HHS Public Access

Author manuscript

Cancer Causes Control. Author manuscript; available in PMC 2025 June 17.

Published in final edited form as:

Cancer Causes Control. 2025 May ; 36(5): 539–550. doi:10.1007/s10552-024-01953-6.

Registry versus claims-based index dates for studies of cancer diagnosis in administrative data

Sarah E. Soppe[1], **Sharon Peacock Hinton**[1], **Jamie C. Halula**[1], **Jennifer L. Lund**[1], **Chris D. Baggett**[1,2], **Sandi L. Pruitt**[3,4], **Megan A. Mullins**[3,4], **Ellis C. Dillon**[5], **Matthew E. Barclay**[6], **Matthew Thompson**[7], **Nicholas Pettit**[8], **Georgios Lyratzopoulos**[6], **Caroline A. Thompson**[1,2]

[1]Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Division of Cancer Epidemiology, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[3]Peter O'Donnell School of Public Health, UT Southwestern Medical Center, Dallas, TX, USA

[4]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA

[5]Center on Aging, University of Connecticut Health Center, Farmington, CT, USA

[6]Epidemiology of Cancer Healthcare and Outcomes, Department of Behavioural Science and Health, Institute of Epidemiology and Health Care (IEHC), University College London, London, UK

[7]Department of Family Medicine, University of Washington, Seattle, WA, USA

[8]Department of Emergency Medicine, Indiana University School of Medicine, Indiana University, Indianapolis, IN, USA

Abstract

Purpose: Studies of healthcare encounters leading to cancer diagnosis have increased over recent years. While some studies examine healthcare utilization before the cancer registry date of diagnosis, relevant pre-diagnosis interactions are not always immediately prior to this date due to date abstraction guidelines. We evaluated agreement of a registry date with a claims-based index and examined Emergency Department (ED) involvement in cancer diagnosis as an example of possible pre-diagnostic healthcare misclassification that could arise from improper date choice.

Methods: We implemented an algorithm to define a claims-based index as the date of the earliest International Classification of Diseases code for the cancer in Medicare and estimated agreement with the date of diagnosis from a North Carolina registry for patients diagnosed aged 66 or older

✉ Caroline A. Thompson, caroline.thompson@unc.edu.

Author contributions CAT, SES, and JCH contributed to the study conception and design. Data preparation and derivation of the index date were performed by SPH, while analysis was performed by SES. The first draft of the manuscript was written by SES and all authors commented on multiple previous versions of the manuscript. All authors read and approved the final manuscript.

Conflict of interest Sandi Pruitt reports consulting fees from Pfizer and Gilead unrelated to the current study; there are no other potential conflicts to report for other authors for the past 3 years.

with 16 cancer types from 2008 to 2017 (n = 92,056). We then classified whether each cancer was initially diagnosed through care originating in the ED using each date.

Results: The index date was identical to the cancer registry date for 47% of patients and preceded the registry date for 28%, with extent of agreement varying by cancer- and patient-specific characteristics. Agreement in ED-involved diagnosis classification using each date varied by cancer site, with sensitivity of classifications using the registry date relative to the index having a minimum of 86% for prostate and kidney cancer.

Conclusion: Studies assessing healthcare utilization proximal to cancer diagnosis should carefully consider the relevant assessment window and be aware that the use of cancer registry versus claims-based dates may impact variable classification.

Keywords

Diagnosis date; SEER–Medicare; Early diagnosis; Emergency diagnosis; Routes to diagnosis

Introduction

Early cancer diagnosis is a prevailing objective in cancer research, with observational studies of pre-diagnostic healthcare utilization critical for realizing this goal [1–5]. To assess relevant healthcare preceding cancer diagnosis in large administrative data like Medicare claims, researchers must define the time interval for healthcare encounters that likely initiated the cancer identification process. Many studies examine the period before the cancer registry-reported diagnosis date [6]. According to the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) coding and staging manual, this corresponds to the date a physician documents a suspicion of cancer or in the absence of such documentation, the date of biopsy or histologic confirmation [7], and is determined through medical chart review [8]. However, the registry diagnosis date and the date of the true start of the diagnostic episode could be separated by several months [9], with relevant patient visits preceding one date and not the other. Further, the selection of the assessment window defining the patient journey to diagnosis may hinge arbitrarily on the extent of the practitioner's documentation. For example, cancer registry guidelines indicate the date of diagnosis for prostate cancer could be the date of biopsy or, if the physician documented a suspicion of cancer, the date a prostate antigen (PSA) test was ordered before cancer was confirmed [10]. Thus, even though two patients may have identical pathways to diagnosis, one may have a much earlier registry date due to differences in physician notes. Healthcare utilization after the PSA test would be missed for one of these patients with a lookback window based on the registry date of diagnosis.

Because the registry diagnosis date is based on information from the medical record, its ability to define the beginning of the diagnostic episode may depend on referral patterns, resulting in systematic error in the classification of pre-diagnosis metrics, such as the route to diagnosis [7]. The registry date may be precipitated by many different patient encounters commensurate with the diversity of diagnostic pathways, which may vary according to cancer site, biology, healthcare access, and other patient characteristics. Early diagnosis research using national SEER registry data may be further impacted by random error, as

only the month and year of the registry date of diagnosis is provided to investigators to prevent patient reidentification, resulting in some researchers arbitrarily selecting the 1st or 15th of the month to define the pre-diagnosis assessment window [11].

Due to both this systematic and random variability arising from the choice of date used to investigate pre-diagnostic healthcare utilization, we sought to assess the use of a claims-based index date to mark the start of the relevant lookback interval in studies of the pre-diagnosis period. We implemented an NCI-recommended algorithm that identifies the date of the earliest cancer-related claim in linked Medicare claims using International Classification of Diseases (ICD) codes [8, 11], hereafter referred to as the “index date”, and evaluated its proximity to the cancer registry date by cancer site and stage at diagnosis. The index date was previously evaluated as a valid approximation for the month of diagnosis in the absence of an exact registry date [11], yet comprehensive assessment of this date for characterizing the pre-diagnosis period is lacking.

One example of a research topic relying on accurate ascertainment of the pre-diagnosis window is studies of emergency presentation (EP) of cancer, or cancer diagnosis resulting from healthcare originating in the Emergency Department (ED). EP is an important predictor of poorer cancer outcomes in international contexts [12] and is a growing topic in United States (US) literature [13–17]. Diagnosis as an EP may be related to socioeconomic disparities and patients diagnosed through an emergency route may be less likely to be treated with curative intent, although additional studies on related pre-diagnostic healthcare and symptoms are needed [18]. We evaluate use of the claims-based index vs. registry date in the classification of EPs to assess potential for misclassification when defining aspects of the pre-diagnostic interval. Most studies define EPs as ED care received in the month before the identification of cancer [12, 19, 20]. Because it relies on a narrow time interval, accurate ascertainment of this precise assessment window is especially critical for future studies to reduce EPs and improve cancer diagnosis.

Methods

Study population

This study used the North Carolina (NC) Cancer Information and Population Health Resource which includes all cancer diagnoses from the state-wide NC Central Cancer Registry (NCCCR) linked to multiple payer claims covering roughly 80% of the NC cancer population [21]. Data from the NCCCR were used for the registry date of diagnosis, while linked Medicare claims were used to derive the index date. The NCCCR abstracts dates of diagnosis using a similar process as SEER registries and exact dates are made available to investigators approved to access the data [10]. Included patients were NC residents diagnosed while alive with their first primary malignant cancer of the following types: bladder, female breast, colon, esophageal, kidney, leukemia, lung, liver, myeloma, non-Hodgkin’s lymphoma, ovarian, pancreas, prostate, rectum, stomach, and uterine from 2008 to 2017 (n = 341,559) [22]. Additional inclusion criteria were selected to match a typical SEER–Medicare population in which studies of pre-diagnostic healthcare are frequently conducted. Included individuals were at least 66 years of age at diagnosis (n = 153,521) with continuous enrollment in Medicare Parts A and B and no concurrent Medicare

Advantage enrollment − 12 to + 2 months from the diagnosis month (or until death), based on the Medicare Beneficiary Summary File (n = 98,723). Finally, patients with no relevant claim to calculate the index date were excluded, resulting in a final population of 92,056 patients. Characteristics of these patients relative to those with claims for an index date were compared in a sensitivity analysis (Online Resource 1).

Abstraction of dates

We created the index date using a previously published algorithm that identifies the earliest ICD Clinical Modification code, Ninth or Tenth Revision, for the cancer site of interest in either inpatient, outpatient, or physician claims [11, 23], considering claims ± 60 days from the NCCCR date of diagnosis (see Online Resource 2 for ICD codes). Discordance between the index date and registry date was examined in Table 1 by comparing the difference in days overall and by cancer site, stage, and patient characteristics, using the χ^2 test to evaluate the statistical significance of differences across groups. We tested whether the mean difference between the dates was significantly different from zero using a one-sample t test with a α level of 0.05. To assess potential reasons for differences between these dates, we compared them to the earliest of the date of surgical diagnostic and staging procedures or the date of first surgical treatment from the NCCCR [10]. Additionally, to compare with a previous study of this algorithm, we calculated the proportion of index dates in the same calendar month as the NCCCR date of diagnosis for breast, colorectal, and lung cancer patients [11].

Classification of emergency presentations

Previously, EPs were operationalized as patients receiving care from EDs within about one month before cancer diagnosis, so we defined EPs as those with an ED visit in the 30 days prior to cancer identification [14, 19, 20]. We assessed the dependence of the classification on the specific assessment window by identifying EPs looking back 30 days from the following dates: (1) the claims-based index date, (2) the exact cancer registry date, (3) the first of the cancer registry diagnosis month, and (4) the 15th of the cancer registry diagnosis month. To highlight random error incurred by imputing the date of diagnosis as the 1st or 15th of the month, sensitivity and specificity of the EP classifications using dates 3 and 4 were calculated using the exact registry date as the reference standard. To highlight non-random error with use of a registry date, sensitivity and specificity of the EP classification using date 2 was calculated with the index date as the reference standard. Statistics were reported separately for each cancer site. The percent misclassified was compared for each date discordance category (Online Resource 3). All statistical analyses were conducted in SAS v9.4 (SAS Institute, Cary NC), with a published macro used to generate the Sankey diagram in Fig. 2 [24].

Results

Date concordance among 16 cancer types

Relative to the registry date of diagnosis, the claims-based index date was identical for 47% of patients, earlier by ≤ 7 days for 19%, later by ≤ 7 days for 9%, and earlier or later by > 7 days for 10 and 16%, respectively (Fig. 1). The difference between the two dates was

significantly different from zero at an α level of 0.05 (Table 1). Patients were more likely to have an index date earlier than the registry date rather than identical if they were older or had a distant stage at diagnosis (Table 1). Breast, prostate, and uterus cancer patients had the greatest proportion of date agreement while patients with bladder, pancreatic, and ovarian cancer were least likely to have matching dates (Fig. 1). Patients with bladder, myeloma, and colon cancer had the highest proportion of earlier index dates, while patients with liver, ovary, and lung cancer had the highest proportion of later index dates relative to the registry date. When the index date was earlier than the registry date, it was more likely to be early by only 1 to 3 days, while patients with an index date later than the registry date tended to be later by 8 to 30 days. Less than 4% of patients had an index date that differed by more than 30 days from the registry date of diagnosis, with within-month concordance exceeding 75% for breast, lung, and colorectal cancers as previously reported (Online Resource 4) [11].

These two dates were also compared to the date of the earliest surgical procedure reported by the registry to explore potential events precipitating either date, excluding 6% of patients without microscopic confirmation unlikely to have a date of surgical diagnostic or staging procedure (n = 86,525). Of these patients, 94% had an available date of surgical diagnostic and staging procedure or a date of first surgical treatment. The earliest of these two dates matched the cancer registry date of diagnosis for 72% and the index date for 47% (Fig. 2). Among cases whose date of diagnosis matched the earliest surgical date but not the index date, the index date preceded both other dates in 69% of cases. In general, the index date was earlier than the surgical date for 36% and later for 12% (Online Resource 5). Among the 12% of total patients with a surgical date that preceded the index date, the index date fell within one week of the surgical procedure for roughly 33% and within 30 days for 83% (Online Resource 5). Patients with breast, uterus, and prostate cancer had the highest concordance between the three dates and patients with pancreatic, kidney, lung, and liver cancer had the lowest (Fig. 2). All dates in this analysis were summarized in Online Resource 6.

Classification of emergency presentations for 16 cancer types using each date

Based on the registry date of diagnosis, 26.5% of all patients were classified as EP compared to 27.1% when using the index date, with some individuals classified as an EP only when using one date versus the other (Fig. 3). Misclassification was observed even when the two dates differed by less than 7 days (Online Resource 3). EP classifications using the 1st and 15th of the registry month to define the assessment window substantially underestimated the prevalence of EPs (Fig. 3). With the exact registry date as the reference standard, sensitivity of the EP measurement using the inexact registry dates were 31 to 51% for the 1st of the month and 66 to 74% for the 15th of the month (Fig. 4; exact values in Online Resource 7). False-positive rates did not exceed 7% for any cancer site in either comparison (Fig. 4).

With the claims-based index date as the reference standard, EP classifications relying on the exact cancer registry date had sensitivities ranging from 86 to 96% (Fig. 5; exact values in Online Resource 8). Patients with prostate, kidney, and liver cancer had the lowest sensitivity for EP with the registry date while patients with liver and pancreatic cancer had the highest false-positive rate wherein a patient was classified as an EP using the registry date but not

the index date (Fig. 5). Misclassification of EP varied by stage at diagnosis, with more advanced stage cancers having higher false-positive rates (Fig. 5). Cancer sites associated with more ED visits in the claim search period tended to have more false-positive EPs. Overall, the cancer registry date had a sensitivity of 91.5% (95% CI 91.2, 91.9) and a specificity of 97.7% (95% CI 97.5, 97.8) compared to the index date.

Discussion

Choice of registry vs. claims-based index date

Registry-linked insurance claims data sources like SEER–Medicare are valuable for research on cancer diagnosis to identify interventions to reduce disparities and improve outcomes. However, identifying the relevant lookback period in these studies is subject to misclassification error, particularly for those with a narrow search window. Here, we employed the earliest ICD code for the cancer as a proxy for when cancer was first suspected by the healthcare system, serving as our index date. This date was identical to the date of diagnosis reported by the registry for 47% of patients and earlier than the registry date for 28%. The extent of variation between these dates differed by type of cancer and stage at diagnosis, indicating that disease-specific factors influence the timing of the registry date of diagnosis. This information may inform choices of which date is most appropriate to define the lookback period, a study-specific choice that may lead to misclassification if not appropriately considered.

The registry and index date differed by 7 days or less for 75% of patients, indicating that healthcare utilization documented over long periods of time may be minimally affected by the choice of date. However, acute factors directly precipitating identification of cancer may be misclassified with date discordance of even a few days, as these would appear prior to cancer diagnosis under one date and after diagnosis using the other. Our example EP measure assessing ED visits over a 30-day period showed that patients with an index date early or late by only 7 days or less were still 3–7% misclassified, respectively (Online Resource 3), although misclassification is expected to be greater for measures with a narrower search window. For example, researchers evaluating prognosis of patients diagnosed during an inpatient hospitalization need to ascertain the exact date that suspicion of cancer began, as an inpatient stay immediately prior to this date reflects diagnostic care, while a visit immediately after may simply reflect symptom management.

One quarter of the patients had a registry date of diagnosis earlier than the index date, yet this may not necessarily indicate cancer was suspected before the index. According to cancer registration guidelines [10], physician documentation that a patient likely had cancer at an earlier date, such as the date of a previous surgical or imaging finding previously believed to be benign, should be used for the diagnosis date even if providers had no suspicion of cancer at that time [10]. In our example of EP, ED utilization contributing to the ultimate identification of cancer would occur after this retrospective date, indicating that the earliest available “date of diagnosis” may not always confer the least misclassification. While researchers should carefully consider the most appropriate date for their specific study, the index date directly corresponds to healthcare interaction where cancer was suspected and may better approximate events leading to the identification of cancer than the registry date.

Previously, Lin and Virgo showed high concordance between the SEER month of diagnosis and the Medicare claims index date for breast, colorectal, and lung cancer cases among SEER–Medicare patients diagnosed 2002 to 2007 [11]. Our 2007 to 2018 cohort in NC displayed similar trends, as index dates for breast and lung cancer patients were more frequently later than the registry date, while those for colorectal cancer patients tended to be earlier when the two dates disagreed. Within-month concordance was slightly higher in our cohort, which may be indicative of changes in administrative, coding, or diagnostic practices over time. Setoguchi et al. reported lower concordance between the registry and index dates for leukemia, lymphoma, lung, colorectal, stomach, and breast cancers than those shown here, though their study population was diagnosed earlier (1997–2000) [25]. Another study examined use of the date of the earliest ICD code in electronic health records for prostate cancer and found substantially lower agreement with the registry date of diagnosis, although critically their cohort did not require pre-diagnostic enrollment and some patients may have had patient-provided index dates not corresponding with a specific medical encounter [26]. Finally, Lin and Virgo also reported that patients whose index date preceded the SEER month were more likely to be older and non-White relative to those with a concordant date [11], which matches our findings.

Concordance between the index and registry date may be influenced by differences in the method of detection for each cancer site and variations in technical reporting practices by cancer registries. The site- and stage-specific variation we observed may be related to the type of diagnosis abstracted by the registrar—clinical, pathologic, or both. For instance, according to the NCCCR reporting manual, the date of diagnosis for a patient with elevated prostate antigen (PSA) should correspond to the date of the PSA only if the physician documents a suspicion of cancer, otherwise the date of biopsy should be used [10]. Thus, the registry date may differ across two patients with identical healthcare utilization, whereas the claims-based index based on billing is less vulnerable to variabilities in the extent of documentation. For 62% of prostate cancer diagnoses, the registry date of diagnosis matched the index date, likely the date of biopsy. The majority of the remaining 38% had an index date later than the date reported by the cancer registry, suggesting the date a physician documented suspicion of cancer may have been used as the registry date, which would precede the date of biopsy and other diagnostic testing. Alternatively, our results for liver cancer suggest index dates frequently occurred both before and after the registry date. Radiologic imaging can be used to diagnose liver cancer and may replace the need for biopsy in some cases, while serum-based testing for liver cancer is also possible [27], suggesting there may be multiple ways the initial ICD code for liver cancer appears in claims. However, the NCCCR coding manual currently states the date of biopsy should be selected as the liver cancer date of diagnosis even when radiologic imaging was previously performed, although these Standards for Oncology Registry Entry instructions have fluctuated over time [10].

We hypothesized the registry date of diagnosis would correspond to dates of surgical diagnostic and staging procedures or surgical treatment for a large majority of patients, with the index date often matching or preceding this date due to earlier healthcare encounters with suspicions of cancer. This trend was observed for 62% of patients with microscopic confirmation, although 10% had an index date that occurred after the matching surgical

and diagnosis dates. We suspect this 10% may be explained by a combination of the following: (1) use of an ICD code for a condition other than the conclusive cancer type at the time of surgical procedure, such as a related suspected cancer type or mass or (2) random fluctuations and errors in the timing of claims or notes in the health record. The first explanation suggests healthcare workup for cancer could have started before the earliest cancer-specific ICD code and highlights the dependence of the index date on the specific codes used. These fluctuations should be considered and explored in subsequent analysis. Nevertheless, 22% of patients with microscopic confirmation had a date of diagnosis corresponding to the date of earliest surgery with an ICD code for the cancer before either of these dates. Thus, the index date is likely a better approximation of initial healthcare encounters leading to diagnosis for more patients, while reliance on the registry date likely misses initial patient visits.

Importantly, researchers estimating time-to-treatment should be aware that the registry date of diagnosis is frequently equivalent to the date of first surgery [28]. For this reason the index date might be preferred for measuring timeliness of diagnostic workup and referral for treatment. Using the registry date of diagnosis or date of surgery to measure treatment delay would likely underestimate the time from initial workup for cancer to treatment, with the impact of long delays between the time of symptomatic presentation and surgery overlooked.

Classification of emergency presentations

We classified EPs by looking back 30 days from each date to determine whether the ED was involved in cancer diagnosis. Researchers employ a variety of dates to identify EPs depending on available data. Because SEER does not provide an exact registry date to investigators to protect patient privacy, some researchers have imputed the 1st or 15th of the registry-reported month as the date of diagnosis [11]. We showed reliance on these dates leads to substantial underestimation of EP, as relevant ED visits preceding the exact cancer registry date were frequently missed. Thus, inexact imputed dates subject to random error should be avoided when measuring pre-diagnostic healthcare.

When exact registry dates are available, as in our state-wide NCCCR data, they may be subject to disease-specific variation driven by differences in healthcare encounters, as discussed above. Here, we demonstrated that relying on the registry date of diagnosis resulted in a different classification of EP than when using the index date, with a false negative proportion of 8.5% and a false-positive proportion of 2.3%. Sensitivity was particularly low for patients with prostate, kidney, and liver cancers. Assuming the index date better reflects the start of the diagnostic episode, reliance on the cancer registry date may result in non-random misclassification of the route to diagnosis, as it does not always immediately precede the relevant healthcare encounters that ultimately led to identification of cancer.

Strengths and limitations

While our findings are robust and provide support for the use of a claims-based index date when studying pre-diagnostic healthcare utilization, some limitations should be considered.

Roughly 9% of patients were excluded due to lack of cancer-specific claims to assign an index date. Thus, the index may not be available for all patients which could result in selection bias if the exposure or outcome of interest is related to factors resulting in few claims. Patients with no index date were more likely to be older males and less likely to have a cancer of the breast, colon, or prostate, indicating cancers detectable through screening may be less likely to lack a diagnostic code around the time of the registry date (Online Resource 1).

We also note that the index date was assigned based on claims within 60 days of the NCCCR date of diagnosis. Although unlikely for the majority of patients, the earliest ICD code for the cancer may be more than 60 days before the diagnosis, which would be missed under the algorithm described here. It is also possible that patients may have healthcare interactions missing from Medicare claims such as due to care received through Veterans Affairs facilities, which could lead to potential inaccuracies in the index date. Our inclusion criteria were designed to minimize this by ensuring all patients had continuous fee-for-service claims in the 1-year pre-diagnosis period and were not concurrently enrolled in Medicare Advantage. This issue emphasizes the need for cancer registry linkage in our approach, as diagnostic codes may not correspond to the timing of initial cancer diagnosis if the general timeframe is not provided by the registry.

While this study provided assessment of date discordance in a Medicare population of older adults, generalizability to younger populations and electronic health record (EHR) data was not possible due to our study data and sample definitions. Further, we excluded Medicare Advantage beneficiaries, a population that grew over the study period, limiting the number of individuals that could be included and potentially impacting demographics of the study population [29]. Sensitivities and specificities presented here could also change over time as routes to diagnosis and administrative practices change.

Some unmeasurable factors may have also influenced the reported proportion of EPs, such as individuals unaccounted for in the registry who presented to the ED and passed away before tissue sampling. Because ED involvement in routing potential cancer patients may be several days or weeks before cancer is ultimately identified through referrals or later workup, not all ED claims of EP patients include an ICD code for the cancer. Because we followed prior literature and classified EPs as those with any ED claim in the 30 days before diagnosis, some ED visits leading up to diagnosis could be incidental and unrelated to the cancer, although we expect these to be rare.

Conclusion

Understanding pre-diagnostic healthcare utilization is critical to increase early detection of cancer [1–5]. There are many possible dates that might approximate the start of the diagnostic interval, with the validity of the observation window hinging on the date choice. Nevertheless, how this “date of diagnosis” was defined is often overlooked. We assessed agreement between different dates to define the window for pre-diagnosis healthcare utilization and examined how use of these different dates impacts the classification of EPs. Our findings illustrate that selection of the relevant assessment window is critical in

accurately defining the route to diagnosis. While the appropriate date should be considered on a case-by-case basis, use of the registry date may lead to misclassification in studies of the healthcare utilization that ultimately led to cancer diagnosis, a growing area of research to improve outcomes and reduce disparities in cancer care.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sanah Vohra, PhD, should be acknowledged for contributions at the study design phase. We thank Ellen Green, MD, for her comments and feedback.

Funding

This research was funded by the National Cancer Institute through Grant R01-CA-264176. Work on this study was supported by the Cancer Information and Population Health Resource, UNC Lineberger Comprehensive Cancer Center, with funding provided by the University Cancer Research Fund via the State of North Carolina. The findings and conclusions in this publication are those of the author(s) and do not necessarily represent the views of the North Carolina Department of Health and Human Services, Division of Public Health. SES was supported by a grant from UNC Lineberger Comprehensive Cancer Center, which is funded by the Susan G Komen Foundation (TREND21686258). MEB is supported by a Cancer Research UK International Alliance for Cancer Early Detection (ACED) Pathway Award (EDDAPA-2022/100002) and NP is supported by an ACS K Award (CSDG-23-1025372-01-CPSH).

Data availability

The data analyzed during the current study are not publicly available due to protections for individual privacy by the North Carolina Central Cancer Registry.

References

1. Crosby D, Lyons N, Greenwood E et al. (2020) A roadmap for the early detection and diagnosis of cancer. Lancet Oncol 21(11):1397–1399. 10.1016/S1470-2045(20)30593-3 [PubMed: 33031732]
2. Whitfield E, White B, Denaxas S, Barclay ME, Renzi C, Lyratzopoulos G (2023) A taxonomy of early diagnosis research to guide study design and funding prioritisation. Br J Cancer 129(10):1527–1534. 10.1038/s41416-023-02450-4 [PubMed: 37794179]
3. Sarma EA, Kobrin SC, Thompson MJ (2020) A proposal to improve the early diagnosis of symptomatic cancers in the United States. Cancer Prev Res 13(9):715–719. 10.1158/1940-6207.CAPR-20-0115
4. Flaubert JL, Cooper R, Snair M, Nass S (2022) Achieving excellence in cancer diagnosis: proceedings of a workshop—in brief, 2022. National Academies Press. 10.17226/26505
5. Delamare Fauvel A, Bischof JJ, Reinbolt RE et al. (2023) Diagnosis of cancer in the Emergency Department: a scoping review. Cancer Med 12(7):8710–8728. 10.1002/cam4.5600 [PubMed: 36622062]
6. Hornbrook M, Fishman P, Ritzwoller D, Elston-Lafata J, O'Keeffe-Rosetti M, Salloum R (2013) When does an episode of care for cancer begin? Med Care 51(4):324–329. 10.1097/MLR.0b013e3182731277 [PubMed: 23222500]
7. Adamo M, Dickie L, Ruhl J (2018) SEER program coding and staging manual. US Department of Health and Human Services, National Institutes of Health, National Cancer Institute, Bethesda
8. National Cancer Institute (2019) SEER–Medicare: defining the date of diagnosis and treatment. National Cancer Institute Division of Cancer Control and Population Sciences web site. https://healthcaredelivery.cancer.gov/seermedicare/considerations/date.html. Accessed 6 May 2023

9. Myklebust TÅ, Andersson T, Bardot A et al. (2020) Can different definitions of date of cancer incidence explain observed international variation in cancer survival? An ICBP SURVMARK-2 study. Cancer Epidemiol. 10.1016/j.canep.2020.101759
10. N.C. Division of Public Health (2023) CCARM cancer collection and reporting manual effective with 1/1/2023 diagnoses. North Carolina Central Cancer Registry. N.C. Division of Public Health web site. https://schs.dph.ncdhhs.gov/schs/CCR/docs/CCARM2018_05012019.pdf. Accessed 6 May 2023
11. Lin CC, Virgo KS (2014) Diagnosis date agreement between SEER and Medicare claims data: impact on treatment. Med Care 52(1):32–37. 10.1097/MLR.0000000000000004 [PubMed: 24322987]
12. McPhail S, Swann R, Johnson SA et al. (2022) Risk factors and prognostic implications of diagnosis of cancer within 30 days after an emergency hospital admission (emergency presentation): an International Cancer Benchmarking Partnership (ICBP) population-based study. Lancet Oncol 23(5):587–600. 10.1016/S1470-2045(22)00127-9 [PubMed: 35397210]
13. Thompson CA, Sheridan P, Metwally E et al. (2024) Emergency department involvement in the diagnosis of cancer among older adults: a SEER–Medicare study. JNCI Cancer Spectr 8(3):pkae039. 10.1093/jncics/pkae039/7682377 [PubMed: 38796687]
14. Kapadia P, Zimolzak AJ, Upadhyay DK et al. (2024) Development and implementation of a digital quality measure of emergency cancer diagnosis. J Clin Oncol. 10.1200/JCO.23.01523
15. Ray EM, Elston Lafata J, Reeder-Hayes KE, Thompson CA (2024) Predicting the future by studying the past for patients with cancer diagnosed in the emergency department. J Clin Oncol. 10.1200/JCO.24.00480
16. Kronenfeld JP, Ryon EL, Goldberg D et al. (2021) Disparities in presentation at time of hepatocellular carcinoma diagnosis: a United States Safety-Net Collaborative Study. Ann Surg Oncol 28(4):1929–1936. 10.1245/s10434-020-09156-4 [PubMed: 32975686]
17. Pettit N, Cinbat M (2021) Retrospective review of patients with lung cancer identified in the emergency department. Am J Emerg Med 50:394–398. 10.1016/j.ajem.2021.08.040 [PubMed: 34478945]
18. Zhou Y, Abel GA, Hamilton W et al. (2017) Diagnosis of cancer as an emergency: a critical review of current evidence. Nat Rev Clin Oncol 14(1):45–56. 10.1038/nrclinonc.2016.155 [PubMed: 27725680]
19. Elliss-Brookes L, McPhail S, Ives A et al. (2012) Routes to diagnosis for cancer—determining the patient journey using multiple routine data sets. Br J Cancer 107(8):1220–1226. 10.1038/bjc.2012.408 [PubMed: 22996611]
20. Sikka V, Ornato JP (2012) Cancer diagnosis and outcomes in Michigan EDs vs other settings. Am J Emerg Med 30(2):283–292. 10.1016/j.ajem.2010.11.029 [PubMed: 21247723]
21. Meyer AM, Olshan AF, Green L, Wheeler SB, Basch E, Carpenter WR (2014) Big data for population-based cancer research: the integrated cancer information and surveillance system. NC Med J 75(4):265–269. 10.18043/ncm.75.4.265
22. National Cancer Institute (nd) Site recode ICD-O-3/WHO 2008 definition. National Cancer Institute Surveillance, Epidemiology, and End Results Program web site. https://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html. Accessed 6 May 2023
23. McClish D, Penberthy L (2004) Using Medicare data to estimate the number of cases missed by a cancer registry a 3-source capture–recapture model. Med Care 42(11):1111–1116. 10.1097/00005650-200411000-00010 [PubMed: 15586838]
24. Rosanbalm S (2015) Getting Sankey with bar charts (PharmaSUG web site). https://www.pharmasug.org/proceedings/2015/DV/PharmaSUG-2015-DV07.pdf. Accessed 19 June 2023
25. Setoguchi S, Solomon DH, Glynn RJ, Cook EF, Levin R, Schneeweiss S (2007) Agreement of diagnosis and its date for hematologic malignancies and solid tumors between Medicare claims and cancer registry data. Cancer Causes Control 18(5):561–569. 10.1007/s10552-007-0131-1 [PubMed: 17447148]
26. Porter KR, Chao C, Quinn VP, Hsu JWY, Jacobsen SJ (2014) Variability in date of prostate cancer diagnosis: a comparison of cancer registry, pathology report, and electronic health data sources. Ann Epidemiol 24(11):855–860. 10.1016/j.annepidem.2014.09.004 [PubMed: 25282324]

27. Yang JD, Heimbach JK (2020) New advances in the diagnosis and management of hepatocellular carcinoma. BMJ. 10.1136/bmj.m3544

28. Hanna TP, King WD, Thibodeau S et al. (2020) Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ 371:m4087. 10.1136/bmj.m4087 [PubMed: 33148535]

29. Enewold L, Parsons H, Zhao L et al. (2020) (2020) Updated overview of the SEER–Medicare data: enhanced content and applications. J Natl Cancer Inst Monogr 55:3–13. 10.1093/jncimonographs/lgz029

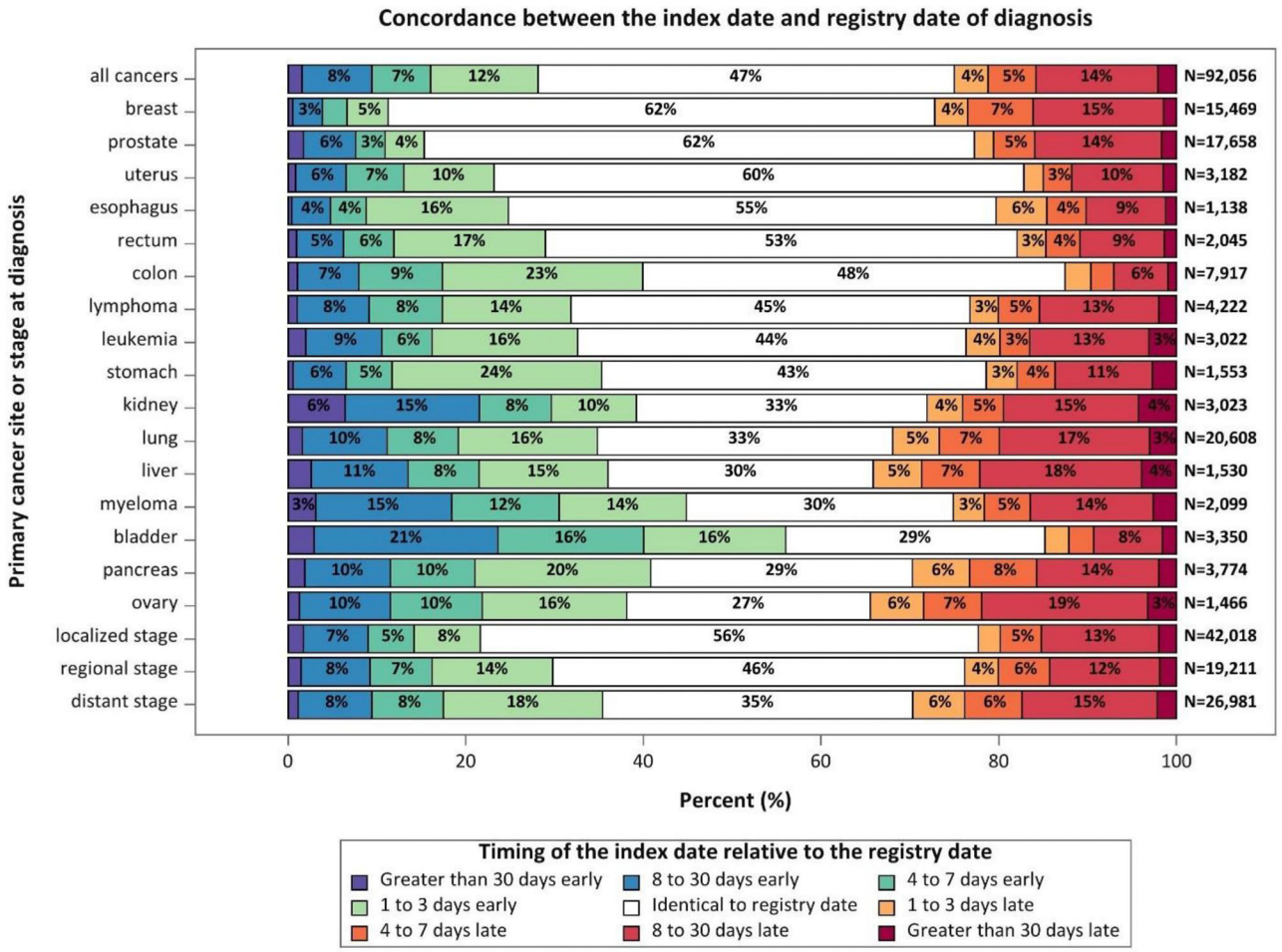

Fig. 1.

Timing of the index date relative to the North Carolina Central Cancer Registry (NCCCR) date of diagnosis by cancer site and stage. Bars are ordered by those with the highest proportion of agreement between the dates to the lowest proportion. Segments to the left are indicative of an index date earlier than the NCCCR date while bars to the right are indicative of index dates later than the NCCCR date

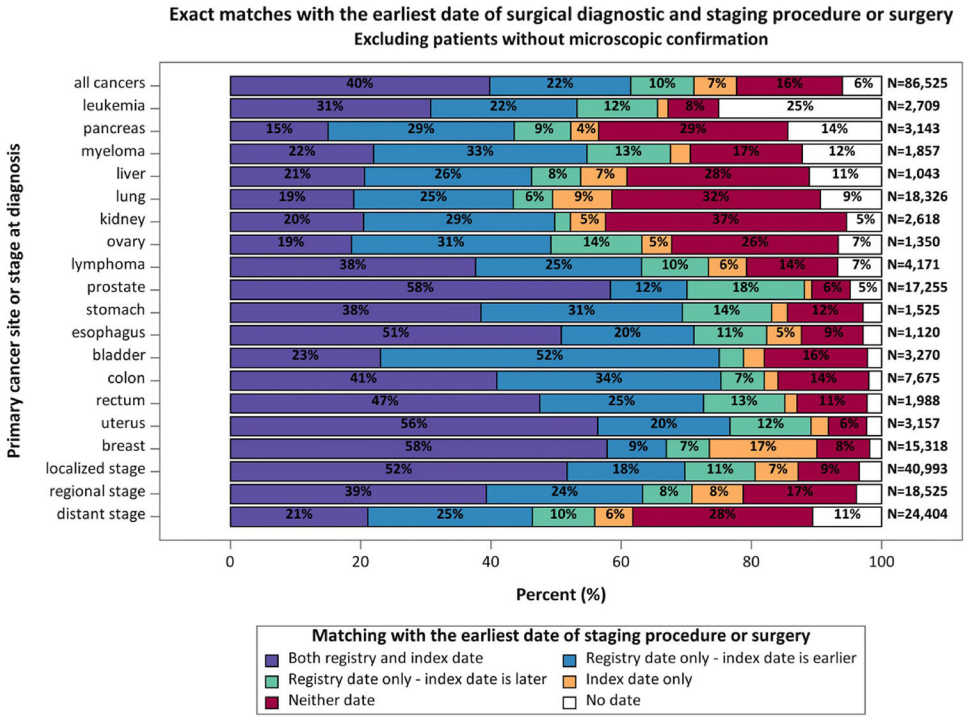

Fig. 2.

Concordance of the cancer registry date and the index date relative to the date of the earliest surgical diagnostic and staging procedure or the date of first surgical treatment. Patients without microscopic confirmation of the cancer diagnosis are excluded

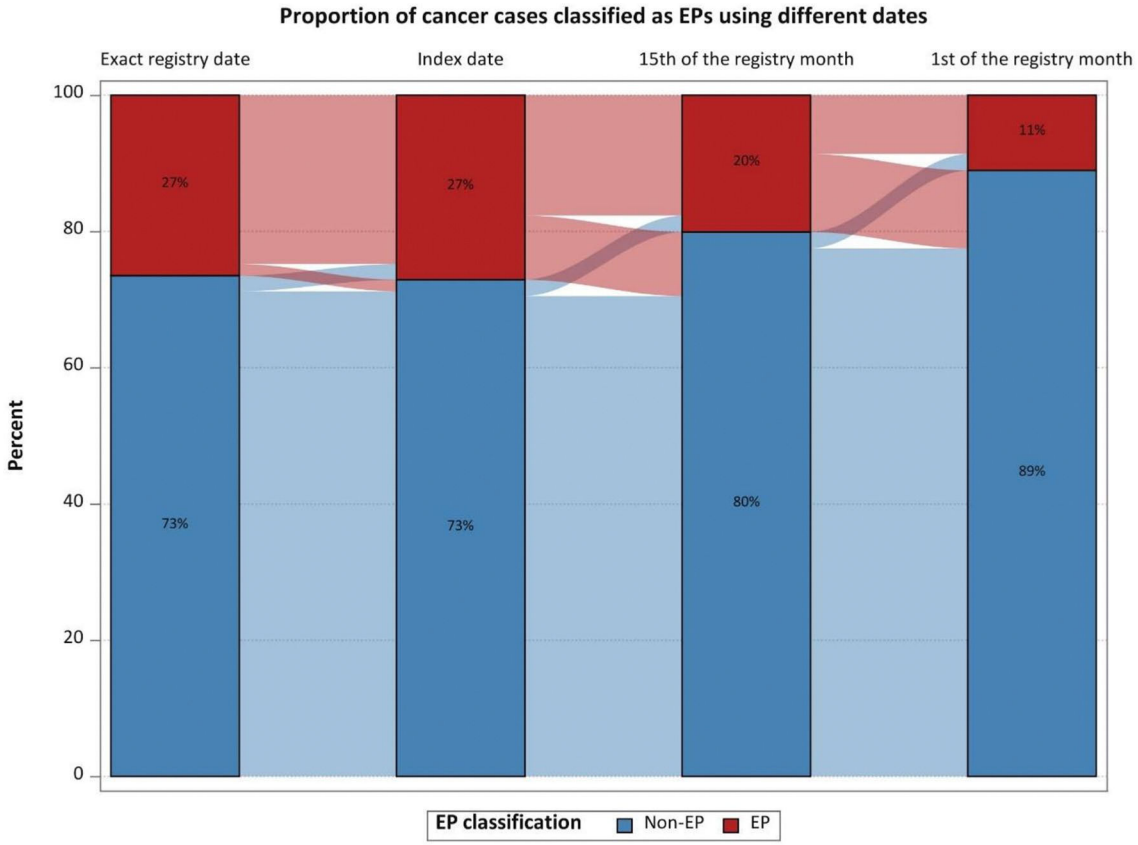

Fig. 3.

The proportion of individuals classified as an EP using different dates to define the pre-diagnostic assessment window

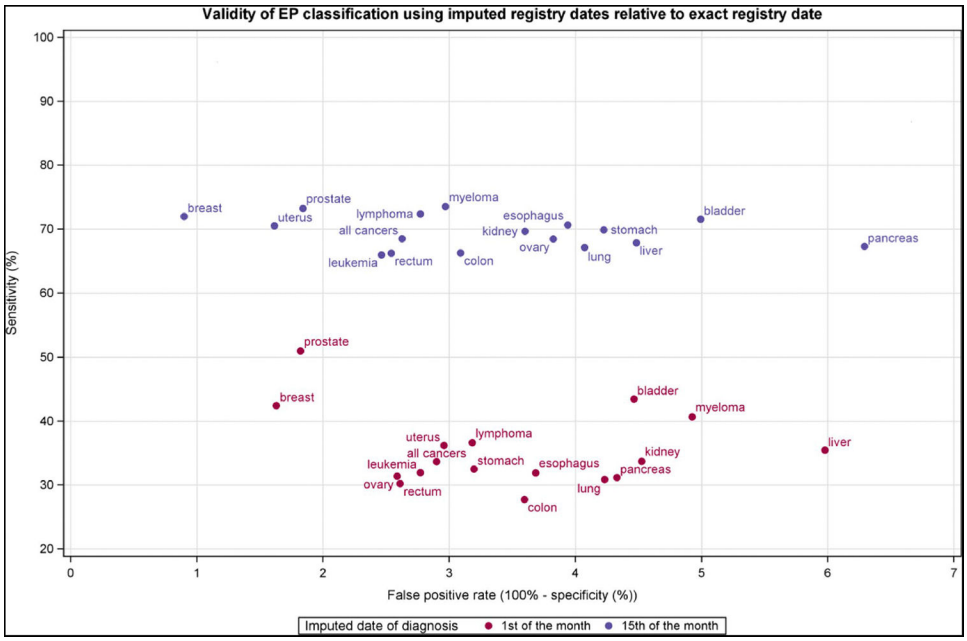

Fig. 4.

Specificity by false-positive rate of the EP classification using the inexact cancer registry dates of diagnosis relative to the exact registry date by cancer site. EP classifications using the inexact date for cancer sites in the upper left-hand quadrant have the best performance relative to the exact registry date, while those in the lower right have the poorest performance

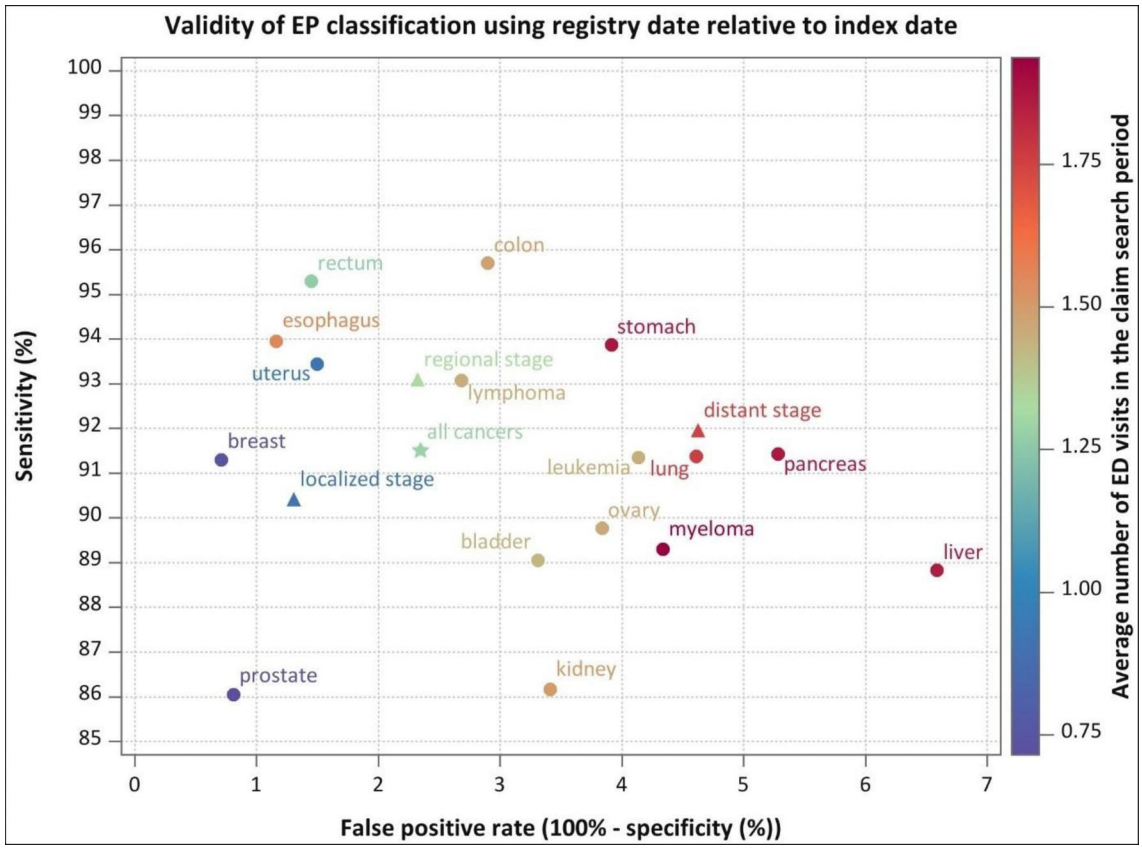

Fig. 5.

Specificity by false-positive rate of the EP classification using the registry date of diagnosis relative to the index date by cancer site or stage. Points are colored by the average number of ED visits in the claim search period (− 12 to + 2 months from the registry month of diagnosis). EP classifications using the registry date in the upper left-hand quadrant have the best performance relative to using the index date, while those in the lower right have the poorest performance.

Table 1

Characteristics of the study population by timing of the claims-derived index date relative to the cancer registry date of diagnosis

	Overall		Index date > 7 days earlier	Index date earlier by ≤ 7 days	Index date the same as registry date	Index date later by ≤ 7 days	Index date > 7 days later	*p* value[b]
	n	(%)	(%)	(%)	(%)	(%)	(%)	
All participants	92,056	100	9.5	18.7	46.9	9.2	15.8	< 0.0001
Sex[a]								
Male	46,221	50.2	10.4	18.3	46.7	8.6	16.1	< 0.0001
Female	45,834	49.8	8.6	19.1	47	9.8	15.5	
Age at diagnosis								
66–69	23,071	25.1	9.1	15.5	49.8	9.1	16.5	< 0.0001
70–74	25,705	27.9	9	16.7	48.4	9.6	16.3	
75–79	19,304	21	9.8	19.2	45.8	9.4	15.9	
80–84	13,347	14.5	10.2	21.8	44.6	8.4	15	
85+	10,629	11.5	10.1	25.7	41.4	8.8	14	
Race/ethnicity (Medicare)								
White	75,110	81.6	9.4	18.4	47	9.4	15.8	< 0.0001
Black	14,963	16.3	9.9	20.1	45.6	8.3	16	
Asian	382	0.4	9.2	21.5	45.5	9.9	13.9	
Hispanic	240	0.3	10	17.5	52.1	5.4	15	
North American Native	365	0.4	8.5	21.9	48.5	9.6	11.5	
Another racial/ethnic identity	642	0.7	7.9	17.4	50	8.7	15.9	
Unknown	354	0.4	7.1	13.6	60.5	4.5	14.4	
Quartile of Yost Index (block group 2009–2013)								
First quartile (lowest neighborhood-level SES)	13,594	14.8	9.3	15.8	49.3	9.8	15.9	< 0.0001
Second quartile	21,430	23.3	9.4	17.5	47.5	9.4	16.2	
Third quartile	29,571	32.1	9.3	19.1	46.5	9.3	15.8	
Fourth quartile (highest neighbor-hood-level SES)	26,982	29.3	9.7	20.7	45.5	8.6	15.5	
Unknown	479	0.5	11.5	16.3	50.1	7.9	14.2	
Primary cancer site								

	Overall		Index date > 7 days earlier	Index date earlier by ≤ 7 days	Index date the same as registry date	Index date later by ≤ 7 days	Index date > 7 days later	*p* value[b]
	n	(%)	(%)	(%)	(%)	(%)	(%)	
Bladder	3,350	3.6	23.6	32.4	29.2	5.5	9.3	< 0.0001
Breast	15,469	16.8	3.9	7.3	61.5	11.1	16.1	
Colon	7,917	8.6	7.9	32	47.5	5.5	7	
Esophagus	1,138	1.2	4.7	20	54.9	10.1	10.2	
Kidney	3,023	3.3	21.5	17.7	32.7	8.6	19.5	
Leukemia	3,022	3.3	10.5	22.1	43.7	7.2	16.5	
Liver	1,530	1.7	13.5	22.5	29.9	12	22.2	
Lung	20,608	22.4	11.2	23.6	33.2	12	19.9	
Myeloma	2,099	2.3	18.4	26.4	30.1	8.7	16.4	
Non-Hodgkin’s lymphoma	4,222	4.6	9.1	22.7	44.9	7.8	15.4	
Ovary	1,466	1.6	11.5	26.6	27.4	12.6	21.9	
Pancreas	3,774	4.1	11.5	29.3	29.5	14	15.7	
Prostate	17,658	19.2	7.7	7.7	61.9	6.8	16	
Rectum	2,045	2.2	6.3	22.7	53.1	7.1	10.9	
Stomach	1,553	1.7	6.5	28.8	43.3	7.8	13.7	
Uterus	3,182	3.5	6.5	16.7	59.6	5.4	11.8	
SEER Summary stage								
Localized	42,018	45.6	9	12.7	56	7.1	15.2	< 0.0001
Regional	19,211	20.9	9.2	20.6	46.3	9.6	14.2	
Distant	26,981	29.3	9.4	26	34.9	12.3	17.4	
Unknown	3,846	4.2	16.4	23.2	33.3	7.6	19.4	

[a] Includes one patient with missing data

[b] *p* value for the difference between dates (first row) was calculated using a *t* test. Remaining *p* values were calculated using the χ^2 test