, wc, twp, twc, mp, mc = days_matching_with_confidence_interval(zeroDay, oneDay, sevenDay, fifteenday, thirtyDay, TP)

print('zero day')

print(zdp)

print('(' + ', '.join(zdc) + ')')

print('one day')

print(odp)

print('(' + ', '.join(odc) + ')')

print('week')

print(wp)

print('(' + ', '.join(wc) + ')')

print('two week')

print(twp)

print('(' + ', '.join(twc) + ')')

print('month')

print(mp)

print('(' + ', '.join(mc) + ')')

# Chi-Squared calculation

from scipy.stats import chi2_contingency

print("Chi squared with Yates' Correction is - ", chi2_contingency(ar)[0], "\nWith a p-value of - ", chi2_contingency(ar)[1])

print("Chi squared without Yates Correction is - ", chi2_contingency(ar, correction=False)[0], "\nWith a p-value of - ", chi2_contingency(ar, correction=False)[1])
```


TABLE A1. Summary Comparison of Sensitivity for Patient Subsegments

	Sensitivity *v* NDI								Specificity *v* NDI		Date Concordance *v* NDI Within 1 Day	
Patient Subsegments	SEER	Digitized Obituaries	SSDI	Manual Curation	EHRs	Hospital Tumor Registries	Composite Score With All Sources	Composite Score Without Manual Curation	Composite Score With All Sources	Composite Score Without Manual Curation	Composite Score With All Sources	Composite Score Without Manual Curation
Race												
White	99.1	50.9	11.9	94.9	77.5	89.4	96.1	95.5	95.0	95.2	99.1	98.0
95% CI; patient count	98.7 to 99.4; 10,320	49.9 to 51.9; 55,624	11.2 to 12.6; 55,624	93.6 to 95.9; 3,835	76.6 to 78.3; 54,419	88.1 to 90.6; 9,776	95.7 to 96.5; 55,624	95.0 to 95.9; 55,624	94.8 to 95.2; 55,624	95.0 to 95.4; 55,624	98.9 to 99.3; 55,624	97.7 to 98.3; 55,624
Black or African American	95.9	31.7	12.9	90.4	75.5	82.4	95.3	95.1	94.0	94.7	99.1	98.0
95% CI; patient count	91.3 to 98.1; 498	28.4 to 35.1; 3,437	10.6 to 15.5; 3,437	84.5 to 94.2; 379	72.3 to 78.5; 3,318	75.5 to 87.6; 524	93.5 to 96.6; 3,437	92.7 to 96.0; 3,437	93.1 to 94.9; 3,437	93.8 to 95.5; 3,437	98.9 to 99.3; 3,437	97.7 to 98.3; 3,437
Asian or Pacific Islander	95.2	26.9	7.50	89.3	66.2	88.0	94.6	93.5	94.6	95.2	99.3	96.4
95% CI; patient count	91.7 to 97.3; 811	22.1 to 32.2; 1,726	5.0 to 11.1; 1,726	72.8 to 96.3; 145	60.6 to 71.4; 1,694	82.6 to 91.8; 725	91.3 to 96.6; 1,726	90.1 to 95.8; 1,726	93.3 to 95.7; 1,726	94.0 to 96.2; 1,726	98.3 to 100.3; 1,726	94.2 to 98.6; 1,726
American Indian or Alaskan Native	100.0	31.7	14.6	71.4	65.9	100.0	95.1	95.1	96.9	97.3	100.0	100.0
95% CI; patient count	80.6 to 100.0; 96	19.6 to 47.0; 296	6.9 to 28.4; 296	45.4 to 88.3; 28	50.5 to 78.4; 291	78.5 to 100.0; 79	83.9 to 98.7; 296	83.9 to 98.7; 296	93.9 to 98.4; 296	94.4 to 98.7; 296	100.0 to 100.0; 296	100.0 to 100.0; 296
Unspecified race	98.5	46.7	18.9	92.9	76.8	86.2	93.8	91.4	87.9	89.4	98.1	96.6
95% CI; patient count	97.3 to 99.2; 3,087	45.6 to 47.8; 29,910	18.1 to 19.8; 29,910	91.6 to 94.1; 6,624	75.8 to 77.8; 26,458	85.2 to 87.2; 13,265	93.3 to 94.3; 29,910	90.7 to 92.0; 29,910	87.5 to 88.3; 29,910	89.0 to 89.8; 29,910	97.7 to 98.4; 29,910	96.1 to 97.0; 29,910
Sex												
Male	98.3	53.5	15.9	93.8	76.8	86.5	95.0	94.6	94.5	94.7	98.4	96.8
95% CI; patient count	97.5 to 98.8; 5,951	52.4 to 54.6; 35,392	15.1 to 16.7; 35,392	91.7 to 95.3; 1,449	75.8 to 77.8; 33,630	85.3 to 87.5; 10,103	94.5 to 95.4; 35,392	94.0 to 95.0; 35,392	94.2 to 94.8; 35,392	94.4 to 95.0; 35,392	98.1 to 98.7; 35,392	96.4 to 97.2; 35,392
Female	98.7	49.7	15.8	94.0	76.8	87.3	94.9	94.1	95.3	95.6	98.1	96.8
95% CI; patient count	98.0 to 99.2; 7,716	48.5 to 50.9; 43,295	15.0 to 16.7; 43,295	92.4 to 95.3; 3,188	75.7 to 77.8; 41,385	86.1 to 88.4; 12,509	94.4 to 95.4; 43,295	93.5 to 94.6; 43,295	95.1 to 95.5; 43,295	95.3 to 95.8; 43,295	97.8 to 98.5; 43,295	96.4 to 97.2; 43,295
Unspecified sex	100.0	29.1	10.1	93.2	77.5	92.0	94.8	88.7	77.9	81.4	100.0	100.0
95% CI; patient count	98.9 to 100.0; 1,145	27.5 to 30.8; 12,306	9.1 to 11.3; 12,306	91.8 to 94.3; 6,374	75.9 to 78.9; 11,165	89.1 to 94.1; 1,757	94.0 to 95.6; 12,306	87.6 to 89.8; 12,306	77.1 to 78.8; 12,306	80.6 to 82.2; 12,306	100.0 to 100.0; 12,306	100.0 to 100.0; 12,306
Age												
Younger than 30	100.0	46.2	23.1	100.0	60.0	100.0	84.6	84.6	99.2	99.2	100.0	100.0
95% CI; patient count	34.2 to 100.0; 75	23.2 to 70.9; 508	8.2 to 50.3; 508	100.0 to 100.0; 3	31.3 to 83.2; 489	64.6 to 100.0; 102	57.8 to 95.7; 508	57.8 to 95.7; 508	97.9 to 99.7; 508	97.9 to 99.7; 508	100.0 to 100.0; 508	100.0 to 100.0; 508
30-59	98.5	45.1	11.0	92.0	72.3	85.2	94.0	92.5	96.1	96.6	98.6	97.2
95% CI; patient count	96.8 to 99.2; 2,891	42.8 to 47.4; 16,952	9.6 to 12.5; 16,952	88.6 to 94.5; 1,420	70.1 to 74.3; 16,257	82.5 to 87.5; 4,429	92.8 to 95.0; 16,952	91.2 to 93.6; 16,952	95.8 to 96.4; 16,952	96.2 to 96.8; 16,952	98.1 to 99.2; 16,952	96.4 to 98.0; 16,952
60 and older	98.7	48.2	15.3	93.7	77.4	87.4	95.1	93.5	91.9	92.6	98.5	97.4
95% CI; patient count	98.2 to 99.0; 11,846	47.4 to 48.9; 73,533	14.8 to 15.9; 73,533	92.8 to 94.5; 9,588	76.8 to 78.1; 69,434	86.6 to 88.2; 19,838	94.7 to 95.4; 73,533	93.1 to 93.9; 73,533	91.7 to 92.1; 73,533	92.3 to 92.8; 73,533	98.3 to 98.7; 73,533	97.1 to 97.6; 73,533
Metastatic status												
Metastatic	99.4	50.0	13.4	97.5	81.9	88.8	95.9	95.0	86.4	87.1	98.6	97.5
95% CI; patient count	98.8 to 99.7; 2,545	48.8 to 51.2; 17,614	12.6 to 14.2; 17,614	96.5 to 98.3; 2,715	81.0 to 82.8; 16,527	87.4 to 90.1; 3,976	95.4 to 96.3; 17,614	94.5 to 95.5; 17,614	85.7 to 87.0; 17,614	86.5 to 87.8; 17,614	98.3 to 98.9; 17,614	97.1 to 97.9; 17,614
Nonmetastatic	98.2	46.4	15.9	91.3	73.3	86.5	94.3	92.3	93.9	94.5	98.5	97.2
95% CI; patient count	97.5 to 98.7; 12,267	45.5 to 47.3; 73,379	15.2 to 16.6; 73,379	90.1 to 92.5; 8,296	72.4 to 74.2; 69,653	85.6 to 87.4; 20,393	93.8 to 94.7; 73,379	91.8 to 92.8; 73,379	93.7 to 94.1; 73,379	94.3 to 94.7; 73,379	98.3 to 98.8; 73,379	96.9 to 97.5; 73,379
Socioeconomic status												
Less than $30,000 USD	100.0	35.1	22.8	89.7	70.6	80.1	90.0	86.8	85.5	86.9	95.7	96.0
95% CI; patient count	87.1 to 100.0; 92	31.1 to 39.2; 2,374	19.4 to 26.6; 2,374	82.5 to 94.2; 452	66.5 to 74.3; 2,246	74.5 to 84.7; 583	87.2 to 92.3; 2,374	83.6 to 89.4; 2,374	83.8 to 87.1; 2,374	85.3 to 88.4; 2,374	93.9 to 97.6; 2,374	94.2 to 97.8; 2,374
$30,000-$59,999 USD	98.8	46.9	15.4	93.5	78.0	85.6	94.9	95.4	92.2	93.0	98.8	97.6
95% CI; patient count	98.2 to 99.2; 7,153	45.9 to 47.9; 49,766	14.7 to 16.2; 49,766	92.3 to 94.5; 6,557	77.2 to 78.8; 47,183	84.5 to 86.7; 13,279	94.5 to 95.3; 49,766	95.0 to 95.8; 49,766	91.9 to 92.4; 49,766	92.7 to 93.2; 49,766	98.6 to 99.0; 49,766	97.3 to 97.9; 49,766
$60,000-$99,999 USD	98.6	47.3	11.8	93.8	74.6	88.6	95.6	95.0	94.0	94.5	99.0	97.6
95% CI; patient count	98.0 to 99.1; 7,537	46.1 to 48.6; 36,813	11.0 to 12.7; 36,813	92.3 to 95.1; 3,801	73.5 to 75.7; 34,761	87.2 to 89.8; 10,248	95.0 to 96.0; 36,813	94.4 to 95.5; 36,813	93.8 to 94.3; 36,813	94.3 to 94.8; 36,813	98.7 to 99.2; 36,813	97.2 to 98.0; 36,813
$100,000 and higher USD	100.0	51.3	18.4	100.0	79.3	81.0	92.0	90.4	94.4	95.1	97.5	97.5
95% CI; patient count	56.5 to 100.0; 30	45.3 to 57.4; 2,040	13.9 to 23.3; 2,040	91.6 to 100.0; 201	73.8 to 83.7; 1,990	69.2 to 89.1; 259	87.9 to 94.6; 2,040	86.1 to 93.4; 2,040	93.2 to 95.4; 2,040	94.0 to 96.0; 2,040	95.5 to 99.5; 2,040	95.5 to 99.5; 2,040

Abbreviations: EHR, electronic health record; NDI, National Death Index; USD, US dollars.