Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health, SEER, and NPCR

Xinran Ma, MS; Lura Long; Sharon Moon; Blythe J.S. Adamson, PhD, MPH; Shrujal S. Baxi, MD, MPH

Flatiron Health, Inc., New York, NY

Corresponding author:

Xinran Ma, MS
Flatiron Health
233 Spring St
New York, NY 10013
E-mail: xma@flatiron.com

Keywords: cancer; real-world data; cancer registries

Running title: Real-world oncology data sources in the US

Manuscript length: Abstract, 366 words. Text, 3913 words. Tables, 20. Figures, 1. Appendices, 2.

Funding: This study was sponsored by Flatiron Health, Inc. (Flatiron Health), which is an independent member of the Roche group.

Disclosures:
At the time of the study, all authors report employment at Flatiron Health, Inc., which is an independent member of the Roche Group, and stock ownership in Roche. SSB, LL own equity in Flatiron Health.

Author roles and contributions:
Study design and concept: XM, LL, SM, BJSA, SSB
Data collection: Flatiron Health
Data analysis and interpretation: XM, LL, SSB
Manuscript writing, review and approval: All

ABSTRACT

Background and Objective

The Surveillance, Epidemiology, and End Results Program (SEER) program and the National Program of Cancer Registries (NPCR), are authoritative sources for population cancer surveillance and research in the US. An increasing number of recent oncology studies are based on the electronic health record (EHR)-derived de-identified databases created and maintained by Flatiron Health. This report describes the differences in the originating sources and data development processes, and compares baseline demographic characteristics in the cancer-specific databases from Flatiron Health, SEER, and NPCR, to facilitate interpretation of research findings based on these sources.

Methods

Patients with documented care from January 1, 2011 through May 31, 2019 in a series of EHR-derived Flatiron Health de-identified databases covering multiple tumor types were included. SEER incidence data (obtained from the SEER 18 database) and NPCR incidence data (obtained from the US Cancer Statistics public use database) for malignant cases diagnosed from January 1, 2011 to December 31, 2016 were included. Comparisons of demographic variables were performed across all disease-specific databases, for all patients and for the subset diagnosed with advanced-stage disease.

Results

As of May 2019, a total of 201,570 patients with 19 different cancer types were included in Flatiron Health datasets. In an overall comparison to national cancer registries, patients in the Flatiron Health databases had similar sex, age at initial diagnosis, and geographic distributions but appeared to be diagnosed with later stages of disease compared with patients in other

datasets. For variables such as stage and race, Flatiron Health databases had a greater degree of incompleteness. There are variations in these trends by cancer types.

Conclusions

These three databases present general similarities in demographic and geographic distribution, but there are overarching differences across the populations they cover. Differences in data sourcing (medical oncology EHRs vs cancer registries), and disparities in sampling approaches and rules of data acquisition may explain some of these divergences. Furthermore, unlike the steady information flow entered into registries, the availability of medical oncology EHR-derived information reflects the extent of involvement of medical oncology clinics at different points in the specialty management of individual diseases, resulting in inter-disease variability. These differences should be considered when interpreting study results obtained with these databases.

INTRODUCTION

The field of oncology is undergoing rapid evolution as our understanding of the pathophysiology of cancer expands and therapeutic progress accelerates. This environment calls for the development of tools that facilitate the translation of these advances into improvements in patient care. Recent years have seen increasing use of real-world data (RWD) as a clinical research source, from descriptive epidemiology to intervention effectiveness studies. RWD analyses generate real-world evidence (RWE) that could supplement and complement the evidence for new drug approvals (traditionally gathered from prospective studies), for health services research, policy evaluation, or as a pharmacovigilance tool (1,2).

RWD can be obtained from many different sources, including billing and administrative claim activities, product and disease registries, national surveys, and electronic health records (EHRs). The appropriate and optimal source of RWD will differ by the research question. Traditionally, registries have been a key RWD source for epidemiologic and population-based outcomes studies; in the US, the Surveillance, Epidemiology, and End Results program (SEER) registries and the National Program of Cancer Registries (NPCR), have been commonly used for oncology research. The SEER Program (3), supported by the Surveillance Research Program (SRP) in the National Cancer Institute's (NCI) Division of Cancer Control and Population Sciences (DCCPS), collects data on patient demographics, primary tumor site, morphology and stage at diagnosis, and first course of treatment; the SEER program provides cancer incidence and survival data through 16 state level population-based cancer registries currently covering approximately 34.6% of the US population. The NPCR (4) is supported by the US Centers for Disease Control and Prevention (CDC) Division of Cancer Prevention and Control, covers 97% of the US cancer population, spans 46 states, the District of Columbia, Puerto Rico, the US Pacific Island Jurisdictions, and the US Virgin Islands.

Following the passage of the Health Information Technology for Economic and Clinical Health Act (HITECH) (5) in January 2009, EHR systems have been rapidly adopted in the US and are now a key source of RWD. In the field of oncology, adoption has been even swifter than in other fields, and by 2015, approximately 90% of oncology practices had already adopted EHRs (6, 7). This, accompanied by an exponential growth of data storage and mining technology, has allowed for access to detailed clinical data at an unprecedented scale. In oncology, patient-level data can be derived from EHRs to generate granular information about baseline population characteristics as well as longitudinal views of sequential treatments, interventions, and associated outcomes. Flatiron Health is an oncology-focused health technology company that generates RWD from two EHR-derived primary sources: (i) OncoEMR®, a proprietary oncology-specific EHR used by community oncologists throughout the US, and (ii) EHR data integrations with academic research centers that enable bidirectional transmission of RWD.

The underlying data collection procedures and data model architecture make Flatiron Health a fundamentally different data source from SEER or NPCR (Table 1). SEER and NPCR have been well established as research resources since their inception in 1973 and 1992, respectively. As the most recent of the three sources, data derived from Flatiron Health have become a research resource in the last five years (8-18); therefore, it has become increasingly critical to understand their features. SEER and NPCR collect specific incident disease data points in a systematic and ordered fashion, fulfilling a public health reporting mandate justified by the public health burden of cancer as a disease. Flatiron Health data collection mirrors routine oncology EHR documentation practices; upon curation, this approach yields longitudinal clinical data models with considerable depth, including clinical, genetic, and outcome data. This feature, together with its 30-day recency, makes this a suitable source for detailed investigation of contemporary trends in cancer management, including sequential time-to-event endpoints.

Researchers considering Flatiron Health data for their studies may be interested in gaining insights into the nature of the originating source, on the principles of data collection, and on the characteristics of its patient population. Descriptive statistics of the Flatiron Health, SEER, and NPCR data can be informative for the interpretation of findings when these large data sets are used in cancer research. This study aims to describe data sources and collection procedures, and to provide a detailed comparison of the demographic characteristics in the disease-specific databases from Flatiron Health, SEER, and NPCR, identifying similarities and differences within the data elements common across the three.

METHODS

Flatiron Health databases

Flatiron Health has developed de-identified disease-specific dynamic databases (termed enhanced data marts) derived from information available in an EHR. These databases combine curated manually-abstracted unstructured data with structured data. The starting point is a single database composed only of structured data elements available within an EHR (termed the Flatiron Health research database) refreshed on a monthly basis. This large general cross-tumor cohort includes all patients with at least one International Classification of Diseases (ICD)-9 or ICD-10 cancer code and at least one unique-date clinic encounter documented in the EHR (reflected by records of vital signs, treatment administration, and/or laboratory tests) on or after January 1, 2011, from both academic and community care sites combined. From there, patient data are sampled each month into each disease-specific database with randomized approaches implemented through software code, to ensure uniform application of the sampling approach and to avoid the potential for bias. Inclusion is based on cancer-specific cohort inclusion and exclusion criteria (i.e., relevant ICD-9 and ICD-10 codes and technology-assisted abstraction of unstructured information, such as metastatic status).

All manual abstraction of unstructured information, including confirmation of diagnosis and stage, is carried out by abstractors (i.e., clinical oncology nurses or tumor registrars). Clinically-relevant details specific to each cancer type are abstracted from any form of clinical documentation available in the EHR including clinic visit notes, radiology reports, pathology reports, etc. Abstractors are trained to identify and extract relevant information by following policies and procedures tested and optimized for reliability and reproducibility through iterative processes, and oversight is provided by medical oncologists. Each month, the datasets grow with new cases as well as incremental abstraction of newly-available clinical documentation from pre-existing patients. Therefore, at any given cutoff time, the size of a dataset is dependent on the initiation date for that disease-specific dataset and the overall population prevalence (which determines the growth rate). Typical data recency is 30 days. Practically speaking, a data cutoff of December 31, 2019 would include all information entered into the EHR through November 30, 2019, and subsequent cutoffs (i.e., January 31, 2020, February 29, 2020, etc) will render increasingly larger sample sizes. In addition, each database undergoes continuous audit procedures to monitor abstractor performance while proprietary technology links each curated data variable to its source documentation within the EHR, enabling subsequent review, when necessary. At the individual patient level, this approach provides a recent and robust longitudinal view into the clinical course, capturing new clinical information as it is documented within the EHR. Flatiron Health data are available for research via Institutional review board (IRB) approval of a master study protocol with waiver of informed consent (IRB # RWE-001, “The Flatiron Health Real-World Evidence Parent Protocol”, Tracking # FLI1-18-044 by the Copernicus Group IRB), obtained prior to study conduct, which covers the data from all sites represented.

As of May 31, 2019, there were 19 disease-specific databases available at Flatiron Health: advanced urothelial cancer, metastatic breast cancer, early breast cancer, chronic lymphocytic

leukemia (CLL), metastatic colorectal cancer, diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), advanced gastric/esophageal carcinoma, advanced hepatocellular carcinoma (HCC), advanced head and neck cancer, advanced melanoma, malignant pleural mesothelioma, multiple myeloma (MM), advanced non-small cell lung cancer (NSCLC), ovarian carcinoma, metastatic pancreatic carcinoma, metastatic prostate cancer, advanced renal-cell carcinoma (RCC), and small-cell lung cancer (SCLC) (Appendix I). At least two unique-date clinic encounters documented in the EHR in the Flatiron Health database (reflected by records of vital signs, treatment administration, and/or laboratory tests) on or after January 1, 2011, are required for patient data to be entered into a given dataset, with the exception of a few diseases with different start dates (Appendix I).

At the time of this analysis, the Flatiron Health EHR-derived database included de-identified data from over 280 cancer practices representing more than 2.2 million patients and about 800 distinct sites of care from all 50 states and Puerto Rico. The distribution of patients across community and academic practices largely reflects patterns of care in the US, where most patients are treated in community clinics, but can vary for each disease. Mortality information is captured via a composite variable that uses multiple data sources (structured and unstructured EHR content, commercial sources, Social Security Death Index) and is benchmarked against the National Death Index data as a gold standard (19).

SEER and NPCR databases

The SEER Program supports most aspects of cancer surveillance research, providing analytical tools, and methodological expertise in collecting, analyzing, interpreting and disseminating population-based statistics. SEER population-based data include cancer incidence and survival data by age, sex, race, year of diagnosis, and geographic areas (including SEER registry and

county). SEER releases new research data each spring based on the previous November's data submission. Data are available across various registries and versions of the SEER Program from 1975 through 2016. Mortality information in SEER is obtained from the CDC's National Center for Health Statistics' National Vital Statistics System (20) and includes mortality data along with cause-specific death classification information.

The NPCR cancer registries routinely capture data elements including the type, extent, and location of the cancer, the type of initial treatment, and outcomes of newly diagnosed cancers. Medical facilities such as hospitals, physician offices, and pathology laboratories send information about cancer cases to their respective central cancer registry, and each central cancer registry submits electronically de-identified demographic and clinical information to the NPCR on a yearly basis (4). Mortality information in the NPCR is obtained from the CDC's National Center for Health Statistics' National Vital Statistics System (20). As of May 31, 2019, the most recent information available from NPCR included new incident malignancies diagnosed through December 31, 2016. NPCR data is made available through the US Cancer Statistics dataset, which combines NPCR data and data from 4 SEER-funded states (Connecticut, Hawaii, Iowa, and New Mexico). This data provides information on 100% of the US population. In this paper, "NPCR data" was obtained through the US Cancer Statistics public use research dataset and was restricted to the 46 NPCR funded states and D.C.

Comparative analysis

Variables

For each cancer type, demographic and clinical characteristics including race, age, region, year and stage at diagnosis were compared between the Flatiron Health and the SEER and NPCR databases. To overcome coding discrepancies across databases, cancer types were matched using ICD-9, ICD-10, and histology codes (e.g. ICD-0-3). All comparisons were unadjusted.

Variables not available across two or more data sources were not included in the comparison (e.g., smoking status, treatment detail, real world progression [rwP] information).

In the Flatiron Health databases, cancer staging information was collected as entered into the EHR by the treating physician or otherwise as assessed by Flatiron Health abstractors; during the study time period, the applicable staging criteria for solid tumors were those of the American Joint Commission of Cancer (AJCC) 7th edition and 8th edition manuals (21, 22), Rai staging for CLL and the International Staging System (ISS) for MM. For SEER and NPCR, diagnosis and staging information was abstracted from various sources including medical records and pathology reports. For both programs, staging information followed the Collaborative Stage coding systems (23). See additional information on variable definitions in Appendix I.

Patient eligibility and time frames

Diagnostic codes and eligibility criteria used to select the patients eligible for each database are listed in Appendix I. For SEER and NPCR, only malignant cases diagnosed on or after January 1st, 2011 were included for all the analyses (benign, uncertain behavior, carcinoma in situ, secondary malignancy cancers were excluded). For the Flatiron Health databases, patients diagnosed between January 1, 2011 and May 31, 2019, were included (except those missing diagnosis year and/or birth year).

In order to compare the particular data segments common across all three databases (the most recent SEER and NPCR data releases reach through 2016 as initial diagnosis year), descriptive analyses were performed not only for all patients available for analysis across the entire time frame of January 2011 - May 2019 in the Flatiron Health databases, but also in the subset available for analysis from January 2011 - December 2016.

As sensitivity analyses to address potential biases related to temporal drifts, we performed separate comparisons for the patient subgroups who had stage IV disease at diagnosis in each

cancer type, for whom survival times would be expected to be shorter and the date of diagnosis would be expected to be closer to the database entry point. The potential biases to address were twofold: (i) as noted above, SEER and NPCR only collect specific incident disease data points, whereas Flatiron Health databases include both incident and prevalent cases. Therefore, Flatiron Health databases may receive patients at the time of diagnosis but also patients with initial diagnosis dates in the past; these cases may have long intervening periods between the initial diagnosis date and the date of entry into the Flatiron Health database, introducing a potential bias for patient characteristics associated with longer survival times (when compared with strictly incident cases in cancer registries); (ii) in addition, temporal trends where certain patient characteristics (i.e., sex, age) may be associated with cancer diagnoses during discrete time periods and can affect distributions depending on diagnosis year.

Analyses

Case-level data for patients in SEER were extracted from the SEER 18 November 2018 data submission dataset to the SEER Program by using the Case Listing Session feature in SEER*Stat software (Version 8.3.6, Information Management Services, Inc., Silver Spring, MD) and processed by using R 3.6.1. For patients in NPCR, case listing is not publicly available in the US Cancer Statistics public use SEER*Stat dataset, and case-level data cannot be accessed or downloaded. Frequencies by demographic and clinical characteristics for all malignant cases were calculated in SEER*Stat software using the November 2018 data submission.

The analysis of patients in the Flatiron Health databases was refreshed in April 2023 to incorporate an update to the birth year variable, reflecting best practices in patient de-identification (see Appendix II for more detail). This refresh resulted in updates to the distributions in calculated age at initial diagnosis in Tables 2 through 20 and A3.A through A3.S.

RESULTS

Among the 2.2 million patients with cancer in the Flatiron Health database as of May 2019, 201,570 were included in this analysis, as well as 1,719,277 and 6,308,342 cases from the SEER and NPCR, respectively. The disease-specific databases vary in size, depending on the incidence and prevalence of the disease, and in the case of Flatiron Health, on the length of time the database has been active. The largest comparisons corresponded to 55,554 vs 273,742 and 903,355 patients with NSCLC, and the smallest to 1,116 vs 18,148 and 72,575 for patients with FL.

Tables 2-20 present the comparisons for the following tumor types: advanced urothelial (or bladder), metastatic breast cancer, early breast cancer, CLL, metastatic colorectal cancer, DLBCL, FL, advanced gastric/esophageal carcinoma, HCC, advanced head and neck cancer, advanced melanoma, malignant pleural mesothelioma, MM, advanced NSCLC, ovarian carcinoma, metastatic pancreatic carcinoma, metastatic prostate cancer, metastatic RCC, and SCLC.

DISCUSSION

Results reported in this descriptive study provide an overview of the originating sources, data collection methods, and comparative population characteristics for three oncology-specific RWD sources in the US: the Flatiron Health, SEER, and NPCR data. We focused our comparisons on baseline demographic and clinical variables at the time of initial cancer diagnosis that describe the populations included in these data sources, based on the data elements commonly found in all three of them. Each of these three data sources relies on different collection approaches (Table 1), and our findings reveal population differences likely stemming from those distinct collection strategies. These differences in data collection methods and resulting populations should be considered when determining whether a dataset is fit-for-use for a particular research

question and can help to contextualize research results obtained when using each data source. To further assist in that contextualization, this discussion highlights some of the potential underlying explanations for the differences observed.

The distribution of patients according to sex/gender in the three data sources was comparable, but there were noticeable differences for other variables. Regarding regional distribution, Flatiron Health and the NPCR data were most closely aligned to the regional population distribution in the most recent US census (24). Due to its design, SEER data diverges the most from the census, particularly overrepresenting the West, while the Flatiron Health database provides a convenience sample that is slightly weighted towards the South and underrepresents the Western region.

The three data sources had overall similar age distributions, with a trend toward a modestly lower proportion of patients over 80 years at diagnosis in the Flatiron Health database (generally ≤5% lower across most diseases). These slight differences could be secondary to the algorithmic transformation of birth year for select elderly patients, which is required to reduce the risk of re-identification (see Appendix II). Another potential reason for modest discrepancies in the oldest age category is the difference in information sources that feed each one of them. State registries collect information regardless of patients' site of care and from death certificates and autopsy reports (25), while Flatiron Health databases accrue information only via oncology clinics. By focusing on specialized care, Flatiron Health has limited reach into general hospice or other geriatric care settings, where some elderly patients may be referred before they complete two visits to an oncology clinic (therefore excluding them from eligibility into Flatiron Health databases). This mechanism may account for the larger age discrepancies seen in some diseases such as pancreatic cancer, as this is an aggressive cancer in which elderly patients may be more likely to be referred to hospice early in the disease course. Of note, the larger age

discrepancy seen in CLL may be related to further differences in clinical inclusion criteria between the three databases, as patients are only included in the Flatiron Health database if they have a record of CLL therapy, and elderly patients have been observed to be less likely than younger patients to receive CLL treatment (26).

For information on race, the different data collection approaches result in expected differences in completeness and in population distribution across the three databases. The proportion of incomplete records for race in Flatiron Health data is greater than in the other databases. During routine oncology care in the US (i.e., in the source clinics for Flatiron Health), collection of race data is not mandatory or incentivized; on the other hand, the registries feeding both SEER and NPCR have a mandate to reach certain levels of completeness for this variable and thus, this information is collected both directly from self-reports and indirectly using algorithms (27). In addition, Flatiron Health relies on self-reported information by patients, which adds complexity and variability to an information category that is in constant evolution within a broader social context. For example, while “Hispanic or Latino” is standardly recorded as an ethnicity value, in some patients this may be coded as a race value instead; in such cases, these values were mapped to “Unknown”. These challenges probably contribute to the apparent lower completeness of this variable in the Flatiron Health databases. Furthermore, due to the purposeful design of the SEER program, there is an overrepresentation of certain groups compared to the US census (24, 28), a finding consistent with prior representativeness studies (29-31).

Lastly, a combination of differences in sources and in data collection approaches across the three data sources leads to noticeable differences in their information about AJCC disease stage at diagnosis. In some diseases, particularly in HCC, malignant pleural mesothelioma, RCC, SCLC, prostate cancer, and DLBCL, detailed AJCC disease stage information is missing from Flatiron Health databases to a substantially larger extent than in SEER, although that

incompleteness is mitigated in the simplified category metastatic/non-metastatic disease. To understand that finding, it is important to note that Flatiron Health data are generated from a pipeline of medical oncology EHR-derived data, where stage information is mostly as documented in unstructured notes by the treating oncology team. In contrast, registries rely on multiple sites of care as sources, and disease stage is intentionally entered into their databases via mandated calculation and coding by trained tumor registrars (32, 33). Ultimately, these fundamental differences lead to idiosyncratic fluctuations in information completeness in EHR-derived vs systematically-collected data. For instance, clinical scenarios where medical oncologists tend to be involved in initial diagnosis (when staging takes place) are more likely to have a complete capture of initial staging in the medical oncology EHR. To wit, compare cancers commonly diagnosed at an advanced stage (e.g., SCLC) or that are eligible for systemic adjuvant therapy from early stages (e.g., breast cancer) with diseases where the medical oncologist tends to be less involved in initial diagnosis (e.g., RCC, which is often managed surgically upon initial diagnosis).

Completeness/incompleteness of EHR-derived data is also affected by practical patterns of clinical documentation; as staging algorithms become increasingly complex, practitioners may tend to be more attentive to staging specifics in settings where the link between staging and treatment is more crucial (i.e., local or locally-advanced settings, where patients are candidates for multi-modality therapy), and less stringent in other settings (i.e., advanced disease) where staging information is less critical to clinical decision making. For example, AJCC staging is not a key consideration for initial HCC treatment decisions; clinical and laboratory data to assess underlying liver functional status and inform potential transplant eligibility are far more clinically relevant. Treating clinicians may prefer to document and rely on clinically actionable, non-AJCC staging systems for the routine management of some diseases, like HCC and SCLC, resulting in less AJCC-based information available in those databases.

In conclusion, the disease-specific Flatiron Health databases provide deep demographic, clinical, and treatment data models derived from EHR information. Several of the data elements in the Flatiron Health databases cannot be found in SEER or NPCR, such as date of metastatic diagnosis and sites of metastatic disease, comprehensive standard biomarker status, longitudinal treatment sequences, and disease progression dates. Within the portfolio of data elements commonly found across the three databases, comparing Flatiron Health to SEER and NPCR shows that Flatiron Health has a regional distribution closer to the general US census than SEER, less complete racial information, and disease-dependent variability in the capture of staging data. These differences stem from the originating EHR-source and from the rules for data capture and processing. Investigators should consider these inter-database demographic differences when designing studies and interpreting results obtained with Flatiron Health data, and when contextualizing their findings relative to SEER- or NPCR-based research.

Acknowledgements

Authors wish to thank Julia Saiz, PhD, Cody Patton, Hannah Gilham, and Jennifer Swanson from Flatiron Health, for editorial support, and Neil McQuarrie from Flatiron Health, for analytical support. We also wish to thank Mary Elizabeth O'Neil, MPH, from the CDC DCPC; Angela Mariotto, PhD, and Donna R. Rivera, PharmD, MSc, from the DCCPS at the NIH NCI; and Olivier Humblet, ScD, Emily Castellanos, MD, MPH, and Roxanne Diaz, BS, from Flatiron Health, for their review and valuable feedback.

REFERENCES

1. US Food and Drug Administration (b). Framework for FDA's real-world evidence program. December 2018. Accessed at https://www.fda.gov/media/120060/download on December 23, 2019

2. Eichler H□G, Bloechl□Daum B, Broich K et al. Data rich, information poor: Can we use electronic health records to create a learning healthcare system for pharmaceuticals? Clin Pharmacol Ther. 2019;105: 912-922.

3. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Overview of the SEER Program. Accessed at https://seer.cancer.gov/about/overview.html on February 18, 2020

4. Center for Disease Control and Prevention. National Program of Cancer Registries. Accessed at https://www.cdc.gov/cancer/npcr/index.htm on February 18, 2020

5. Health Information Technology (HITECH Act). Accessed at https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf on February 18 2020

6. American Society of Clinical Oncology. The state of cancer care in America, 2015: a report by the American Society of Clinical Oncology. J Oncol Practice. 2015; 11:79-113

7. The Office of the national Coordinator for Health Information Technology. 2016 Report to congress on health IT progress: examining the hitech era and the future of health it. Accessed at https://www.healthit.gov/sites/default/files/2016_report_to_congress_on_healthit_progress.pdf on February 18 2020

8. Parikh RB, Feld EK, Galsky MD et al. First-line immune checkpoint inhibitor use in cisplatin-eligible patients with advanced urothelial carcinoma: results from a real-world analysis. Futur Oncol. 2019;16:4341-4345.

9. Parikh RB, Adamson BJS, Khozin S et al. Association between FDA label restriction and immunotherapy and chemotherapy use in bladder cancer. JAMA. 2019;322:1209-1211.

10. Khozin S, Miksad RA, Adami J et al. Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non-small cell lung cancer. Cancer. 2019;125:4019-4032.

11. Feld EK, Harton J, Meropol NJ et al. Effectiveness of first-line immune checkpoint blockade versus carboplatin-based chemotherapy for metastatic urothelial cancer. Eur Urol. 2019;76:524-532.

12. Steuten LM, Goulart BHL, Meropol NJ, et al. Cost effectiveness of multigene panel sequencing for patients with advanced non-small-cell lung cancer. JCO Clin Cancer Informatics. 2019;3:1-0.

13. Bagley SJ, Talento S, Mitra N et al. Comparative effectiveness of carboplatin/pemetrexed with versus without bevacizumab for advanced nonsquamous non-small cell lung cancer. J Natl Compr Cancer Netw. 2019;17:469-477

14. Singal G, Miller PG, Agarwala V et al. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinicogenomic database. JAMA. 2019;321:1391-1399.

15. Parikh RB, Galsky MD, Gyawali B et al. Trends in checkpoint inhibitor therapy for advanced urothelial cell carcinoma at the end of life: Insights from real□world practice. Oncologist. 2019;24:397-399.

16. Riaz F, Presley CJ, Chiang AC et al. Disparities in broad-based genomic sequencing for patients with advanced non-small cell lung cancer. J Geriatr Oncol. 2019;10:669-672.

17. Winfree KB, Torres AZ, Zhu YE et al. Treatment patterns, duration, and outcomes of pemetrexed maintenance therapy in patients with advanced NSCLC in a real-world setting. Curr Med Res Opin. 2018;35:817-827.

18. Presley CJ, Tang D, Soulos PR et al. Association of broad-based genomic sequencing with survival among patients with advanced non-small cell lung cancer in the community oncology setting. JAMA. 2018;320:469-477.

19. Curtis MD, Griffith SD, Tucker MG et al. Development and validation of a high□quality composite real□world mortality endpoint. Health Serv Res. 2018;53:4460-4476.

20. Center for Disease Control and Prevention. National Vital Statistics System. Accessed at https://www.cdc.gov/nchs/nvss/index.htm on February 18, 2020

21. Edge SB, Byrd DR, Compton CC, et al, editors: AJCC cancer staging manual (7th ed). New York, NY: Springer; 2010.

22. Greene FL, Byrd DR, Brookland RK, et al, editors: AJCC cancer staging manual (8th ed). New York, NY: Springer; 2017.

23. American Joint Committee on Cancer. Collaborative stage data collection system. Accessed at http://www.cancerstaging.org/cstage/Pages/default.aspx on February 19, 2020

24. United States Census Bureau. Accessed at https://www.census.gov/quickfacts/fact/map/US/POP010210 on February 18, 2020

25. National Cancer Institute. SEER training modules. Accessed at https://training.seer.cancer.gov/abstracting/ on February 18, 2020

26. Mato A, Jahnke J, Li P, et al. Real-world treatment and outcomes among older adults with chronic lymphocytic leukemia before the novel agents era. *Haematologica*. 2018;103(10):e462-e465. doi:10.3324/haematol.2017.185868

27. National Cancer Institute. Surveillance Epidemiology and End Results Program. Race recode changes. Accessed at https://seer.cancer.gov/seerstat/variables/seer/race_ethnicity/ on February 18, 2020.

28. National Cancer Institute. Surveillance Epidemiology and End Results Program. Population Characteristics. Accessed at https://seer.cancer.gov/registries/data.html on March 11, 2020

29. Warren JL, Klabunde CN, Schrag D et al. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care. 2002;40(8 Suppl):IV–18.

30. Nattinger AB, McAuliffe TL, Schapira MM. Generalizability of the surveillance, epidemiology, and end results registry population: factors relevant to epidemiologic and health care research. J Clin Epidemiol. 1997;50:939–945.

31. Kuo TM, Mobley LR. How generalizable are the SEER registries to the cancer populations of the USA? Cancer Causes Control. 2016;27: 1117-26.

32. National Cancer Institute. Surveillance Epidemiology and End Results Program. Cancer Stage Variable Documentation. Accessed at https://seer.cancer.gov/analysis/stage.html on February 18, 2020

33. National Cancer Institute. Surveillance Epidemiology and End Results Program. Registry Operations. Accessed at https://seer.cancer.gov/registrars/ on February 19, 2020.

34. NPCR and SEER Incidence – U.S. Cancer Statistics 2001-2016 Public Use Database – Data Standards and Data Dictionary. Accessed at https://www.cdc.gov/cancer/uscs/public-use on February 19, 2020

35. Printz C. Changes underway for SEER: Program leaders work to increase the breadth and depth of information. Cancer. 2015;121: 3183-3184.

36. Griffith SD, Tucker M, Bowser B et al. Generating real-world tumor burden endpoints from electronic health record data: comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. Adv Ther. 2019; 36, 2122–2136.

Table 1. General features of SEER, NPCR and Flatiron Health databases.

Feature	SEER	NPCR	Flatiron Health
Source	State disease registries (3) receive information about cancer diagnoses and treatment from public health reporting entities (including healthcare organizations, facilities and providers). Geographic distribution of participating states is based on maintaining a high-quality population-based cancer reporting system, and pre-specified and adjusted to reach specific regional and racial representation. Mortality information is obtained from the CDC's National Center for Health Statistics' National Vital Statistics System (20), which collects mortality data along with cause-specific death classification information.	46 state and D.C. disease registries, receive information about cancer diagnoses from healthcare providers from any specialty.| All-cause mortality is assessed from the CDC's National Center for Health Statistics' National Vital Statistics System (20).	EHRs from community and academic oncology clinics participating in the Flatiron network. Databases receive routinely collected information (i.e., without intentional or pre-specified collection) from non-oncology care settings or sites outside of participating clinics via scanned in documentation. Mortality is assessed via a composite variable that integrates multiple data sources (EHR information, a commercial death data source, and the Social Security Death Index) and is benchmarked against the National Death Index (NDI).
Collection approach	Data collection is based in SEER and NAACCR data standards, and is intentional in order to meet pre-	Data collection is intentional in order to meet pre-specified registry program standards for	Originating information is collected as routinely documented or entered by participating practitioners.

	specified registry program standards for completeness and quality.	completeness.	
Curation	Unstructured and structured data are abstracted by tumor registrars, who enter text manually into a data collection template. In order to facilitate pooling of data, all SEER and NPCR cancer registries use uniform data items and codes as documented by the North American Association of Central Cancer Registries (3,34).	Unstructured and structured data are abstracted by tumor registrars, who enter text manually into a data collection template. Some pathology reports are received electronically. In order to facilitate pooling of data, all NPCR and SEER cancer registries use uniform data items and codes as documented by the North American Association of Central Cancer Registries (3,34).	Structured data are harmonized and mapped to common units and terminology and combined with unstructured data processed by technology-enabled manual abstraction. Abstractors (clinical oncology nurses and tumor registrars) are trained to identify and extract relevant information by following optimized policies and procedures, and oversight is provided by medical oncologists.
Recency	Releases data with a 2-year delay	Releases data with a 2-year delay	Data are available with a 1-month delay
Key data model elements			
Demographics [a]	Age, sex, race, ethnicity, insurance type, geographic location	Age, sex, race, ethnicity, geographic location	Age, gender,[b] race, insurance type, geographic location
Tumor details	Date of initial diagnosis, primary tumor site, limited biomarker information (34,35),[c] morphology/histology, behavior, laterality and stage at diagnosis	Date of initial diagnosis, primary tumor site, limited biomarker information,[c] a morphology/histology and stage at diagnosis	Date of initial diagnosis, primary tumor site, morphology/histology and stage at diagnosis, standard biomarker information,[c] date of advanced/metastatic diagnosis, sites of metastatic disease[d]

Treatment	First course of treatment at initial diagnosis: • Diagnostic/procedure • Surgery • Radiation • Systemic therapy (chemotherapy, hormone therapy, biological response modifier [immunotherapy], or other)	First course of treatment at initial diagnosis: • Surgery	Sequence of treatments received after advanced/metastatic disease:[e] • Surgery, limited to the documentation in the oncology EHR, in certain diseases. • Radiation, limited to the documentation in the oncology EHR, in certain diseases. • Systemic therapy: type, dosing, dates of individual drug episodes (e.g. ordered dates, administered dates) including abstracted oral therapies
Variables	Cause specific mortality	Mortality	Mortality (19) Real world progression (rwP) (36).

[a]In the May 2023 update of this paper, "ethnicity" was removed as a data model element from Flatiron Health for clarity, as race values that were documented as ethnicity ("Hispanic or Latino") were recoded as "Unknown," and ethnicity was not separately reported in the results. This change is also reflected in Tables 2-20 and Table A2.
[b]Gender values likely reflect a patient's sex assigned at birth. However, it is unknown whether these values represent sex assigned at birth or gender identity or whether these values are self- or clinician-reported. These values vary by practices at individual clinical sites and the limitations of the EHR software used.
[c]The biomarker portfolio offered by each database is different, with SEER providing information about a limited set of specific biomarkers that might predict outcomes or response to specific therapies, the US Cancer Statistics database providing ER, PR and HER2 status, and Flatiron Health offering a more extensive portfolio of biomarkers for which testing is considered standard of care.
[d]Flatiron Health databases provide information on specific sites of metastatic disease for a selected set of tumor types.
[e]Flatiron Health has assembled early-stage disease datasets (for which information about early-stage treatment is collected) for certain selected tumor types.

Tables 2-20: Shaded areas are groups with no reported results. NA = not available from the source data, or numbers under the reporting suppression value to preserve patient confidentiality

Table 2. Characteristics for patients with bladder cancer

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	119010	453393	5197	7779	10558	24531	1841	2729
Age at initial diagnosis	0-19	60 (0.05)	237 (0.05)	0 (0.00)	0 (0.00)	1 (0.01)	23 (0.09)	0 (0.00)	0 (0.00)
	20-34	536 (0.45)	1827 (0.40)	4 (0.08)	9 (0.12)	38 (0.36)	94 (0.38)	1 (0.05)	4 (0.15)
	35-49	3733 (3.14)	14199 (3.13)	130 (2.50)	203 (2.61)	403 (3.82)	889 (3.62)	53 (2.88)	65 (2.38)
	50-64	26381 (22.17)	101927 (22.48)	1215 (23.38)	1810 (23.27)	2714 (25.71)	6107 (24.90)	442 (24.01)	616 (22.57)
	65-79	53815 (45.22)	210014 (46.32)	2750 (52.92)	4088 (52.55)	4751 (45.00)	10967 (44.71)	944 (51.28)	1408 (51.59)
	80+	34485 (28.98)	125189 (27.61)	1098 (21.13)	1584 (20.36)	2651 (25.11)	6451 (26.30)	401 (21.78)	636 (23.31)
	Unknown	0 (0)	0 (0)	0 (0.00)	85 (1.09)				
Sex	Female	29829 (25.06)	113919 (25.13)	1360 (26.17)	2025 (26.03)	3358 (31.81)	8037 (32.76)	551 (29.93)	808 (29.61)
	Male	89181 (74.94)	339474 (74.87)	3836 (73.81)	5753 (73.96)	7200 (68.19)	16494 (67.24)	1289 (70.02)	1920 (70.36)
	Unknown	0 (0)	0 (0)	1 (0.02)	1 (0.01)	0 (0)	0 (0)	1 (0.05)	1 (0.04)
Region	Midwest	12351 (10.38)	105212 (23.21)	707 (13.60)	1024 (13.16)	1119 (10.60)	5616 (22.89)	233 (12.66)	348 (12.75)
	Northeast	23368 (19.64)	97581 (21.52)	1191 (22.92)	1769 (22.74)	1715 (16.24)	5175 (21.10)	425 (23.09)	610 (22.35)
	South	25259 (21.22)	161064 (35.52)	2288 (44.03)	3449 (44.34)	2340 (22.16)	8689 (35.42)	824 (44.76)	1237 (45.33)
	West	58032 (48.76)	89536 (19.75)	895 (17.22)	1360 (17.48)	5384 (50.99)	5051 (20.59)	308 (16.73)	463 (16.97)
	Other	NA	NA	39 (0.75)	64 (0.82)	NA	NA	14 (0.76)	26 (0.95)
	Unknown	0 (0)	0 (0)	77 (1.48)	113 (1.45)	0 (0)	0 (0)	37 (2.01)	45 (1.65)
Race	Asian	5381 (4.52)	7998 (1.76)	72 (1.39)	99 (1.27)	549 (5.20)	513 (2.09)	29 (1.58)	47 (1.72)
	Black/Afr. American	7143 (6.00)	26290 (5.80)	214 (4.12)	301 (3.87)	961 (9.10)	2317 (9.45)	77 (4.18)	111 (4.07)
	White	104317 (87.65)	411164 (90.69)	4004 (77.04)	5838 (75.05)	8966 (84.92)	21499 (87.64)	1408 (76.48)	2011 (73.69)
	Other	444 (0.37)	3169 (0.70)	453 (8.72)	744 (9.56)	62 (0.59)	177 (0.72)	154 (8.37)	272 (9.97)

	Unknown	1725 (1.45)	4772 (1.05)	454 (8.74)	797 (10.25)	20 (0.19)	25 (0.10)	173 (9.40)	288 (10.55)
AJCC stage at diagnosis	0	54641 (45.91)	NA	23 (0.44)	39 (0.5)				
	I	27083 (22.76)	NA	82 (1.58)	120 (1.54)				
	II	13905 (11.68)	NA	312 (6.00)	467 (6.00)				
	III	6172 (5.19)	NA	334 (6.43)	493 (6.34)				
	IV	10558 (8.87)	NA	1841 (35.42)	2729 (35.08)				
	Unknown	6651 (5.59)	NA	2605 (50.13)	3931 (50.53)				
Metastatic at diagnosis	Yes	6378 (5.36)	24531 (5.41)	1436 (27.63)	2134 (27.43)				
	No	104405 (87.73)	410490 (90.54)	1937 (37.27)	2877 (36.98)				
	Unknown	8227 (6.91)	18372 (4.05)	1824 (35.10)	2768 (35.58)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	970 (12.47)	NA	NA	0 (0)	0 (0)
	2011 onward	119010 (100)	453393 (100)	5197 (100)	6724 (86.44)	10558 (100.00)	24531 (100.00)	1841 (100.00)	2729 (100.00)
	Unknown	0 (0)	0 (0)	0 (0)	85 (1.09)	0 (0)	0 (0)	0 (0)	0 (0)

Table 3. Characteristics for patients with metastatic breast cancer in the Flatiron Health database, compared with patients with breast cancer (any stage) in SEER and NPCR[a]

		N	All eligible, n (%)				Stage IV at diagnosis, n (%)			
			SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
					2011-16	2011-19			2011-16	2011-19
		388064	1379999	10219	19890	22092	81977	4534	6236	
Age at initial diagnosis	0-19		34 (0.01)	114 (0.01)	0 (0.00)	1 (0.01)	1(0.00)	0 (0)	0 (0.00)	0 (0.00)
	20-34		7254 (1.87)	24838 (1.80)	351 (3.43)	655 (3.29)	611 (2.77)	2092 (2.55)	130 (2.87)	177 (2.84)
	35-49		67561 (17.41)	230112 (16.67)	1945 (19.03)	4357 (21.90)	3368 (15.25)	12241 (14.93)	657 (14.49)	887 (14.22)
	50-64		143787 (37.05)	504791 (36.58)	3812 (37.30)	7804 (39.23)	8235 (37.28)	30516 (37.23)	1626 (35.86)	2195 (35.20)
	65-79		125955 (32.46)	465909 (33.76)	3110 (30.43)	5563 (27.97)	6794 (30.75)	25980 (31.70)	1579 (34.83)	2188 (35.09)
	80+		43473 (11.20)	154235 (11.18)	1001 (9.80)	1405 (7.06)	3083 (13.96)	11137 (13.59)	542 (11.95)	789 (12.65)
	Unknown		0 (0)	0 (0)	0 (0.00)	106 (0.53)				
Sex	Female		384979 (99.21)	1367658 (99.11)	10076 (98.60)	19641 (98.75)	21826 (98.8)	80872 (98.65)	4486 (98.94)	6158 (98.75)
	Male		3085 (0.79)	12341 (0.89)	143 (1.40)	249 (1.25)	266 (1.20)	1105 (1.35)	48 (1.06)	78 (1.25)
Region	Midwest		33976 (8.76)	301029 (21.81)	1435 (14.04)	2856 (14.36)	2065 (9.35)	18044 (22.01)	598 (13.19)	816 (13.09)
	Northeast		64594 (16.65)	265553 (19.24)	2307 (22.58)	4653 (23.39)	3941 (17.84)	15835 (19.32)	1068 (23.56)	1471 (23.59)
	South		84024 (21.65)	519910 (37.67)	4010 (39.24)	7583 (38.12)	5327 (24.11)	32341 (39.45)	1746 (38.51)	24441 (39.14)
	West		205470 (52.95)	293507 (21.27)	2024 (19.81)	4042 (20.32)	10759 (48.70)	15757 (19.22)	933 (20.58)	1273 (20.41)
	Other		NA	NA	269 (2.63)	465 (2.34)	NA	NA	101 (2.23)	139 (2.23)
	Unknown		0 (0)	0 (0)	174 (1.7)	291 (1.46)	0 (0)	0 (0)	88 (1.94)	96 (1.54)
Race	Asian		33303 (8.58)	51133 (3.71)	245 (2.4)	433 (2.18)	1563 (7.07)	2549 (3.11)	95 (2.10)	145 (2.33)
	Black/Afr. American		43750 (11.27)	164927 (11.95)	1279 (12.52)	2119 (10.65)	3633 (16.44)	14136 (17.24)	524 (11.56)	723 (11.59)
	White		305274 (78.67)	1142844 (82.81)	6629 (64.87)	13077 (65.75)	16660 (75.41)	64223 (78.34)	2976 (65.64)	3956 (63.44)
	Other		2309 (0.6)	14059 (1.02)	1166 (11.41)	2244 (11.28)	131 (0.59)	855 (1.04)	502 (11.07)	709 (11.37)
	Unknown		3428 (0.88)	7036 (0.51)	900 (8.81)	2017 (10.14)	105 (0.48)	214 (0.26)	437 (9.64)	703 (11.27)
AJCC stage at diagnosis	0		550 (0.14)	NA	1 (0.01)	8 (0.04)				
	I		182660 (47.07)	NA	748 (7.32)	2088 (10.5)				
	II		127547 (32.87)	NA	2098 (20.53)	4988 (25.08)				

	III	39857 (10.27)	NA	2257 (22.09)	4146 (20.84)				
	IV	22092 (5.69)	NA	4534 (44.37)	6237 (31.36)				
	Unknown	15358 (3.96)	NA	581 (5.69)	2423 (12.18)				
Metastatic at diagnosis	Yes	22806 (5.88)	81977 (5.94)	4534 (44.37)	6237 (31.36)				
	No	357347 (92.08)	1266005 (91.74)	5104 (49.95)	11230 (56.46)				
	Unknown	7911 (2.04)	32017 (2.32)	581 (5.69)	2423 (12.18)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	7484 (37.63)	NA	NA	NA	NA
	2011 onward	388064 (100.00)	1379999 (100.00)	10219 (100.00)	12301 (61.84)	22092 (100.00)	81977 (100.00)	4534 (100.00)	6236 (100.00)
	Unknown	NA	NA	0 (0)	105 (0.53)	NA	NA	NA	NA

[a] All patients with breast cancer included for the SEER and NPCR datasets; in this Flatiron Health dataset, technology enabled selection is applied to include patients with metastatic disease.

Table 4. Characteristics for patients with early breast cancer in the Flatiron Health database, compared with patients with breast cancer (any stage) in SEER and NPCR[a]

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
	N	SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
		388064	1379999	2253	3030				
Age at initial diagnosis	0-19	34 (0.01)	114 (0.01)	1 (0.04)	1 (0.03)				
	20-34	7254 (1.87)	24838 (1.80)	29 (1.29)	39 (1.29)				
	35-49	67561 (17.41)	230112 (16.67)	372 (16.51)	506 (16.70)				
	50-64	143787 (37.05)	504791 (36.58)	846 (37.55)	1099 (36.27)				
	65-79	125955 (32.46)	465909 (33.76)	833 (36.97)	1138 (37.56)				
	80+	43473 (11.20)	154235 (11.18)	172 (7.63)	247 (8.15)				
Sex	Female	384979 (99.21)	1367658 (99.11)	2226 (98.80)	2998 (98.94)				
	Male	3085 (0.79)	12341 (0.89)	26 (1.15)	31 (1.02)				
	Unknown	0 (0)	0 (0)	1 (0.04)	1 (0.03)				
Region	Midwest	33976 (8.76)	301029 (21.81)	348 (15.45)	15.31 (464)				
	Northeast	64594 (16.65)	265553 (19.24)	567 (25.17)	25.18 (763)				
	South	84024 (21.65)	519910 (37.67)	811 (36.00)	36.73 (1113)				
	West	205470 (52.95)	293507 (21.27)	451 (20.02)	19.67 (596)				
	Other	NA	NA	51 (2.26)	2.24 (68)				
	Unknown	0 (0)	0 (0)	25 (1.11)	0.86 (26)				

Race	Asian	33303 (8.58)	51133 (3.71)	64 (2.84)	89 (2.94)				
	Black/Afr. American	43750 (11.27)	164927 (11.95)	211 (9.37)	273 (9.01)				
	White	305274 (78.67)	1142844 (82.81)	1557 (69.11)	2049 (67.62)				
	Other	2309 (0.60)	1452059 (1.02)	237 (10.52)	342 (11.29)				
	Unknown	3428 (0.88)	7036 (0.51)	184 (8.17)	277 (9.14)				
AJCC stage at diagnosis	0	550 (0.14)	NA	NA	NA				
	I	182660 (47.07)	NA	1083 (48.07)	1516 (50.03)				
	II	127547 (32.87)	NA	713 (31.65)	926 (30.56)				
	III	39857 (10.27)	NA	255 (11.32)	319 (10.53)				
	IV	22092 (5.69)	NA	NA	NA				
	Unknown	15358 (3.96)	NA	202 (8.97)	269 (8.88)				
Metastatic at diagnosis	No	35734 (92.08)	1266005 (91.74)	2051 (91.03)	2761 (91.12)				
	Unknown	7911 (2.04)	32017 (2.32)	202 (8.97)	269 (8.88)				
	Yes	22806 (5.88)	81977 (5.94)	NA	NA				
Year of initial diagnosis	Pre-2011								
	2011 onward	388064 (100.00)	1379999 (100.00)	2253 (100.00)	3030 (100.00)				

[a] All patients with breast cancer included for the SEER and NPCR datasets; in this Flatiron Health dataset, technology enabled selection is applied to include patients with early-stage disease only.

Table 5. Characteristics for patients with chronic lymphocytic leukemia

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	18295	66384	4308	10722				
Age at initial diagnosis	0-19	9 (0.05)	33 (0.05)	0 (0.00)	0 (0.00)				
	20-34	47 (0.26)	186 (0.28)	9 (0.21)	31 (0.29)				
	35-49	769 (4.2)	2797 (4.21)	220 (5.11)	765 (7.13)				
	50-64	5247 (28.68)	18899 (28.47)	1343 (31.17)	3691 (34.42)				
	65-79	7944 (43.42)	29456 (44.37)	2112 (49.03)	4640 (43.28)				
	80+	4279 (23.39)	15013 (22.62)	624 (14.48)	925 (8.63)				
	Unknown	0 (0)	0 (0)	0 (0.00)	670 (6.25)				
Sex	Female	7100 (38.81)	25684 (38.69)	1504 (34.91)	4040 (37.68)				
	Male	11195 (61.19)	40700 (61.31)	2804 (65.09)	6682 (62.32)				
Region	Midwest	2035 (11.12)	13911 (20.96)	682 (15.83)	1671 (15.58)				
	Northeast	3385 (18.50)	14188 (21.37)	1040 (24.14)	2715 (25.32)				
	South	43879 (23.99)	26115 (39.34)	1673 (38.83)	4002 (37.33)				
	West	8466 (46.38)	12170 (18.33)	846 (19.64)	2164 (20.18)				
	Other	NA	NA	39 (0.91)	74 (0.69)				

	Unknown	0 (0)	0 (0)	28 (0.65)	96 (0.9)				
Race	Asian	394 (2.15)	668 (1.01)	39 (0.91)	73 (0.68)				
	Black/Afr. American	1362 (7.44)	4718 (7.11)	341 (7.92)	726 (6.77)				
	White	15690 (85.76)	58473 (88.08)	3247 (75.37)	8127 (75.80)				
	Other	60 (0.33)	484 (0.73)	383 (8.89)	960 (8.95)				
	Unknown	789 (4.31)	2041 (3.07)	298 (6.92)	836 (7.80)				
Rai stage at diagnosis	0	NA	NA	832 (19.31)	2136 (19.92)				
	I	NA	NA	574 (13.31)	1234 (11.51)				
	II	NA	NA	291 (6.75)	572 (5.33)				
	III	NA	NA	244 (5.66)	471 (4.39)				
	IV	NA	NA	451 (10.47)	897 (8.37)				
	Unknown	18295 (100.00)	NA	1916 (44.48.)	5412 (50.48)				
Metastatic at diagnosis	Yes	NA	66322 (99.91)	4308 (100.00)	10722 (100.00)				
	No	NA	NA	NA	NA				
	Unknown	18295 (100)	60 (0.09)	NA	NA				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	4821 (44.96)				
	2011 onward	18295 (100.00)	66384 (100.00)	4308 (100.00)	5231 (48.79)				
	Unknown	NA	NA	0 (0)	670 (6.25)				

Table 6. Characteristics for patients with colorectal cancer

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	159140	569605	15137	21914	30852	120049	8153	12613
Age at initial diagnosis	0-19	420 (0.26)	1280 (0.22)	7 (0.05)	10 (0.05)	2 (0.08)	74 (0.06)	5 (0.06)	6 (0.05)
	20-34	2803 (1.76)	9280 (1.63)	250 (1.65)	363 (1.66)	572 (1.85)	2054 (1.71)	151 (1.85)	226 (1.79)
	35-49	15523 (9.75)	54174 (9.51)	1934 (12.78)	2878 (13.13)	3830 (12.41)	14274 (11.89)	1096 (13.44)	1752 (13.89)
	50-64	53494 (33.61)	187319 (32.89)	5414 (35.77)	7831 (35.74)	10745 (34.83)	41885 (34.89)	2989 (36.66)	4543 (36.02)
	65-79	55382 (34.80)	205947 (36.16)	5718 (37.77)	8260 (37.69)	10239 (33.19)	41225 (34.34)	2931 (35.95)	4548 (36.06)
	80+	31518 (19.81)	111605 (19.59)	1814 (11.98)	2547 (11.62)	5441 (17.64)	20537 (17.11)	981 (12.03)	1538 (12.19)
	Unknown	0 (0)	0 (0)	0 (0.00)	25 (0.11)				
Sex	Female	76739 (48.22)	274648 (48.22)	6852 (45.27)	9845 (44.93)	14294 (46.33)	56414 (46.99)	3759 (46.11)	5739 (45.50)
	Male	82401 (51.78)	294957 (51.78)	8281 (54.71)	12065 (55.06)	16558 (53.67)	63635 (53.01)	4392 (53.87)	6872 (54.48)
	Unknown	0 (0)	0 (0)	4 (0.03)	4 (0.02)	0 (0)	0 (0)	2 (0.02)	2 (0.02)
Region	Midwest	15200 (9.55)	127281 (22.35)	2077 (13.72)	2987 (13.63)	2948 (9.56)	26305 (21.91)	1117 (13.70)	1717 (13.61)
	Northeast	25022 (15.72)	104750 (18.39)	3569 (23.58)	5034 (22.97)	4991 (16.18)	22445 (18.70)	1936 (23.75)	2879 (22.83)
	South	39150 (24.6)	224791 (39.46)	5783 (38.20)	8561 (39.07)	7792 (25.26)	48202 (40.15)	3104 (38.07)	4968 (39.39)
	West	79768 (50.12)	112783 (19.80)	3046 (20.12)	4425 (20.19)	15121 (49.01)	23097 (19.24)	1637 (20.08)	2548 (20.20)
	Other	NA	NA	424 (2.80)	625 (2.85)	NA	NA	194 (2.38)	316 (2.51)
	Unknown	0 (0)	0 (0)	238 (1.57)	282 (1.29)	0 (0)	0 (0)	165 (2.02)	185 (1.47)
Race	Asian	13150 (8.26)	19734 (3.46)	421 (2.78)	609 (2.78)	2516 (8.16)	4094 (3.41)	210 (2.58)	335 (2.66)
	Black/Afr. American	19386 (12.18)	71643 (12.58)	1580 (10.44)	2244 (10.24)	4458 (14.45)	17836 (14.86)	880 (10.79)	1339 (10.62)
	White	123419 (77.55)	467146 (82.01)	10076 (66.57)	14275 (65.14)	23502 (76.18)	96537 (80.41)	5400 (66.23)	8104 (64.25)
	Other	1320 (0.83)	6585 (1.16)	1775 (11.73)	2617 (11.94)	271 (0.88)	1352 (1.13)	924 (11.33)	1465 (11.62)
	Unknown	1865 (1.17)	4497 (0.79)	1285 (8.49)	2169 (9.90)	105 (0.34)	230 (0.19)	739 (9.06)	1370 (10.86)
AJCC stage at	0	4084 (2.57)	NA	0 (0)	2 (0.01)				
	I	35617 (22.38)	NA	399 (2.64)	584 (2.66)				
	II	35164 (22.10)	NA	1827 (12.07)	2410 (11.00)				

diagnosis	III	38868 (24.42)	NA	4163 (27.50)	5321 (24.28)				
	IV	30852 (19.39)	NA	8153 (53.86)	12613 (57.56)				
	Unknown	14555 (9.15)	NA	595 (3.93)	984 (4.49)				
Metastatic at diagnosis	Yes	32861 (20.65)	120049 (21.08)	8153 (53.86)	12613 (57.56)				
	No	116547 (73.24)	413843 (72.65)	6389 (42.21)	8317 (37.95)				
	Unknown	9732 (6.12)	35713 (6.27)	595 (3.93)	984 (4.49)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	1362 (6.22)				
	2011 onward	159140 (100.00)	569605 (100.00)	15137 (100.00)	20527 (93.67)	30852 (100.00)	120049 (100.00)	8153 (100.00)	12613 (100.00)
	Unknown	0 (0)	0 (0)	0 (0)	25 (0.11)				

Table 7. Characteristics for patients with diffuse large B-cell lymphoma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	38953	139511	3989	5659	13071	71788	1122	1581
Age at initial diagnosis	0-19	360 (0.92)	1237 (0.89)	12 (0.30)	14 (0.25)	88 (0.67)	550 (0.77)	1 (0.09)	1 (0.06)
	20-34	1422 (3.65)	4925 (3.53)	158 (3.96)	219 (3.87)	463 (3.54)	2066 (2.88)	36 (3.21)	57 (3.61)
	35-49	3606 (9.26)	12426 (8.91)	341 (8.55)	460 (8.13)	1182 (9.04)	6042 (8.42)	98 (8.73)	133 (8.41)
	50-64	10709 (27.49)	37313 (26.75)	1132 (28.38)	1538 (27.18)	3770 (28.84)	20062 (27.95)	322 (28.70)	448 (28.34)
	65-79	14733 (37.82)	54467 (39.04)	1669 (41.84)	2392 (42.27)	5009 (38.32)	28857 (40.20)	495 (44.12)	689 (43.58)
	80+	8123 (20.85)	29143 (20.89)	677 (16.97)	1036 (18.31)	2559 (19.58)	14211 (19.80)	170 (15.15)	253 (16.00)
Sex	Male	17319 (44.46)	62523 (44.82)	1828 (45.83)	2562 (45.27)	5682 (43.47)	31679 (44.13)	492 (43.85)	686 (43.39)
	Female	21634 (55.54)	76988 (55.18)	2161 (54.17)	3097 (54.73)	7389 (56.53)	40109 (55.87)	630 (56.15)	895 (56.61)
Region	Midwest	3922 (10.07)	31915 (22.88)	520 (13.04)	716 (12.65)	1407 (10.76)	17472 (24.34)	151 (13.46)	202 (12.78)
	Northeast	5952 (15.28)	27014 (19.36)	1144 (28.68)	1637 (28.93)	1948 (14.90)	14206 (19.79)	306 (27.27)	436 (27.58)
	South	7945 (20.40)	50512 (36.21)	1484 (37.20)	2143 (37.87)	2669 (20.42)	24864 (34.64)	439 (39.13)	619 (39.15)
	West	21134 (54.26)	30070 (21.55)	743 (18.63)	1031 (18.22)	7047 (53.91)	15246 (21.24)	190 (16.93)	280 (17.71)
	Other	NA	NA	45 (1.13)	70 (1.24)	NA	NA	22 (1.96)	28 (1.77)
	Unknown	0 (0)	0 (0)	53 (1.33)	62 (1.10)	0 (0)	0 (0)	14 (1.25)	16 (1.01)
Race	Asian	3396 (8.72)	5253 (3.77)	96 (2.41)	133 (2.35)	1060 (8.11)	2555 (3.56)	26 (2.32)	34 (2.15)
	Black/Afr. American	2775 (7.12)	10637 (7.62)	229 (5.74)	332 (5.87)	1050 (8.03)	5985 (8.34)	78 (6.95)	112 (7.08)
	White	32233 (82.75)	121430 (87.04)	2894 (72.55)	3986 (70.44)	10822 (82.79)	62326 (86.82)	817 (72.82)	1119 (70.78)
	Other	226 (0.58)	1408 (1.01)	392 (9.83)	589 (10.41)	77 (0.59)	710 (0.99)	101 (9.00)	160 (10.12)
	Unknown	323 (0.83)	783 (0.56)	378 (9.48)	619 (10.94)	62 (0.47)	212 (0.30)	100 (8.91)	156 (9.87)
AJCC stage at diagnosis	I	9582 (24.60)	NA	523 (13.11)	677 (11.96)				
	II	7190 (18.46)	NA	652 (16.34)	883 (15.60)				
	III	6810 (17.48)	NA	721 (18.07)	1036 (18.31)				
	IV	13071 (33.56)	NA	1122 (28.13)	1581 (27.94)				
Metastatic at	Yes	20024 (51.41)	71788 (51.46)	1843 (46.2)	2617 (46.24)				

diagnosis	No	16983 (43.60)	57864 (41.48)	1175 (29.46)	1560 (27.57)				
	Unknown	1946 (5.00)	9859 (7.07)	971 (24.34)	1482 (26.19)				
Year of initial diagnosis	Pre-2011								
	2011 onward	38953 (100.00)	139511 (100.00)	3989 (100.00)	5659 (100.00)	13071 (100.00)	71788 (100.00)	1122 (100.00)	1581 (100.00)

Table 8. Characteristics for patients with follicular lymphoma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	18199	72575	804	1116	4838	35933	315	440
Age at initial diagnosis	0-19	53 (0.29)	201 (0.28)	2 (0.25)	2 (0.18)	1 (0.02)	18 (0.05)	0 (0.00)	0 (0.00)
	20-34	306 (1.69)	1167 (1.61)	12 (1.49)	14 (1.25)	82 (1.69)	574 (1.60)	5 (1.59)	6 (1.36)
	35-49	2193 (12.05)	8056 (11.10)	82 (10.20)	112 (10.04)	680 (14.04)	4513 (12.56)	37 (11.75)	51 (11.59)
	50-64	6442 (35.47)	24867 (34.26)	286 (35.57)	374 (33.51)	1806 (37.33)	13115 (36.50)	107 (33.97)	144 (32.73)
	65-79	6784 (37.25)	28224 (38.89)	329 (40.92)	467 (41.85)	1783 (36.85)	13500 (37.57)	139 (44.13)	196 (44.55)
	80+	2408 (13.22)	10060 (13.86)	93 (11.57)	147 (13.17)	486 (10.05)	4213 (11.72)	27 (8.57)	43 (9.77)
Sex	Female	8891 (48.85)	36130 (49.78)	407 (50.62)	557 (49.91)	2401 (49.63)	17958 (49.98)	159 (50.48)	219 (49.77)
	Male	9308 (51.15)	36445 (50.22)	397 (49.38)	559 (50.09)	2437 (50.38)	17975 (50.02)	156 (49.52)	221 (50.23)
Region	Midwest	1767 (9.71)	16620 (22.90)	119 (14.80)	163 (14.61)	507 (10.48)	9189 (25.57)	46 (14.60)	62 (14.09)
	Northeast	2993 (16.45)	13977 (19.26)	220 (27.36)	295 (26.43)	707 (14.61)	7064 (19.66)	78 (24.76)	101 (22.95)
	South	3833 (21.06)	27591 (38.02)	285 (35.45)	408 (36.56)	1048 (21.66)	12340 (34.34)	110 (34.92)	159 (36.14)
	West	9606 (52.78)	14387 (19.82)	160 (19.9)	221 (19.80)	2576 (53.25)	7340 (20.43)	74 (23.49)	106 (24.09)
	Other	NA	NA	10 (1.24)	16 (1.43)	NA	NA	4 (1.27)	8 (1.82)
	Unknown	0 (0)	0 (0)	10 (1.24)	13 (1.16)	0 (0)	0 (0)	3 (0.95)	4 (0.91)
Race	Asian	945 (5.2)	1614 (2.22)	9 (1.12)	16 (1.43)	263 (5.44)	794 (2.21)	3 (0.95)	6 (1.36)

	Black/Afr. American	855 (4.7)	3622 (4.99)	33 (4.10)	44 (3.94)	248 (5.13)	1876 (5.22)	13 (4.13)	20 (4.55)
	White	16029 (88.08)	65832 (90.71)	615 (76.49)	825 (73.92)	4258 (88.01)	32748 (91.14)	248 (78.73)	330 (75.00)
	Other	85 (0.47)	690 (0.95)	77 (9.58)	126 (11.29)	27 (0.56)	349 (0.97	32 (10.16)	52 (11.82)
	Unknown	284 (1.56)	817 (1.13)	70 (8.71)	105 (9.41)	42 (0.87)	166 (0.46)	19 (6.03)	7.27 (32)
Stage at diagnosis	I	4611 (25.34)	NA	139 (17.29)	190 (17.03)				
	II	2667 (14.64)	NA	101 (12.56)	140 (12.54)				
	III	4713 (25.9)	NA	246 (30.6)	343 (30.73)				
	IV	4838 (26.58)	NA	315 (39.18)	440 (39.43)				
	Unknown	1367 (7.53)	NA	3 (0.37)	3 (0.27)				
Metastatic at diagnosis	Yes	9610 (52.81)	35933 (49.51)	561 (69.78)	783 (70.16)				
	No	7492 (41.17)	28936 (39.87)	240 (29.85)	330 (29.57)				
	Unknown	1097 (6.03)	7706 (10.62)	3 (0.37)	3 (0.27)				
Year of initial diagnosis	Pre-2011								
	2011 onward	18199 (100.00)	72575 (100.00)	804 (100.00)	1116 (100.00)	4838 (100.00)	35933 (100.00)	315 (100.00)	440 (100.00)

Table 9. Characteristics for patients with esophageal/gastric carcinoma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	67821	244200	6181	8837	22747	82765	2867	4230
Age at initial diagnosis	0-19	52 (0.08)	158 (0.06)	0 (0.00)	0 (0.00)	21 (0.09)	56 (0.07)	0 (0.00)	0 (0.00)
	20-34	851 (1.25)	2670 (1.09)	55 (0.89)	73 (0.83)	419 (1.84)	1200 (1.45)	40 (1.40)	54 (1.28)
	35-49	5119 (7.55)	17665 (7.23)	420 (6.80)	581 (6.57)	2245 (9.87)	7373 (8.91)	237 (8.27)	342 (8.09)
	50-64	20955 (30.9)	77326 (31.67)	1988 (32.16)	2765 (31.29)	7973 (35.05)	29581 (35.74)	1039 (36.24)	1472 (34.80)
	65-79	27045 (39.88)	99849 (40.89)	2691 (43.54)	3882 (43.93)	8590 (37.76)	32469 (39.23)	1215 (42.38)	1830 (43.26)
	80+	13899 (20.35)	46532 (19.05)	1027 (16.62)	1530 (17.31)	3499 (15.38)	12086 (14.60)	336 (11.72)	532 (12.58)
	Unknown	0 (0)	0 (0)	0 (0.00)	6 (0.07)				
Sex	Female	22698 (33.47)	77548 (31.76)	1641 (26.55)	2343 (26.51)	6758 (29.71)	22740 (27.48)	745 (25.99)	1097 (25.93)
	Male	45123 (66.53)	166652 (68.24)	4540 (73.45)	6494 (73.49)	15989 (70.29)	60025 (72.52)	2122 (74.01)	3133 (74.07)
Region	Midwest	6046 (8.91)	53062 (21.73)	817 (13.22)	1163 (13.16)	2057 (9.04)	18647 (22.53)	370 (12.91)	551 (13.03)
	Northeast	11655 (17.20)	49838 (20.41)	1626 (26.31)	2302 (26.05)	3657 (16.08)	16872 (20.39)	761 (26.54)	1081 (25.56)
	South	14613 (21.55)	91094 (37.3)	2158 (34.91)	3117 (35.27)	4824 (21.21)	29902 (36.13)	1011 (35.26)	1518 (35.89)
	West	35497 (52.34)	50206 (20.56)	1354 (21.91)	1949 (22.05)	12209 (53.67)	17344 (20.96)	620 (21.63)	929 (21.96)
	Other	NA	NA	95 (1.54)	147 (1.66)	NA	NA	36 (1.26)	64 (1.51)
	Unknown	0 (0)	0 (0)	131(2.12)	159 (1.80)	0 (0)	0 (0)	69 (2.41)	87 (2.06)
Race	Asian	7031 (10.37)	11744 (4.81)	186 (3.01)	285 (3.23)	2265 (9.96)	3492 (4.22)	94 (3.28)	141 (3.33)
	Black/Afr. American	8304 (12.24)	32289 (13.22)	449 (7.26)	673 (7.62)	2779 (12.22)	10877 (13.14)	193 (6.73)	307 (7.26)
	White	51402 (75.79)	196008 (80.27)	4106 (66.43)	5713 (64.65)	17392 (76.46)	67246 (81.25)	1874 (65.36)	2670 (63.12)
	Other	563 (0.83)	2882 (1.18)	694 (11.23)	1039 (11.76)	219 (0.96)	1000 (1.21)	347 (12.1)	545 (12.88)
	Unknown	521 (0.77)	1277 (0.52)	746 (12.07)	1127 (12.75)	92 (0.40)	150 (0.18)	359 (12.52)	567 (13.40)
AJCC stage at diagnosis	0	170 (0.25)	NA	NA	NA				
	I	13414 (19.78)	NA	257 (4.16)	393 (4.45)				
	II	9327 (13.75)	NA	779 (12.60)	1103 (12.48)				
	III	10951 (16.15)	NA	1229 (19.88)	1664 (18.83)				

	IV	22747 (33.54)	NA	2867 (46.38)	4230 (47.87)				
	Unknown	11212 (16.53)	NA	1049 (16.97)	1447 (16.37)				
Metastatic at diagnosis	Yes	23441 (34.56)	82765 (33.89)	2867 (46.38)	4204 (47.57)				
	No	36456 (53.75)	133037 (54.48)	2928 (47.37)	4061 (45.95)				
	Unknown	7924 (11.68)	28398 (11.63)	386 (6.24)	572 (6.47)				
Year of initial diagnosis	Pre-2011			0 (0)	240 (2.72)				
	2011 onward	67821 (100.00)	244200 (100.00)	6181 (100.00)	8591 (97.22)	22747 (100.00)	82765 (100.00)	2867 (100.00)	4230 (100.00)
	Unknown	0 (0)	0 (0)	0 (0)	6 (0.07)				

Table 10. Characteristics for patients with hepatocellular carcinoma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	41443	131285	2744	4003	6908	19028	347	566
Age at initial diagnosis	0-19	69 (0.17)	258 (0.20)	0 (0.00)	0 (0.00)	24 (0.35)	68 (0.36)	0 (0.00)	0 (0.00)
	20-34	244 (0.59)	819 (0.62)	15 (0.55)	18 (0.45)	76 (1.10)	173 (0.91)	2 (0.58)	2 (0.35)
	35-49	1875 (4.52)	5883 (4.48)	93 (3.39)	121 (3.02)	339 (4.91)	929 (4.88)	8 (2.31)	15 (2.65)
	50-64	20494 (49.45)	66437 (50.61)	1212 (44.17)	1671 (41.74)	3495 (50.59)	9588 (50.39)	156 (44.96)	233 (41.17)
	65-79	14625 (35.29)	45378 (34.56)	1117 (40.71)	1752 (43.77)	2358 (34.13)	6477 (34.04)	148 (42.65)	253 (44.70)
	80+	4136 (9.98)	12510 (9.53)	307 (11.19)	441 (11.02)	616 (8.92)	1793 (9.42)	33 (9.51)	63 (11.13)
Sex	Female	9617 (23.21)	29896 (22.77)	660 (24.05)	941 (23.51)	1303 (18.86)	3674 (19.31)	66 (9.02)	102 (18.02)
	Male	31826 (76.79)	101389 (77.23)	2083 (75.91)	3061 (76.47)	5605 (81.14)	15354 (80.69)	281 (80.98)	464 (81.98)
	Unknown	0 (0)	0 (0)	1 (0.04)	1 (0.02)				
Region	Midwest	2965 (7.15)	22598 (17.21)	197 (7.18)	317 (7.92)	522 (7.56)	3361 (17.66)	7.49 (26)	51 (9.01)
	Northeast	5224 (12.61)	23610 (17.98)	860 (31.34)	1238 (30.93)	912 (13.20)	3419 (17.97)	24.78 (86)	135 (23.85)
	South	8332 (20.10)	52854 (40.26)	967 (35.24)	1443 (36.05)	1629 (23.58)	8057 (42.34)	155 (44.67)	254 (44.88)
	West	24922 (60.14)	32223 (24.54)	634 (23.10)	888 (22.18)	3845 (55.66)	4191 (22.03)	71 (20.46	111 (19.61)
	Other	NA	NA	39 (1.42)	61 (1.52)	NA	NA	4 (1.15)	8 (1.41)
	Unknown	0 (0)	0 (0)	47 (1.71)	56 (1.40)	0 (0)	0 (0)	5 (1.44)	7 (1.24)

Race	Asian	6098 (14.71)	9879 (7.52)	135 (4.92)	190 (4.75)	934 (13.52)	1348 (7.08)	20 (5.76)	25 (4.42)
	Black/Afr. American	5781 (13.95)	21529 (16.40)	297 (10.82)	438 (10.94)	1141 (16.52)	3562 (18.72)	37 (10.66)	61 (10.78)
	White	28722 (69.30)	96995 (73.88)	1597 (58.20)	2318 (57.91)	4703 (68.08)	13753 (72.28)	208 (59.94)	328 (57.95)
	Other	582 (1.40)	2486 (1.89)	379 (13.81)	563 (14.06)	105 (1.52)	335 (1.76)	41 (11.82)	77 (13.60)
	Unknown	260 (0.63)	396 (0.30)	336 (12.24)	494 (12.34)	25 (0.36)	30 (0.16)	41 (11.82)	75 (13.25)
AJCC stage at diagnosis	I	14388 (34.72)	NA	179 (6.52)	235 (5.87)				
	II	7252 (17.50)	NA	122 (4.45)	183 (4.57)				
	III	6754 (16.30)	NA	189 (6.89)	274 (6.84)				
	IV	6908 (16.67)	NA	347 (12.65)	566 (14.14)				
	Unknown	6141 (14.82)	NA	1907 (69.50)	2745 (68.57)				
Metastatic at diagnosis	Yes	5716 (13.79)	19028 (14.49)	307 (11.19)	496 (12.39)				
	No	31841 (76.83)	99927 (76.11)	908 (33.09)	1310 (32.73)				
	Unknown	3886 (9.38)	12330 (9.39)	1529 (55.72)	2197 (54.88)				
Year of initial diagnosis	Pre-2011								
	2011 onward	41443 (100.00)	131285 (100.00)	2744 (100.00)	4003 (100.00)	6908 (100.00)	19028 (100.00)	347 (100.00)	566 (100.00)

Table 11. Characteristics for patients with head and neck cancer

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	78674	304366	4448	6298	32396	50486	2707	3517
Age at initial diagnosis	0-19	251 (0.32)	957 (0.31)	0 (0.00)	0 (0.00)	59 (0.18)	287 (0.57)	0 (0.00)	0 (0.00)
	20-34	1004 (1.28)	3599 (1.18)	21 (0.47)	40 (0.64)	300 (0.93)	609 (1.21)	8 (0.30)	9 (0.26)
	35-49	7420 (9.43)	28618 (9.40)	329 (7.40)	489 (7.76)	3304 (10.20)	5027 (9.96)	211 (7.79)	254 (7.22)
	50-64	33683 (42.81)	133310 (43.80)	2054 (46.18)	2908 (46.17)	16067 (49.60)	23700 (46.94)	1326 (48.98)	1692 (48.11)
	65-79	27192 (34.56)	105890 (34.79)	1672 (37.59)	2335 (37.08)	10225 (31.56)	16500 (32.68)	985 (36.39)	1321 (37.56)
	80+	9124 (11.6)	31992 (10.51)	372 (8.36)	492 (7.81)	2441 (7.53)	4363 (8.64)	177 (6.54)	241 (6.85)
	Unknown	0 (0)	0 (0)	0 (0.00)	34 (0.54)				
Sex	Female	20646 (26.24)	79644 (26.17)	1026 (23.07)	1468 (23.31)	6753 (20.85)	11602 (22.98)	560 (20.69)	731 (20.78)
	Male	58028 (73.76)	224722 (73.83)	3422 (76.93)	4830 (76.69)	25643 (79.15)	38884 (77.02)	2147 (79.31)	2786 (79.22)
Region	Midwest	7861 (9.99)	68245 (22.42)	509 (11.44)	718 (11.40)	3212 (9.91)	11143 (22.07)	328 (12.12)	416 (11.83)
	Northeast	11918 (15.15)	53446 (17.56)	862 (19.38)	1206 (19.15)	4818 (14.87)	8917 (17.66)	537 (19.84)	690 (19.62)
	South	21418 (27.22)	126840 (41.67)	2265 (50.92)	3235 (51.37)	9034 (27.89)	21664 (42.91)	1369 (50.57)	1814 (51.58)
	West	37477 (47.64)	55835 (18.34)	655 (14.73)	939 (14.91)	15332 (47.33)	8762 (17.36)	389 (14.37)	499 (14.19)
	Other	NA	NA	106 (2.38)	133 (2.11)	NA	NA	52 (1.92)	64 (1.82)
	Unknown	0 (0)	0 (0)	51 (1.15)	67 (1.06)	0 (0)	0 (0)	32 (1.18)	34 (0.97)

Race	Asian	4766 (6.06)	8218 (2.70)	72 (1.62)	90 (1.43)	1815 (5.60)	1781 (3.53)	41 (1.51)	49 (1.39)
	Black/Afr. American	7855 (9.98)	30393 (9.99)	298 (6.70)	437 (6.94)	3809 (11.76)	7381 (14.62)	196 (7.24)	280 (7.96)
	White	64657 (82.18)	260560 (85.61)	3240 (72.84)	4510 (71.61)	26376 (81.42)	40649 (80.52)	1954 (72.18)	2487 (0.71)
	Other	528 (0.67)	2967 (0.97)	450 (10.12)	649 (10.30)	259 (0.80)	556 (1.10)	279 (10.31)	371 (10.55)
	Unknown	868 (1.10)	2228 (0.73)	388 (8.72)	612 (9.72)	137 (0.42)	119 (0.24)	237 (8.76)	330 (9.38)
AJCC stage at diagnosis	0	NA	NA	0 (0)	2 (0.03)				
	I	16450 (20.91)	NA	231 (5.19)	345 (5.48)				
	II	8715 (11.08)	NA	293 (6.59)	407 (6.46)				
	III	10755 (13.67)	NA	563 (12.66)	762 (12.10)				
	IV	32396 (41.18)	NA	2707 (60.86)	3757 (59.65)				
	Unknown	10358 (13.17)	NA	654 (14.70)	1025 (16.28)				
Metastatic at diagnosis	Yes	6978 (8.87)	50486 (16.59)	2707 (60.86)	3757 (59.65)				
	No	37198 (47.28)	235478 (77.37)	1087 (24.44)	1516 (24.07)				
	Unknown	34498 (43.85)	18402 (6.05)	654 (14.70)	1025 (16.28)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	747 (11.86)				
	2011 onward	78674 (100.00)	304366 (100.00)	4448 (100.00)	5517 (87.60)	32396 (100.00)	50486 (100.00)	2707 (100.00)	3517 (100.00)
	Unknown	0 (0)	0 (0)	0 (0)	34 (0.54)				

Table 12. Characteristics for patients with melanoma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	130722	440999	5480	8682	4925	22281	1175	1777
Age at initial diagnosis	0-19	561 (0.43)	1888 (0.43)	17 (0.31)	31 (0.36)	13 (0.26)	66 (0.30)	0 (0.00)	0 (0.00)
	20-34	7081 (5.42)	23818 (5.40)	277 (5.05)	466 (5.37)	141 (2.86)	807 (3.62)	35 (2.98)	48 (2.70)
	35-49	17818 (13.64)	60341 (13.68)	654 (11.93)	1163 (13.40)	501 (10.17)	2560 (11.49)	112 (9.53)	171 (9.62)
	50-64	40807 (31.24)	135831 (30.80)	1695 (30.93)	2650 (30.52)	1575 (31.98)	7124 (31.97)	386 (32.85)	542 (30.50)
	65-79	43338 (33.18)	148702 (33.72)	2067 (37.72)	3146 (36.24)	1797 (36.49)	7922 (35.55)	438 (37.28)	684 (38.49)
	80+	21017 (16.09)	70419 (15.97)	770 (14.05)	1136 (13.08)	898 (18.23)	3802 (17.06)	204 (17.36)	332 (18.68)
	Unknown	0 (0)	0 (0)	0 (0.00)	90 (1.04)				
Sex	Female	53201 (40.73)	180838 (41.01)	1839 (33.56)	3032 (34.92)	1491 (30.27)	7178 (32.22)	355 (30.21)	552 (31.06)
	Male	77421 (59.27)	260161 (58.99)	3639 (66.41)	5648 (65.05)	3434 (69.73)	15103 (67.78)	820 (69.79)	1225 (68.94)
	Unknown	0 (0)	0 (0)	2 (0.04)	2 (0.02)				
Region	Midwest	10104 (7.74)	93702 (21.25)	763 (13.92)	1215 (13.99)	395 (8.02)	4743 (21.29)	160 (13.62)	256 (14.41)
	Northeast	18850 (14.43)	79006 (17.92)	1202 (21.93)	1893 (21.80)	709 (14.40)	3796 (17.04)	248 (21.11)	360 (20.26)
	South	28865 (22.10)	163515 (37.08)	1810 (33.03)	2977 (34.29)	1047 (21.26)	8827 (39.62)	412 (35.06)	660 (37.14)
	West	72803 (55.74)	104776 (23.76)	1614 (29.45)	2461 (28.35)	2774 (56.32)	4915 (22.06)	336 (28.6)	472 (26.56)
	Other	NA	NA	24 (0.44)	39 (0.45)	NA	NA	6 (0.51)	11 (0.62)

	Unknown	0 (0)	0 (0)	67 (1.22)	97 (1.12)	0 (0)	0 (0)	13 (1.11)	18 (1.01)
Race	Asian	813 (0.62)	1220 (0.28)	14 (0.26)	22 (0.25)	70 (1.42)	130 (0.58)	4 (0.34)	7 (0.39)
	Black/Afr. American	571 (0.44)	2171 (0.49)	24 (0.44)	36 (0.41)	75 (1.52)	350 (1.57)	14 (1.19)	17 (0.96)
	White	122291 (93.62)	416967 (94.55)	4703 (85.82)	7303 (84.12)	4751 (96.47)	21618 (97.02)	960 (81.7)	1427 (80.3)
	Other	306 (0.23)	1918 (0.43)	321 (5.86)	566 (6.52)	15 (0.3)	124 (0.56)	80 (6.81)	142 (7.99)
	Unknown	6641 (5.08)	18723 (4.25)	418 (7.63)	755 (8.70)	14 (0.28)	59 (0.26)	1117 (9.96)	184 (10.35)
AJCC stage at diagnosis	0	2 (0)	NA	15 (0.27)	26 (0.30)				
	I	88348 (67.64)	NA	378 (6.90)	671 (7.73)				
	II	15197 (11.63)	NA	815 (14.87)	1200 (13.82)				
	III	8738 (6.69)	NA	2416 (44.09)	3319 (38.23)				
	IV	4925 (3.77)	NA	1175 (21.44)	1777 (20.47)				
	Unknown	13412 (10.27)	NA	681 (12.43)	1689 (19.45)				
Metastatic at diagnosis	Yes	5607 (4.29)	22281 (5.05)	1176 (21.46)	1781 (20.51)				
	No	118810 (90.96)	380456 (86.27)	3877 (70.75)	5759 (66.33)				
	Unknown	6205 (4.75)	38262 (8.68)	427 (7.79)	1142 (13.15)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	1418 (16.33)				
	2011 onward	130722	440999	5480 (100.00)	7174 (82.63)	4925 (100.00)	22281 (100.00)	1175 (100.00)	1777 (100.00)
	Unknown	0 (0)	0 (0)	0 (0)	90 (1.04)				

Table 13. Characteristics for patients with malignant pleural mesothelioma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	5034	18440	1474	2032	1794	11587	401	576
Age at initial diagnosis	0-19	1 (0.02)	19 (0.10)	0 (0.00)	0 (0.00)	1 (0.06)	0 (0)	0 (0.00)	0 (0.00)
	20-34	51 (1.01)	161 (0.87)	2 (0.14)	2 (0.10)	22 (1.23)	79 (0.68)	0 (0.00)	0 (0.00)
	35-49	207 (4.11)	669 (3.63)	19 (1.29)	27 (1.33)	78 (4.35)	345 (2.98)	5 (1.25)	8 (1.39)
	50-64	907 (18.02)	3280 (17.79)	226 (15.33)	292 (14.37)	369 (20.57)	2067 (17.86)	65 (16.21)	92 (15.97)
	65-79	2286 (45.41)	8724 (47.31)	857 (58.14)	1175 (57.82)	860 (47.94)	5567 (48.09)	221 (55.11)	319 (55.38)
	80+	1582 (31.43)	5587 (30.30)	370 (25.10)	536 (26.38)	464 (25.86)	3518 (30.39)	110 (27.43)	157 (27.26)
Sex	Female	1259 (25.01)	4582 (24.85)	312 (21.17)	447 (22.00)	443 (24.69)	2820 (24.24)	88 (21.95)	131 (22.74)
	Male	3775 (74.99)	13858 (75.15)	1162 (78.83)	1585 (78.00)	1351 (75.31)	8767 (75.66)	313 (78.05)	445 (77.26)
Region	Midwest	441 (8.76)	4307 (23.36)	162 (10.99)	218 (10.73)	142 (7.92)	2727 (23.53)	51 (12.72)	65 (11.28)
	Northeast	971 (19.29)	3891 (21.10)	566 (38.40)	755 (37.16)	342 (19.06)	2482 (21.42)	96 (23.94)	144 (25.00)
	South	962 (19.11)	6247 (33.88)	531 (36.02)	764 (37.60)	341 (19.01)	3882 (33.50)	182 (45.39)	264 (45.83)
	West	2660 (52.84)	3995 (21.66)	189 (12.82)	261 (12.84)	969 (54.01)	2496 (21.54)	59 (14.71)	85 (14.76)
	Other	NA	NA	5 (0.34)	8 (0.39)	NA	NA	3 (0.75)	5 (0.87)
	Unknown	0 (0)	0 (0)	21 (1.42)	26 (1.28)	0 (0)	0 (0)	10 (2.49)	13 (2.26)
Race	Asian	182 (3.62)	265 (1.44)	11 (0.75)	17 (0.84)	73 (4.07)	184 (1.59)	4 (1.00)	6 (1.04)

	Black/Afr. American	260 (5.16)	914 (4.96)	42 (2.85)	54 (2.66)	97 (5.41)	572 (4.94)	11 (2.74)	15 (2.60)
	White	4542 (90.23)	17100 (92.73)	1155 (78.36)	1582 (77.85)	1606 (89.52)	11742 (92.71)	317 (79.05)	446 (77.43)
	Other	28 (0.56)	123 (0.67)	108 (7.33)	159 (7.82)	12 (0.67)	71 (0.61)	26 (6.48)	46 (7.99)
	Unknown	22 (0.44)	38 (0.21)	158 (10.72)	220 (10.83)	6 (0.33)	18 (0.16)	43 (10.72)	63 (10.94)
AJCC stage at diagnosis	I	868 (17.24)	NA	66 (4.48)	108 (5.31)				
	II	416 (8.26)	NA	88 (5.97)	109 (5.36)				
	III	897 (17.82)	NA	262 (17.77)	354 (17.42)				
	IV	1794 (35.64)	NA	401 (27.20)	576 (28.35)				
	Unknown	1059 (21.04)	NA	657 (44.57)	43.55% (885)				
Metastatic at diagnosis	Yes	NA	11587 (62.84)	215 (14.59)	332 (16.34)				
	No	NA	5029 (27.27)	1259 (85.41)	1700 (83.66)				
	Unknown	5034 (100.00)	1824 (9.89)	NA	NA				
Year of initial diagnosis	Pre-2011								
	2011 onward	5034 (100.00)	18440 (100.00)	1474 (100.00)	2032 (100.00)	1794 (100.00)	11587 (100.00)	401 (100.00)	576 (100.00)

Table 14. Characteristics for patients with multiple myeloma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	38519	139422	6927	9696				
Age at initial diagnosis	0-19	10 (0.03)	30 (0.02)	0 (0.00)	0 (0.00)				
	20-34	193 (0.50)	642 (0.46)	27 (0.39)	34 (0.35)				
	35-49	2511 (6.52)	8954 (6.42)	381 (5.50)	522 (5.38)				
	50-64	11557 (30.00)	41842 (30.01)	2130 (30.75)	2871 (29.61)				
	65-79	16660 (43.25)	61575 (44.16)	3180 (45.91)	4552 (46.95)				
	80+	7588 (19.70)	26379 (18.92)	1209 (17.45)	1717 (17.71)				
Sex	Female	16870 (43.80)	61494 (44.11)	3254 (46.98)	4466 (46.06)				
	Male	21649 (56.20)	77928 (55.89)	3673 (53.02)	5230 (53.94)				
Region	West	3852 (10.00)	28852 (20.69)	879 (12.69)	1222 (12.60)				
	Northeast	6420 (16.67)	26944 (19.33)	1936 (27.95)	2681 (27.65)				
	South	9677 (25.12)	57654 (41.35)	2578 (37.22)	3678 (37.93)				
	Midwest	18570 (48.21)	25972 (18.63)	1315 (18.98)	1840 (18.98)				
	Other	NA	NA	108 (1.56)	1.59 (154)				
	Unknown	0 (0)	0 (0)	111 (1.60)	121 (1.251)				
Race	Asian	2182 (5.66)	3358 (2.41)	123 (1.78)	170 (1.75)				

	Black/Afr. American	7759 (20.14)	29076 (20.85)	1115 (16.10)	1564 (16.13)				
	White	27912 (72.46)	104409 (74.89)	4332 (62.54)	5891 (60.76)				
	Other	245 (0.64)	1434 (1.03)	743 (10.73)	1076 (11.10)				
	Unknown	421 (1.09)	1145 (0.82)	614 (8.86)	995 (10.26)				
ISS stage at diagnosis	I	NA	NA	1115 (16.10)	1654 (17.06)				
	II	NA	NA	1065 (15.37)	1598 (16.48)				
	III	NA	NA	1044 (15.07)	1582 (16.32)				
	Unknown	38519 (100.00))	NA	3703 (53.46)	4862 (50.14)				
Metastatic at diagnosis	Yes	36762 (95.44)	132842 (95.28)	6927 (100.00)	9696 (100.00)				
	No	1700 (4.41)	6399 (4.59)	NA	NA				
	Unknown	57 (0.15)	181 (0.13)	NA	NA				
Year of initial diagnosis	Pre-2011								
	2011 onward	38519 (100.00)	139422 (100.00)	6927 (100.00)	9696 (100.00)				

Table 15. Characteristics for patients with non-small cell lung cancer (NSCLC)

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	223742	903355	38782	55554	104816	435937	24147	34530
Age at initial diagnosis	0-19	16 (0.01)	50 (0.01)	2 (0.01)	3 (0.01)	7 (0.01)	30 (0.01)	2 (0.01)	2 (0.01)
	20-34	372 (0.17)	1276 (0.14)	76 (0.20)	107 (0.19)	269 (0.26)	958 (0.22)	59 (0.24)	80 (0.23)
	35-49	6971 (3.12)	28932 (3.20)	1430 (3.69)	1994 (3.59)	4275 (4.08)	18177 (4.17)	1002 (4.15)	1344 (3.89)
	50-64	62207 (27.80)	260318 (28.82)	11879 (30.63)	16916 (30.45)	32016 (30.54)	139497 (32.00)	7681 (31.81)	10794 (31.26)
	65-79	112459 (50.26)	457052 (50.59)	19782 (51.01)	28416 (51.15)	48940 (46.69)	204400 (46.89)	11756 (48.69)	16875 (48.87)
	80+	41717 (18.65)	155727 (17.24)	5613 (14.47)	8084 (14.55)	19309 (18.42)	72875 (16.72)	3647 (15.10)	5435 (15.74)
	Unknown	0 (0)	0 (0)	0 (0.00)	34 (0.06)				
Sex	Female	105777 (47.28)	420421 (46.54)	18389 (47.42)	26330 (47.40)	47681 (45.49)	193871 (44.47)	11312 (46.85)	16071 (46.54)
	Male	117965 (52.72)	482934 (53.46)	20392 (52.58)	29219 (52.60)	57135 (54.51)	242066 (55.53)	12835 (53.15)	18457 (53.45)
	Unknown	0 (0)	0 (0)	1 (0.00)	5 (0.01)	0 (0)	0 (0)	0 (0)	2 (0.01)
Region	Midwest	24555 (10.97)	214378 (23.73)	5788 (14.93)	8279 (14.90)	12012 (11.46)	104741 (24.03)	3536 (14.64)	5107 (14.79)
	Northeast	38444 (17.18)	171817 (19.02)	10395 (26.80)	14967 (26.94)	16637 (15.87)	81940 (18.80)	6412 (26.55)	9157 (26.52)
	South	62488 (27.93)	367990 (40.74)	14952 (38.55)	21649 (38.97)	29020 (27.69)	175273 (40.21)	9151 (37.90)	13280 (38.46)
	West	98255 (43.91)	149170 (16.51)	6502 (16.77)	9255 (16.66)	47147 (44.98)	73983 (16.97)	4240 (17.56)	6029 (17.46)
	Other	NA	NA	259 (0.67)	350 (0.63)	NA	NA	185 (0.77)	253 (0.73)
	Unknown	0 (0)	0 (0)	886 (2.28)	1054 (1.90)	0 (0)	0 (0)	623 (2.58)	704 (2.04)
Race	Asian	15902 (7.11)	24790 (2.74)	945 (2.44)	1395 (2.51)	8573 (8.18)	14043 (3.22)	653 (2.70)	966 (2.80)
	Black/Afr. American	26427 (11.81)	104060 (11.52)	3238 (8.35)	4615 (8.31)	13511 (12.89)	55479 (12.73)	1973 (8.17)	2865 (8.30)
	White	179681 (80.31)	765947 (84.79)	27179 (70.08)	38385 (69.09)	81992 (78.22)	362371 (83.12)	16540 (68.50)	23363 (67.66)
	Other	1148 (0.51)	7091 (0.78)	3435 (8.86)	4991 (8.98)	533 (0.51)	3527 (0.81)	2212 (9.16)	3218 (9.32)
	Unknown	584 (0.26)	1467 (0.16)	3985 (10.28)	6168 (11.10)	207 (0.20)	517 (0.12)	2769 (11.47)	4118 (11.93)
AJCC stage at diagnosis	0	NA	NA	5 (0.01)	5 (0.01)				
	I	53377 (23.86)	NA	3199 (8.25)	4864 (8.76)				
	II	16038 (7.17)	NA	2066 (5.33)	2889 (5.20)				

	III	40833 (18.27)	NA	8206 (21.16)	11306 (20.35)				
	IV	104816 (46.85)	NA	24147 (62.26)	34541 (62.16)				
	Unknown	8628 (3.86)	NA	1159 (2.99)	1959 (3.53)				
Metastatic at diagnosis	Yes	113743 (50.84)	435937 (48.26)	24147 (62.26)	34531 (62.16)				
	No	104852 (46.86)	441376 (48.86)	13476 (34.75)	19064 (34.32)				
	Unknown	5147 (2.30)	26042 (2.88)	1159 (2.99)	1959 (3.53)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	3039 (5.47)				
	2011 onward	223742 (100.00)	903355 (100.00)	38782 (100.00)	52481 (94.47)	104816 (100.00)	435937 (100.00)	24147 (100.00)	34530 (100.00)
	Unknown	0 (0)	0 (0)	0 (0.00)	34 (0.06)				

Table 16. Characteristics for patients with ovarian carcinoma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	41850	147290	4701	6383	11058	79444	887	1280
Age at initial diagnosis	0-19	544 (1.30)	1870 (1.27)	11 (0.23)	11 (0.17)	63 (0.57)	419 (0.53)	0 (0.00)	0 (0.00)
	20-34	1489 (3.56)	5089 (3.46)	93 (1.98)	124 (1.94)	123 (1.11)	1240 (1.56)	4 (0.45)	6 (0.47)
	35-49	5502 (13.15)	18515 (12.57)	545 (11.59)	719 (11.26)	932 (8.43)	7551 (9.50)	72 (8.12)	104 (8.13)
	50-64	14700 (35.13)	51340 (34.86)	1734 (36.89)	2307 (36.14)	3529 (31.91)	26845 (33.79)	280 (31.57)	407 (31.80)
	65-79	13652 (32.63)	50272 (34.13)	1829 (38.91)	2519 (39.46)	4314 (39.01)	31268 (39.36)	397 (44.76)	573 (44.77)
	80+	5963 (14.25)	20204 (13.72)	489 (10.40)	703 (11.01)	2097 (18.96)	12121 (15.26)	134 (15.11)	190 (14.84)
Sex	Female	41850 (100.00)	147290 (100.00)	4700 (99.98)	6382 (99.98)	11058 (100.00)	79444 (100.00)	887 (100.0)	1280 (100.00)
	Unknown	0 (0)	0 (0)	1 (0.02)	1 (0.02)				
Region	Midwest	3830 (9.15)	31718 (21.53)	615 (13.08)	796 (12.47)	939 (8.49)	17013 (21.42)	109 (12.29)	152 (11.88)
	Northeast	6859 (16.39)	28876 (19.60)	1012 (21.53)	1351 (21.17)	1856 (16.78)	15591 (19.63)	195 (21.98)	271 (21.17)
	South	8261 (19.74)	54209 (36.80)	1900 (40.42)	2645 (41.44)	2161 (19.54)	29153 (36.70)	367 (41.38)	533 (41.64)
	West	22900 (54.72)	32487 (22.06)	995 (21.17)	1362 (21.34)	6102 (55.18)	17687 (22.26)	183 (20.63)	278 (21.72)
	Other	NA	NA	79 (1.68)	124 (1.94)	NA	NA	15 (1.69)	26 (2.03)
	Unknown	0 (0)	0 (0)	100 (2.13)	105 (1.64)	0 (0)	0 (0)	18 (2.03)	20 (1.56)
Race	Asian	3633 (8.68)	6026 (4.09)	115 (2.45)	147 (2.30)	835 (7.55)	2878 (3.62)	20 (2.25)	30 (2.34)
	Black/Afr. American	3724 (8.90)	13766 (9.35)	267 (5.68)	358 (5.61)	1207 (10.92)	7644 (9.62)	44 (4.96)	73 (5.70)
	White	33955 (81.14)	125325 (85.09)	3394 (72.20)	4562 (71.47)	8913 (80.60)	67969 (85.56)	641 (72.27)	896 (70.00)
	Other	287 (0.69)	1661 (1.13)	529 (11.25)	744 (11.66)	72 (0.65)	788 (0.99)	97 (10.94)	153 (11.95)
	Unknown	251 (0.6)	512 (0.35)	396 (8.42)	572 (8.96)	31 (0.28)	165 (0.21)	85 (9.58)	128 (10.00)
AJCC stage at diagnosis	I	9038 (21.6)	NA	802 (17.06)	1060 (16.61)				
	II	3358 (8.02)	NA	355 (7.55)	487 (7.63)				
	III	15009 (35.86)	NA	1940 (41.27)	2573 (40.31)				

	IV	11058 (26.42)	NA	887 (18.87)	1280 (20.05)				
	Unknown	3387 (8.09)	NA	717 (15.25)	983 (15.40)				
Metastatic at diagnosis	Yes	19889 (47.52)	79444 (53.94)	886 (18.85)	1279 (20.04)				
	No	12822 (30.64)	58973 (40.04)	3258 (69.30)	4320 (67.68)				
	Unknown	9139 (21.84)	8873 (6.02)	557 (11.85)	784 (12.28)				
Year of initial diagnosis	Pre-2011								
	2011 onward	41850 (100.00)	147290 (100.00)	4701 (100.00)	6383 (100.00)	11058 (100.00)	79444 (100.00)	887 (100.00)	1280 (100.00)

Table 17. Characteristics for patients with pancreatic carcinoma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	38071	134201	4595	7603	17995	67027	2723	5058
Age at initial diagnosis	0-19	49 (0.13)	161 (0.12)	0 (0.00)	0 (0.00)	7 (0.04)	16 (0.02)	0 (0.00)	0 (0.00)
	20-34	260 (0.68)	795 (0.59)	13 (0.28)	20 (0.26)	69 (0.38)	269 (0.40)	11 (0.40)	16 (0.32)
	35-49	1731 (4.55)	6375 (4.75)	181 (3.94)	294 (3.87)	779 (4.33)	3111 (4.64)	108 (3.97)	196 (3.88)
	50-64	10588 (27.81)	38699 (28.84)	1501 (32.67)	2411 (31.71)	5450 (30.29)	20531 (30.63)	864 (31.73)	1564 (30.92)
	65-79	16382 (43.03)	59551 (44.37)	2333 (50.77)	3914 (51.48)	8055 (44.76)	30238 (45.11)	1385 (50.86)	2618 (51.76)
	80+	9061 (23.80)	28620 (21.33)	567 (12.34)	962 (12.65)	3635 (20.20)	12862 (19.19)	355 (13.04)	664 (13.13)
	Unknown	0 (0)	0 (0)	0 (0.00)	2 (0.03)				
Sex	Female	18631 (48.94)	64585 (48.13)	2154 (46.88)	3525 (46.36)	8433 (46.86)	30977 (46.22)	1241 (45.57)	2267 (44.82)
	Male	19440 (51.06)	69616 (51.87)	2441 (53.12)	4078 (53.64)	9562 (53.14)	36050 (53.78)	1482 (54.43)	2791 (55.18)
Region	Midwest	3763 (9.88)	29252 (21.80)	549 (11.95)	903 (11.88)	1869 (10.39)	14554 (21.71)	315 (11.57)	593 (11.72)
	Northeast	6485 (17.03)	26111 (19.46)	1372 (29.86)	2216 (29.15)	3093 (17.19)	13360 (19.93)	800 (29.38)	1442 (28.51)
	South	8618 (22.64)	51739 (38.55)	1798 (39.13)	3060 (40.25)	3980 (22.12)	25698 (38.34)	1078 (39.59)	2064 (40.81)
	West	19205 (50.45)	27099 (20.19)	745 (16.21)	1248 (16.41)	9053 (50.31)	13415 (20.21)	446 (16.38)	16.55 (837)
	Other	NA	NA	42 (0.91)	64 (0.84)	NA	NA	27 (0.99)	46 (0.91)
	Unknown	0 (0)	0 (0)	89 (1.94)	112 (1.47)	0 (0)	0 (0)	57 (2.09)	76 (1.50)

Race	Asian	2918 (7.66)	4149 (3.09)	76 (1.65)	124 (1.63)	1310 (7.28)	2044 (3.05)	48 (1.76)	86 (1.70)
	Black/Afr. American	4594 (12.07)	17264 (12.86)	376 (8.18)	636 (8.37)	2291 (12.73)	8946 (13.35)	228 (8.37)	431 (8.52)
	White	30140 (79.17)	111120 (82.80)	3304 (71.90)	5299 (69.70)	14207 (78.95)	55274 (82.47)	1911 (70.18)	3446 (68.13)
	Other	255 (0.67)	1299 (0.97)	411 (8.94)	751 (9.88)	129 (0.72)	645 (0.96)	227 (8.34)	488 (9.65)
	Unknown	164 (0.43)	369 (0.27)	428 (9.31)	793 (10.43)	58 (0.32)	118 (0.18)	309 (11.35)	607 (12.00)
AJCC stage at diagnosis	0	1 (0.00)	NA	NA	NA				
	I	4146 (10.89)	NA	172 (3.74)	252 (3.31)				
	II	8692 (22.83)	NA	977 (21.26)	1276 (16.78)				
	III	3067 (8.06)	NA	318 (6.92)	464 (6.10)				
	IV	17995 (47.27)	NA	2723 (59.26)	5058 (66.53)				
	Unknown	4170 (10.95)	NA	405 (8.81)	553 (7.27)				
Metastatic at diagnosis	Yes	18862 (49.54)	67027 (49.95)	2723 (59.26)	5058 (66.53)				
	No	15790 (41.48)	58611 (43.67)	1467 (31.93)	1992 (26.20)				
	Unknown	3419 (8.98)	8563 (6.38)	405 (8.81)	553 (7.27)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	53 (0.7)				
	2011 onward	38071 (100.00)	134201 (100.00)	4595 (100.00)	7548 (99.28)	17995 (100.00)	67027 (100.00)	2723 (100.00)	5058 (100.00)
	Unknown	0 (0)	0 (0)	0 (0)	2 (0.03)				

Table 18. Characteristics for patients with prostate cancer

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	200357	714054	4704	10295	20844	49802	3196	4792
Age at initial diagnosis	0-19	7 (0.00)	26 (0.00)	0 (0.00)	0 (0.00)	NA	0 (0)	0 (0.00)	0 (0.00)
	20-34	20 (0.01)	53 (0.01)	0 (0.00)	1 (0.01)	1 (0.00)	0 (0)	0 (0.00)	0 (0.00)
	35-49	4446 (2.22)	15776 (2.21)	77 (1.64)	193 (1.87)	370 (1.78)	777 (1.56)	58 (1.81)	73 (1.52)
	50-64	77595 (38.73)	280356 (39.26)	1260 (26.79)	3456 (33.57)	6591 (31.62)	13188 (26.49)	840 (26.28)	1224 (25.55)
	65-79	99720 (49.77)	358575 (50.22)	2432 (51.70)	5093 (49.47)	9235 (44.31)	21821 (43.83)	1568 (49.06)	2337 (48.78)
	80+	18569 (9.27)	59268 (8.30)	935 (19.88)	1397 (13.57)	4647 (22.29)	14001 (28.12)	730 (22.84)	1157 (24.15)
	Unknown	0 (0)	0 (0)	0 (0.00)	155 (1.51)				
Sex	Male	200357 (100.00)	714054 (100.00)	4704 (100.00)	10295 (100.00)	20844 (100.00)	49802 (100.00)	3196 (100.00)	4792 (100.00)
	Female								
Region	Midwest	19144 (9.55)	154217 (21.60)	630 (13.39)	1320 (12.82)	1964 (9.42)	10992 (22.07)	413 (12.92)	625 (13.04)
	Northeast	35470 (17.70)	139879 (19.59)	947 (20.13)	2169 (21.07)	3043 (14.60)	9739 (19.56)	678 (21.21)	1054 (21.99)
	South	48890 (24.40)	280424 (39.27)	2028 (43.11)	4451 (43.23)	4300 (20.63)	17915 (35.97)	1335 (41.77)	1973 (41.17)
	West	96853 (48.34)	139534 (19.54)	991 (21.07)	2145 (20.84)	11537 (55.35)	11156 (22.40)	706 (22.09)	1054 (21.99)
	Other	NA	NA	69 (1.47)	128 (1.24)	NA	NA	39 (1.22)	53 (1.11)
	Unknown	0 (0)	0 (0)	39 (0.83)	82 (0.80)	0 (0)	0 (0)	25 (0.78)	33 (0.69)
Race	Asian	9708 (4.85)	14660 (2.05)	78 (1.66)	152 (1.48)	1199 (5.75)	1203 (2.42)	63 (1.97)	89 (1.86)
	Black/Afr. American	31316 (15.63)	118387 (16.58)	494 (10.50)	974 (9.46)	3359 (16.11)	9246 (18.57)	357 (11.17)	503 (10.50)
	White	148728 (74.23)	551490 (77.23)	3147 (66.90)	6940 (67.41)	15961 (76.57)	38615 (77.54)	2132 (66.71)	3122 (65.15)

	Other	826 (0.41)	7235 (1.01)	503 (10.69)	1080 (10.49)	134 (0.64)	510 (1.02)	336 (10.51)	533 (11.12)
	Unknown	9779 (4.88)	22282 (3.12)	482 (10.25)	1149 (11.16)	191 (0.92)	228 (0.46)	308 (9.64)	545 (11.37)
AJCC stage at diagnosis	I	11222 (5.60)	NA	22 (0.47)	78 (0.76)				
	II	129755 (64.76)	NA	195 (4.15)	546 (5.30)				
	III	17365 (8.67)	NA	123 (2.61)	305 (2.96)				
	IV	20844 (10.40)	NA	3196 (67.94)	4971 (48.29)				
	Unknown	21171 (10.57)	NA	1168 (24.83)	4395 (42.69)				
Metastatic at diagnosis	Yes	14399 (7.19)	49802 (6.97)	3043 (64.69)	4670 (45.36)				
	No	172918 (86.30)	611692 (85.66)	1056 (22.45)	3152 (30.62)				
	Unknown	13040 (6.51)	52560 (7.36)	605 (12.86)	2473 (24.02)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	3755 (36.47)				
	2011 onward	200357 (100.00)	714054 (100.00)	4704 (100.00)	6386 (62.03)	20844 (100.00)	49802 (100.00)	3196 (100.00)	4792 (100.00)
	Unknown	0 (0)	0 (0)	0 (0)	154 (1.50)				

Table 19. Characteristics for patients with renal-cell carcinoma

		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2011-16	2011-19			2011-16	2011-19
	N	87634	343647	4470	7278	12903	48545	2706	3830
Age at initial diagnosis	0-19	1019 (1.16)	3612 (1.05)	0 (0.00)	3 (0.04)	27 (0.21)	892 (1.84)	0 (0.00)	0 (0.00)
	20-34	1759 (2.01)	6362 (1.85)	30 (0.67)	56 (0.77)	133 (1.03)	451 (0.93)	17 (0.63)	24 (0.63)
	35-49	11023 (12.58)	40803 (11.87)	341 (7.63)	659 (9.05)	1060 (8.22)	3450 (7.11)	188 (6.95)	278 (7.26)
	50-64	32297 (36.85)	123826 (36.03)	1748 (39.11)	2957 (40.63)	4742 (36.75)	17087 (35.20)	1039 (38.40)	1445 (37.73)
	65-79	31918 (36.42)	129839 (37.78)	1929 (43.15)	2989 (41.07)	4893 (37.92)	18961 (39.06)	1161 (42.90)	1655 (43.21)
	80+	9618 (10.98)	39205 (11.41)	422 (9.44)	593 (8.15)	2048 (15.87)	7704 (15.87)	301 (11.12)	428 (11.17)
	Unknown	0 (0)	0 (0)	0 (0.00)	21 (0.29)				
Sex	Female	31767 (36.25)	126876 (36.92)	1372 (30.69)	2285 (31.40)	4162 (32.26)	16195 (33.36)	841 (31.08)	1207 (31.51)
	Male	55867 (63.75)	216771 (63.08)	3098 (69.31)	4993 (68.60)	8741 (67.74)	32350 (66.64)	1865 (68.92)	2623 (68.49)
Region	Midwest	8601 (9.81)	77510 (22.56)	727 (16.26)	1119 (15.38)	1269 (9.83)	11321 (23.32)	448 (16.56)	596 (15.56)
	Northeast	13291 (15.17)	61443 (17.88)	887 (19.84)	1464 (20.12)	1729 (13.40)	8158 (16.81)	541 (19.99)	742 (19.37)
	South	22294 (25.44)	137728 (40.08)	1729 (38.68)	2884 (39.63)	3274 (25.37)	19326 (39.81)	1042 (38.51)	1537 (40.13)
	West	43448 (49.58)	66966 (19.49)	995 (22.26)	1604 (22.04)	6631 (51.39)	9740 (20.406)	591 (21.84)	846 (22.09)
	Other	NA	NA	53 (1.19)	85 (1.17)	NA	NA	29 (1.07)	40 (1.04)
	Unknown	0 (0)	0 (0)	79 (1.77)	122 (1.68)	0 (0)	0 (0)	55 (2.03)	69 (1.80)

Race	Asian	4610 (5.26)	7186 (2.09)	74 (1.66)	106 (1.46)	743 (5.76)	1094 (2.25)	45 (1.66)	60 (1.57)
	Black/Afr. American	11017 (12.57)	42300 (12.31)	319 (7.14)	492 (6.76)	1363 (10.56)	5062 (10.143	186 (6.87)	287 (7.49)
	White	70387 (80.32)	287900 (83.78)	3079 (68.88)	4997 (68.66)	10627 (82.36)	41707 (85.91)	1864 (68.88)	2577 (67.28)
	Other	862 (0.98)	4828 (1.40)	567 (12.68)	911 (12.52)	140 (1.09)	602 (1.24)	329 (12.16)	475 (12.40)
	Unknown	758 (0.86)	1433 (0.42)	431 (9.64)	772 (10.61)	30 (0.23)	80 (0.16)	282 (10.42)	431 (11.25)
AJCC stage at diagnosis	I	50746 (57.91)	NA	NA	NA				
	II	6332 (7.23)	NA	NA	NA				
	III	11131 (12.70)	NA	NA	NA				
	IV	12903 (14.72)	NA	2706 (60.54)	3830 (52.62)				
	Unknown	6522 (7.44)	NA	122 (2.73)	330 (4.53)				
	Unspecified non-metastatic	NA	NA	1642 (36.73)	3118 (42.84)				
Metastatic at diagnosis	Yes	12320 (14.06)	48545 (14.13)	2612 (58.43)	3718 (51.09)				
	No	72025 (82.19)	282876 (82.32)	1726 (38.61)	3219 (44.23)				
	Unknown	3289 (3.75)	12226 (3.56)	132 (2.95)	341 (4.69)				
Year of initial diagnosis	Pre-2011	NA	NA	0 (0)	1442 (19.81)				
	2011 onward	87634	343647	4470 (100.00)	5815 (79.90)	12903 (100.00)	48545 (100.00)	2706 (100.00)	3830 (100.00)
	Unknown	0 (0)	0 (0)	0 (0)	21 (0.29)				

Table 20. Characteristics for patients with small-cell lung cancer

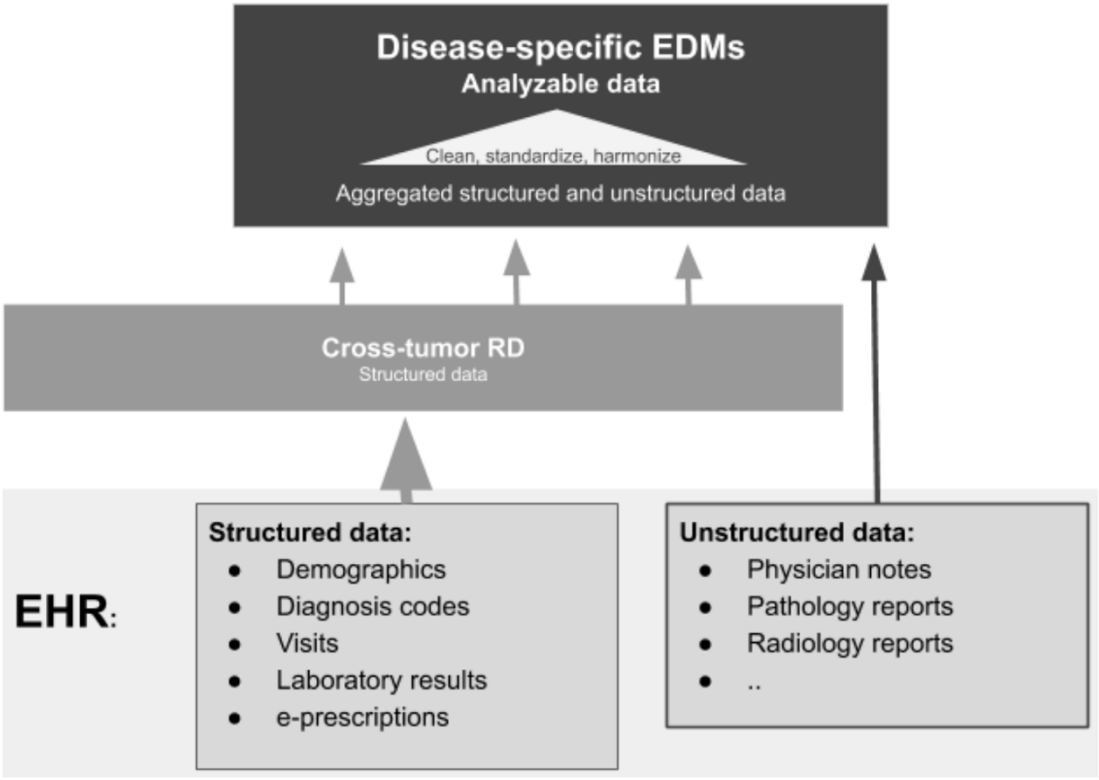
		All eligible, n (%)				Stage IV at diagnosis, n (%)			
		SEER	NPCR	Flatiron Health		SEER	NPCR	Flatiron Health	
				2013-16	2013-19			2013-16	2013-19
	N	23849	105616	3916	6188	16302	74164	2151	3483
Age at initial diagnosis	0-19	2 (0.01)	0 (0)	0 (0.00)	0 (0.00)	1 (0.01)	0 (0)	0 (0.00)	0 (0.00)
	20-34	17 (0.07)	64 (0.06)	2 (0.05)	6 (0.10)	9 (0.06)	43 (0.06)	1 (0.05)	4 (0.11)
	35-49	618 (2.59)	2904 (2.75)	108 (2.76)	161 (2.60)	418 (2.56)	2119 (2.86)	66 (3.07)	103 (2.96)
	50-64	7881 (33.05)	335883 (33.98)	1395 (35.62)	2143 (34.63)	5439 (33.36)	25782 (34.77)	776 (36.08)	1240 (35.60)
	65-79	12231 (51.29)	54162 (51.28)	2024 (51.69)	3253 (52.57)	8298 (50.90)	37422 (50.46)	1108 (51.51)	1790 (51.39)
	80+	3100 (13.00)	12598 (11.93)	387 (9.88)	625 (10.10)	2137 (13.11)	8794 (11.86)	200 (9.30)	346 (9.93)
Sex	Female	12056 (50.55)	53816 (50.95)	2021 (51.61)	3247 (52.47)	7930 (48.64)	36355 (49.02)	1012 (47.05)	1678 (48.18)
	Male	11793 (49.45)	51800 (49.05)	1895 (48.39)	2940 (47.51)	8372 (51.36)	37809 (50.98)	1139 (52.95)	1804 (51.79)
	Unknown	0 (0)	0 (0)	0 (0)	1 (0.02)	0 (0)	0 (0)	0 (0)	1 (0.03)
Region	Midwest	3156 (13.23)	27704 (26.23)	696 (17.77)	1067 (17.24)	2248 (13.79)	19755 (26.64)	409 (19.01)	618 (17.74)
	Northeast	3564 (14.94)	17775 (16.83)	936 (23.90)	1436 (23.21)	2364 (14.50)	12717 (17.15)	529 (24.59)	841 (24.15)
	South	8506 (35.67)	45754 (43.32)	1682 (42.95)	2717 (43.91)	5806 (35.62)	31486 (42.45)	887 (41.24)	1482 (42.55)
	West	8623 (36.16)	14383 (13.62)	516 (13.18)	861 (13.91)	5884 (36.09)	10206 (13.76)	279 (12.97)	484 (13.9)
	Other	NA	NA	14 (0.36)	22 (0.36)	NA	NA	4 (0.19)	7 (0.20)
	Unknown	0 (0)	0 (0)	72 (1.84)	85 (1.37)	0 (0)	0 (0)	43 (2.00)	51 (1.46)

Race	Asian	868 (3.64)	1250 (1.18)	34 (0.87)	54 (0.87)	557 (3.42)	863 (1.16)	14 (0.65)	24 (0.69)
	Black/Afr. American	2138 (8.96)	8460 (8.01)	221 (5.64)	350 (5.66)	1477 (9.06)	5896 (7.95)	111 (5.16)	182 (5.23)
	White	20654 (86.60)	94867 (89.82)	3055 (78.01)	4673 (75.52)	14144 (86.76)	66684 (89.91)	1681 (78.15)	2633 (75.60)
	Other	147 (0.62)	889 (0.84)	327 (8.35)	540 (8.73)	102 (0.63)	631 (0.85)	179 (8.32)	303 (8.70)
	Unknown	42 (0.18)	150 (0.14)	279 (7.12)	571 (9.23)	22 (0.13)	90 (0.12)	166 (7.72)	341 (9.79)
AJCC stage at diagnosis	I	965 (4.05)	NA	122 (3.12)	210 (3.39)				
	II	705 (2.96)	NA	133 (3.40)	217 (3.51)				
	III	5183 (21.73)	NA	592 (15.12)	1005 (16.24)				
	IV	16302 (68.36)	NA	2151 (54.93)	3483 (56.29)				
	Unknown	694 (2.91)	NA	918 (23.44)	1273 (20.57)				
Metastatic at diagnosis	Yes	17449 (73.16)	74164 (70.22)	2151 (54.93)	3483 (56.29)				
	No	5942 (24.92)	28777 (27.25)	847 (21.63)	1432 (23.14)				
	Unknown	458 (1.92)	2675 (2.53)	918 (23.44)	1273 (20.57)				
Year of initial diagnosis	Pre-2013								
	2013 onward	23849 (100.00)	105616 (100.00)	3916 (100.00)	6188 (100.00)	16302 (100.00)	74164 (100.00)	2151 (100.00)	3483 (100.00)

FIGURES

Figure 1. Schema of the structure of the Flatiron Health databases

RD=research database; EDM=enhanced data mart



APPENDIX I

Table A1. Overview of diagnostic and histology codes used to define eligibility, and of additional entry criteria and definitions of 'metastatic at diagnosis'.

Tumor type	Flatiron Health Databases (EDMs)	SEER	NPCR
Advanced urothelial cancer	ICD-9 188x, 189.1, 189.2, 189.3, or ICD-10 C65x, C66x, C67x, C68.0 Metastatic definition: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2011.	Primary Site = "659-680" Metastatic definition: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "in situ", "Local" or "Regional"	Primary Site - Labeled = "C65.9 - C68.0" Metastatic definition: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "In situ", "Local" or "Regional"
Metastatic breast cancer	ICD-9: 174.x or 175.x or ICD-10 C50.x Metastatic definition: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known). Entered into database if two clinical encounters on or after January 1 2011. ICD-9: 174.x or 175.x or ICD-10 C50.x	Primary Site = "500-509" Metastatic: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "Local" or "Regional"	Primary Site - Labeled = "C50.0 - C50.9" Metastatic: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "Local" or "Regional"
Early breast cancer			Primary Site - Labeled = "C50.0 - C50.9" Metastatic: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "Local" or "Regional"
Chronic lymphocytic leukemia	ICD-9: 204.1x or ICD-10: C91.1x, C83.0x Metastatic definition: all stages Entered into database if two clinical encounters on or after January 1 2011.	Site recode ICD-O-3/WHO 2008 = "Chronic Lymphocytic Leukemia" AND Histologic type ICD-O-3 = "9823" Metastatic definition: Summary stage	Site recode ICD-O-3/WHO 2008 = "Chronic Lymphocytic Leukemia" AND Histologic type ICD-O-3 = "9823" Metastatic definition: Summary stage

		= “Distant”	= “Distant”
Metastatic colorectal cancer	ICD-9 153.x or 154.x or ICD-10 C18x, or C19x, or C20x, or C21x Metastatic definition: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2013.	Primary Site = “180-218” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Primary Site - Labeled = “C18.0 - C21.8” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”
Diffuse large B-cell lymphoma	ICD 9: 200x, 202x; ICD 10: C82x, C83x, C84x, C85x, C86x, C88x, C96x Metastatic definition: Stage IV or Stage III Non-Metastatic: Stage I or II Entered into database if two clinical encounters on or after January 1 2011.	Histologic type ICD-O-3 = “9680” Metastatic definition: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Histologic type ICD-O-3 = “9680” Metastatic definition: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”
Follicular lymphoma	ICD 9: 200x, 202x; ICD 10: C82x, C83x, C84x, C85x, C86x, C88x, C96x Metastatic: Stage IV or Stage III Non-Metastatic: Stage I or II Entered into database if two clinical encounters on or after January 1 2011.	Histologic type ICD-O-3 = “9690, 9691, 9695, 9698” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Histologic type ICD-O-3 = “9690, 9691, 9695, 9698” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”
Advanced gastric/esophageal carcinoma	ICD-9: 151.x, 150.x, ICD-9 C16.x, C15.x Metastatic definition: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2011.	Primary Site = “150-169” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Primary Site - Labeled = “C15.0 - C16.9” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”

Hepatocellular carcinoma	ICD-9: 155.x OR ICD-10 C22.x Metastatic: Stage IV (not otherwise specified) or Stage IVB, M1 Non-Metastatic: M0 or non-stage IV/IVB (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2011.	Primary Site = “220, 221” AND Histologic type ICD-O-3 = “8170, 8171, 8172, 8173, 8174, 8175” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Primary Site - Labeled = “C22.0, C22.1” AND Histologic type ICD-O-3 = “8170-8175” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”
Advanced head and neck cancer	140x, 141x, 143x, 144x, 145x, 146x, 147x, 148x, 149x, 161x; ICD 10: C00x, C01x, C02x, C03x, C04x, C05x, C06x, C09x, C10x, C11x, C12x, C13x, C14x, C32x Metastatic definition: Stage IVC, M1 Non-Metastatic: M0 or non-stage IVC (including unknown AJCC group stage if MO is known and IVA, IVB) Entered into database if two clinical encounters on or after January 1 2011.	Primary Site = “0-69, 90-148, 320-329” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Primary Site - Labeled = “ C00.0 - C06.9, C09.0 - C14.8, C32.0 - C32.9” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”
Advanced melanoma	ICD-9 172.x or ICD-10 C43x or D03x Metastatic: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2011.	Site recode ICD-O-3/WHO 2008 = "Melanoma of the Skin" Met: Summary stage 2000 = “Distant”; Non-Met: Summary stage 2000 = “Local” or “Regional”	Site recode ICD-O-3/WHO 2008 = “Melanoma of the skin” Met: Summary stage 2000 = “Distant”; Non-Met: Summary stage 2000 = “Local” or “Regional
Malignant pleural mesothelioma	ICD-9:163* or ICD-10 C45.0 Metastatic: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical	Site recode ICD-O-3/WHO 2008 = “Mesothelioma” AND Histologic Type 9050-9053 Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Site recode ICD-O-3/WHO 2008 = “Mesothelioma” AND Histologic Type 9050-9053 Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”

	encounters on or after January 1 2011.		
Multiple myeloma	ICD-9 203.0x or ICD-10 C90.0x, C90 Metastatic: all known stages. Entered into database if two clinical encounters on or after January 1 2011.	Site recode ICD-O-3/WHO 2008 = "Myeloma" Met: Summary stage = "Distant"	Site recode ICD-O-3/WHO 2008 = "Myeloma" Met: Summary stage = "Distant"
Advanced non-small cell lung cancer (NSCLC)	ICD-9 162.x or ICD-10 C34x, or C39.9 Metastatic: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2011.	Primary Site = "340-349, 390-399" AND Histologic type ICD-O-3 = "8046, 8033, 8022, 8012, 8980, 8140, 8560, 8070, 8550, 8250, 8251, 8252, 8253, 8254, 8255, 8310, 8470, 8083, 8052, 8084, 8071, 8072, 8073, 8480, 8481, 8260, 8490, 8230, 8012, 8013, 8014, 8082, 8123, 8310" Met: Summary stage 2000 = "Distant"; Non-Met: Summary stage 2000 = "Local" or "Regional"	Primary Site - Labeled = "C34.0 - C34.9, C39.9" AND Histologic type ICD-O-3 = "8046, 8033, 8022, 8012, 8980, 8140, 8560, 8070, 8550, 8250, 8251, 8252, 8253, 8254, 8255, 8310, 8470, 8083, 8052, 8084, 8071, 8072, 8073, 8480, 8481, 8260, 8490, 8230, 8012, 8013, 8014, 8082, 8123, 8310" Met: Summary stage 2000 = "Distant"; Non-Met: Summary stage 2000 = "Local" or "Regional"
Ovarian carcinoma	ICD 9: 183x, 158x; ICD 10: C56x, C57.0x, C48x Metastatic: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2011.	Primary Site = "569, 570, 480-488" Metastatic: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "Local" or "Regional"	Primary Site - Labeled = "C48.0 - C48.8, C56.9, C57.0" Metastatic: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "Local" or "Regional"
Metastatic pancreatic carcinoma	ICD-9 157.x; ICD-10 C25.x Metastatic: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical	Primary Site = "250-259" Metastatic: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "Local" or "Regional"	Primary Site - Labeled = "C25.0 - C25.9" Metastatic: Summary stage 2000 = "Distant"; Non-Metastatic: Summary stage 2000 = "Local" or "Regional"

	encounters on or after January 1 2014.		
Metastatic prostate cancer	ICD 9 code: 185x ICD 10 code: C61x Metastatic: Stage IV, M1 Non-Metastatic: M0 or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2013.	Primary Site = “619” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Primary Site - Labeled = “C61.9” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”
Metastatic renal-cell carcinoma	ICD-9 189.x or ICD-10 C64x or C65x Metastatic: Stage IV (excluding patients who are T4NxM0), M1 Non-Metastatic: M0, T4 only Stage IV or non-stage IV (including unknown AJCC group stage if MO is known) Entered into database if two clinical encounters on or after January 1 2011.	Primary Site = “649, 659” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Primary Site - Labeled = “C64.9, C65.9” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”
Small-cell lung cancer (SCLC)	ICD-9 162.x or ICD-10 C34x, or C39.9 Metastatic: Extensive stage and stage IV or MI Non-Metastatic: limited stage or MO or non-Stage IV (including AJCC group stage if MO is unknown) Entered into database if two clinical encounters on or after January 1 2013.	Primary Site = “340-349, 390-399” and Histologic type ICD-O-3 = “8041-8045” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”	Primary Site - Labeled = ”C34.0 - C34.9, C39.9” AND Histologic type ICD-O-3 = “8041- 8045” Metastatic: Summary stage 2000 = “Distant”; Non-Metastatic: Summary stage 2000 = “Local” or “Regional”

Table A2. Overview of variable definitions and steps taken to align variables across the different databases.

	Definitions		
	SEER	NPCR	Flatiron Health
Age at initial diagnosis			
Region	Assigned as: West, Northeast, South, Midwest, other according to US census definitions.		
Race	Defined as: ● Asian = “Asian” or “Pacific Islander” ● Black or African American = “Black” ● Other Race = “American Indian” or “Alaska Native” ● Unknown = “Unknown” ● White = “White”		As self-reported by patients, and captured into the EHR. Defined as: ● White = “White” ● Black or African American = “Black or African American” ● Asian = “Asian” ● Other Race = “Other” ● Unknown = “Unknown”, “Hispanic or Latino” Categories had to be combined to harmonize comparisons across databases
Stage at diagnosis	Defined as: ● Derived AJCC Stage Group, 7th ed for diagnosis before 2016. ● Derived SEER Cmb Stg Grp for diagnosis in 2016.		As reported by the treating clinician in the EHR, following AJCC criteria, 7th or 8th edition. Or as noted by a Flatiron abstractor via assessment of T, N, M components, using AJCC 7th ed guidelines until July, 2018 and AJCC 8th ed guidelines thereafter.

Tables A3.A-S: NA = not available from the source data, or numbers under the reporting suppression value to preserve patient confidentiality.

Table A3. Distribution of patient characteristics excluding patients with unknown/missing status across the categories evaluated

A. Advanced urothelial (bladder) cancer

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Age at initial diagnosis	0-19	0.05	0.05	0.00	0.00
	20-34	0.45	0.40	0.08	0.12
	35-49	3.14	3.11	2.50	2.64
	50-64	22.17	22.46	23.38	23.52
	65-79	45.22	46.28	52.92	53.13
	80+	28.98	27.69	21.13	20.59
Sex	Female	25.06	25.12	26.17	26.03
	Male	74.94	74.88	73.83	73.97
Region	Midwest	10.38	23.52	13.81	13.36
	Northeast	19.64	22.36	23.26	23.08
	South	21.22	34.25	44.69	44.99
	West	48.76	19.87	17.48	17.74
Race	Asian	4.59	1.93	1.52	1.42
	Black	6.09	5.73	4.51	4.31
	White	88.94	91.65	84.42	83.62
	Other	0.38	0.70	9.55	10.66
Stage	0	48.63	NA	0.89	1.01
	I	24.10	NA	3.16	3.12
	II	12.38	NA	12.04	12.14
	III	5.49	NA	12.89	12.81
	IV	9.40	NA	71.03	70.92
Metastatic at diagnosis	Yes	5.76	5.65	42.57	42.59
	No	94.24	94.35	57.43	57.41
Year of diagnosis	2011	16.17	16.16	12.85	8.68
	2012	16.63	16.55	15.57	10.51
	2013	16.52	16.70	16.76	11.32
	2014	16.86	16.93	18.51	12.50
	2015	17.06	16.94	18.38	12.41
	2016	16.75	16.72	17.93	12.11
	2017 - 2019	NA	NA	0	19.85

Prior to 2011		NA	NA	0	12.61

B. Metastatic breast cancer

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	99.21	99.11	98.60	98.75
	Male	0.79	0.89	1.40	1.25
Age at diagnosis	0-19	0.01	0.01	0.00	0.01
	20-34	1.87	1.80	3.43	3.31
	35-49	17.41	16.64	19.03	22.02
	50-64	37.05	36.59	37.30	39.44
	65-79	32.46	33.75	30.43	28.12
	80+	11.20	11.22	9.80	7.10
Region	Midwest	8.76	22.06	14.29	14.57
	Northeast	16.65	19.94	22.97	23.74
	South	21.65	36.38	39.92	38.69
	West	52.95	21.62	20.15	20.62
	Other	NA	NA	2.68)	2.37
Race	Asian	8.66	4.00	2.63	2.42
	Black	11.37)	11.74	13.72	11.86
	White	79.37	83.24	71.13	73.17
	Other	0.60	1.02	12.51	12.56
Stage	0	0.15	NA	0.01	0.05
	I	49.01	NA	7.76	11.95
	II	34.22	NA	21.77	28.56
	III	10.69	NA	23.42	23.74
	IV	5.93	NA	47.04	35.71
Metastatic at diagnosis	Yes	6.00	6.06	47.04	35.71
	No	94.00	93.94	52.96	64.29
Year of diagnosis	2011	15.92	15.89	17.63	9.11
	2012	16.21	16.15	18.36	9.48
	2013	16.56	16.60	17.71	9.15
	2014	16.82	16.86	17.21	8.89
	2015	17.28	17.24	15.51	8.01
	2016	17.21	17.26	13.57	7.01
	2017 - 2019	NA	NA	0	10.52

Prior to 2011	NA	NA	0	37.83

C. Early breast cancer

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	NA	99.11	98.85	98.98
	Male	NA	0.89	1.15	1.02
Age at diagnosis	0-19	NA	0.01	0.04	0.03
	20-34	NA	1.8	1.29	1.29
	35-49	NA	16.64	16.51	16.70
	50-64	NA	36.59	37.55	36.27
	65-79	NA	33.75	36.97	37.56
	80+	NA	11.22	7.63	8.15
Race	Asian	NA	4.00	3.09	3.23
	Black	NA	11.74	10.20	9.92
	White	NA	83.24	75.25	74.43
	Other	NA	1.02	11.45	12.42
Region	Midwest	NA	22.06	15.62	15.45
	Northeast	NA	19.94	25.45	25.40
	South	NA	36.38	36.40	37.05
	West	NA	21.62	20.24	19.84
	Other	NA	NA	2.29	2.26
Metastatic at diagnosis	No	NA	93.94	100.00	100
	Yes	NA	6.06	NA	NA
Stage	Stage I	NA	NA	52.80	54.91
	Stage II	NA	NA	34.76	33.54
	Stage III	NA	NA	12.43	11.55
Year of diagnosis	2011	NA	15.89	15.98	11.88
	2012	NA	16.15	15.05	11.19
	2013	NA	16.60	16.87	12.54
	2014	NA	16.86	17.89	13.30
	2015	NA	17.24	17.04	12.67
	2016	NA	17.26	17.18	12.77
	2017 - 2019	NA	NA	0	25.64

D. Chronic lymphocytic leukemia

		SEER %	NPCR, %	Flatiron Health, %	
				EDM 2011-2016	EDM all time
Age at diagnosis	0-19	0.05	0.05	0.00	0.00
	20-34	0.26	0.28	0.21	0.31
	35-49	4.20	4.19	5.11	7.61
	50-64	28.68	28.48	31.17	36.72
	65-79	43.42	44.33	49.03	46.16
	80+	23.39	22.68	14.48	9.20
Sex	Female	38.81	38.76	34.91	37.68
	Male	61.19	61.24	65.09	62.32
Race	Asian	2.25	1.09	0.97	0.74
	Black	7.78	7.16	8.50	7.34
	White	89.63	91.00	80.97	82.21
	Other	0.34	0.75	9.55	9.71
Region	Midwest	11.12	21.61	15.93	15.73
	Northeast	18.50	21.98	24.30	25.55
	South	23.99	37.92	39.09	37.66
	West	46.38	18.49	19.77	20.37
	Other	NA	NA	0.91	0.70
Stage	0	NA	NA	34.78	40.23
	I	NA	NA	24.00	23.24
	II	NA	NA	12.17	10.77
	III	NA	NA	10.20	8.87
	IV	NA	NA	18.85	16.89
Metastatic at diagnosis	Yes	NA	100.00	100	100.00
	No	NA	0	NA	NA
Year of diagnosis	2011	NA	0	16.62	7.12
	2012	NA	0	16.60	7.11
	2013	25.20	25.50	17.53	7.51
	2014	25.46	25.62	17.01	7.29
	2015	25.36	25.72	17.11	7.33
	2016	23.98	23.16	15.13	6.49
	2017 - 2019	NA	NA	0	9.18
	Prior to 2011	NA	NA	0	47.96

E. Metastatic colorectal cancer

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Age at diagnosis	0-19	0.26	0.22	0.05	0.05
	20-34	1.76	1.62	1.65	1.66
	35-49	9.75	9.48	12.78	13.15
	50-64	33.61	32.86	35.77	35.78
	65-79	34.80	36.12	37.77	37.74
	80+	19.81	19.69	11.98	11.64
Sex	Female	48.22	48.23	45.28	44.93
	Male	51.78	51.77	54.72	55.07
Race	Asian	8.36	3.77	3.04	3.08
	Black	12.33	12.38	11.41	11.36
	White	78.47	82.67	72.74	72.30
	Other	0.84	1.18	12.81	13.25
Region	Midwest	9.55	22.77	13.94	13.81
	Northeast	15.72	18.91	23.95	23.27
	South	24.60	38.13	38.81	39.58
	West	50.12	20.20	20.44	20.46
	Other	NA	NA	2.85	2.89
Stage	0	2.82	NA	0	0.01
	I	24.63	NA	2.74	2.79
	II	24.32	NA	12.56	11.51
	III	26.88	NA	28.63	25.42
	IV	21.34	NA	56.07	60.26
Metastatic at diagnosis	Yes	21.99	22.43	56.07	60.26
	No	78.01	77.57	43.93	39.74
Year of diagnosis	2011	NA	0	4.84	3.34
	2012	NA	0	8.31	5.75
	2013	24.35	24.66	21.91	15.15
	2014	25.27	25.13%	22.57	15.61
	2015	25.22	25.19	21.99	15.20
	2016	25.16	25.02	20.39	14.10
	2017 - 2019	NA	NA	0	24.62
	Prior to 2011	NA	NA	0	6.220

F. Diffuse large B-cell lymphoma (DLBCL)

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	44.46	44.86	45.83	45.27
	Male	55.54	55.14	54.17	54.73
Age at diagnosis	0-19	0.92	0.87	0.30	0.25
	20-34	3.65	3.50	3.96	3.87
	35-49	9.26	8.85	8.55	8.13
	50-64	27.49	26.73	28.38	27.18
	65-79	37.82	39.05	41.84	42.27
	80+	20.85	20.99	16.97	18.31
Race	Asian	8.79	4.00	2.66	2.64
	Black	7.18	7.48	6.34	6.59
	White	83.44	87.49	80.14	79.09
	Other	0.59	1.02	10.86	11.69
Region	Midwest	10.07	23.39	13.21	12.79
	Northeast	15.28	19.84	29.07	29.25
	South	20.40	34.96	37.70	38.29
	West	54.26	21.8%0	18.88	18.42
	Other	NA	NA	1.14	1.25
Metastatic at diagnosis	Yes	54.11	55.30	61.07	62.65
	No	45.89	44.70	38.93	37.35
Stage	Stage I	26.14	NA	17.33	16.21
	Stage II	19.62	NA	21.60	21.14
	Stage III	18.58	NA	23.89	24.8
	Stage IV	35.66	NA	37.18	37.85
Year of diagnosis	2011	15.86	16.04	13.34	9.40
	2012	16.56	16.35	15.44	10.89
	2013	16.45	16.74	16.75	11.80
	2014	16.96	16.97	16.95	11.95
	2015	17.08	17.10	18.68	13.16
	2016	17.08	16.80	18.85	13.29
	2017 - 2019	NA	NA	0	29.51

G. Follicular lymphoma

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	48.85	49.79	50.62	49.91
	Male	51.15	50.21	49.38	50.09
Age at diagnosis	0-19	0.29	0.28	0.25	0.18
	20-34	1.68	1.61	1.49	1.25
	35-49	12.05	11.09	10.20	10.04
	50-64	35.47	34.29	35.57	33.51
	65-79	37.28	38.86	40.92	41.85
	80+	13.23	13.88	11.57	13.17
Race	Asian	5.28	2.40	1.23	1.58
	Black	4.77	4.93	4.50	4.35
	White	89.47	91.73	83.79	81.60
	Other	0.47	0.95	10.49	12.46
Region	Midwest	9.71	23.35	14.99	14.78
	Northeast	16.45	19.84	27.71	26.75
	South	21.06%	36.77	35.89	36.99
	West	52.78	20.04	20.15	20.04
	Other	NA	NA	1.26	1.45
Metastatic at diagnosis	Yes	56.19	55.30	70.04	70.35
	No	43.81	44.7%0	29.96	29.65
Stage	Stage I	27.40	NA	17.35	17.07
	Stage II	15.85	NA	12.61	12.58
	Stage III	28.01	NA	30.71	30.82
	Stage IV	28.75	NA	39.33	39.53
Year of diagnosis	2011	16.25	16.89	15.17	10.93
	2012	16.19	16.33	13.56	9.77
	2013	16.12	16.72	16.67	12.01
	2014	17.22	16.88	16.92	12.19
	2015	16.92	16.79	19.90	14.34
	2016	17.30	16.40	17.79	12.81
	2017 - 2019	NA	NA	0	27.96

H. Advanced gastric/esophageal carcinoma

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	33.47	31.74	26.55	26.51
	Male	66.53	68.26	73.45	73.49
Age at diagnosis	0-19	0.08	0.07	0.00	0.00
	20-34	1.25	1.09	0.89	0.83
	35-49	7.55	7.19	6.80	6.58
	50-64	30.9	31.58	32.16	31.31
	65-79	39.88	40.89	43.54	43.96
	80+	20.35	19.19	16.62	17.33
Race	Asian	10.45	5.13	3.42	3.70
	Black	12.34	13.00	8.26	8.73
	White	76.38	80.66	75.55	74.1
	Other	0.84	1.20	12.77	13.48
Region	Midwest	8.91	21.93	13.50	13.40
	Northeast	17.2	21.09	26.88	26.53
	South	21.55	35.99	35.67	35.92
	West	52.34	20.98	22.38	22.46
	Other	NA	NA	1.57	1.69
Metastatic at diagnosis	Yes	39.14	38.36	49.47	50.87
	No	60.86	61.64	50.53	49.13
Stage	Stage 0	0.30	NA	NA	NA
	Stage I	23.70	NA	5.01	5.32
	Stage II	16.48	NA	15.18	14.93
	Stage III	19.34	NA	23.95	22.52
	Stage IV	40.18	NA	55.87	57.24
Year of diagnosis	2011	16.10	16.19	11.39	7.97
	2012	16.53	16.31	14.53	10.17
	2013	16.48	16.61	16.78	11.74
	2014	16.96	16.88	19.09	13.36
	2015	17.12	16.99	18.95	13.26
	2016	16.82	17.02	19.27	13.49
	2017 - 2019	NA	NA	0	27.29
	Prior to 2011	NA	NA	0	2.72

I. Hepatocellular carcinoma (HCC)

		SEER, %	NPCR, %	Flatiron Health, %	
				EDM 2011-2016	2011-2019
Sex	Female	23.21	22.74	24.06	23.51
	Male	76.79	77.26	75.94	76.49
Age at diagnosis	0-19	0.17	0.19	0.00	0.00
	20-34	0.59	0.62	0.55	0.45
	35-49	4.52	4.49)	3.39	3.02
	50-64	49.45	50.53	44.17	41.74
	65-79	35.29	34.57	40.71	43.77
	80+	9.98	9.59	11.19	11.02
Race	Asian	14.81	7.87	5.61	5.41
	Black	14.04%	16.10	12.33	12.48
	White	69.74%	74.09	66.32	66.06
	Other	1.41	1.93	15.74	16.04
Region	Midwest	7.15	17.38	7.30	8.03
	Northeast	12.61	18.55	31.89	31.37
	South	20.10	38.88	35.85	36.56
	West	60.14	25.19	23.51	22.50
	Other	NA	NA	1.45	1.55
Metastatic at diagnosis	Yes	15.22	15.95	25.27	27.46
	No	84.78	84.05	74.73	72.54
Stage	Stage I	40.76	NA	21.39	18.68
	Stage II	20.54	NA	14.58	14.55
	Stage III	19.13	NA	22.58	21.78
	Stage IV	19.57	NA	41.46	44.99
Year of diagnosis	2011	15.10	14.79	11.01	7.54
	2012	16.01	15.82	15.16	10.39
	2013	16.70	16.49	16.07	11.02
	2014	17.35	17.39	16.98	11.64
	2015	17.73	17.89	20.01	13.71
	2016	17.12	17.64	20.77	14.24
	2017 - 2019	NA	NA	0	31.45

J. Advanced head and neck cancer

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	26.24	26.20	23.07	23.31
	Male	73.76	73.80	76.93	76.69
Age at diagnosis	0-19	0.32	0.31	0.00	0.00
	20-34	1.28	1.18	0.47	0.64
	35-49	9.43	9.38	7.40	7.81
	50-64	42.81	43.76	46.18	46.42
	65-79	34.56	34.78	37.59	37.28
	80+	11.60	10.60	8.36	7.85
Race	Asian	6.13	2.92	1.77	1.58
	Black	10.10	9.84%	7.34	7.69
	White	83.10	86.27	79.80	79.32
	Other	0.68	0.97	11.08	11.41
Region	Midwest	9.99	22.82	11.58	11.52
	Northeast	15.15	18.21	19.60	19.35
	South	27.22	40.29	51.51	51.92
	West	47.64	18.68	14.90	15.07
	Other	NA	NA	2.41	2.13
Metastatic at diagnosis	Yes	15.80	17.69	71.35	71.25
	No	84.20	82.31	28.65	28.75
Stage	Stage I	24.08	NA	6.09	6.54
	Stage II	12.76	NA	7.72	7.72
	Stage III	15.74	NA	14.84	14.45
	Stage IV	47.42	NA	71.35	71.25
	Stage 0	NA	NA	0	0.04
Year of diagnosis	2011	15.85	15.92	12.75	9.05
	2012	16.03	16.03	16.05	11.40
	2013	16.62	16.64	16.25	11.54
	2014	17.02	16.95	17.65	12.53
	2015	17.23	17.21	17.85	12.68
	2016	17.25	17.23	19.45	13.81
	2017 - 2019	NA	NA	0	17.07
	Prior to 2011	NA	NA	0	11.93

K. Advanced melanoma

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	40.73	41.04	33.57	34.93
	Male	59.27	58.96	66.43	65.07
Age at diagnosis	0-19	0.43	0.43	0.31	0.36
	20-34	5.42	5.40	5.05	5.42
	35-49	13.64	13.67	11.93	13.54
	50-64	31.24	30.81	30.93	30.84
	65-79	33.18	33.70	37.72	36.62
	80+	16.09	16.00	14.05	13.22
Race	Asian	0.66	0.32	0.28	0.28
	Black	0.46	0.50	0.47	0.45
	White	98.64	98.72	92.91	92.13
	Other	0.25	0.45	6.34	7.14
Region	Midwest	7.74	21.74	14.10	14.15
	Northeast	14.43	18.49	22.21	22.05
	South	22.10	35.85	33.44	34.68
	West	55.74	23.92	29.82	28.67
	Other	NA	NA	0.44	0.45
Metastatic at diagnosis	Yes	4.51	5.50	23.27	23.62
	No	95.49	94.50	76.73	76.38
Stage	Stage 0	0	NA	0.31	0.37
	Stage I	75.38	NA	7.88	9.60
	Stage II	12.97	NA	16.98	17.16
	Stage III	7.45	NA	50.34	47.46
	Stage IV	4.20	NA	24.48	25.41
Year of diagnosis	2011	14.79	15.13	13.91	8.87
	2012	15.67	15.46	15.46	9.86
	2013	16.25	16.23	16.95	10.81
	2014	17.55	17.18	17.83	11.37
	2015	17.99	17.96	18.32	11.69
	2016	17.74	18.04	17.54	11.18
	2017 - 2019	NA	NA	0	19.72
	Prior to 2011	NA	NA	0	16.50

L. Malignant pleural mesothelioma

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	25.01	24.75	21.17	22.00
	Male	74.99	75.25	78.83	78.00
Age at diagnosis	0-19	0.02	0.10	0.00	0.00
	20-34	1.01	0.87	0.14	0.10
	35-49	4.11	3.65	1.29	1.33
	50-64	18.02	17.80	15.33	14.37
	65-79	45.41	47.31	58.14	57.82
	80+	31.43	30.27	25.10	26.38
Race	Asian	3.63	1.58	0.84	0.94
	Black	5.19	4.83	3.19%	2.98
	White	90.62	92.90	87.77	87.31
	Other	0.56	0.69	8.21	8.77
Region	Midwest	8.76	23.64	11.15	10.87
	Northeast	19.29	21.76	38.95	37.64
	South	19.11	32.80	36.55	38.09
	West	52.84	21.81	13.01	13.01
	Other	NA	NA	0.34	0.40
Metastatic at diagnosis	Yes	NA	69.80	14.59	16.34
	No	NA	30.20	85.41	83.66
Stage	Stage I	21.84	NA	8.08%	9.42
	Stage II	10.47	NA	10.77	9.50
	Stage III	22.57	NA	32.07	30.86
	Stage IV	45.13	NA	49.08	50.22
Year of diagnosis	2011	16.85	17.12	15.54	11.27
	2012	16.51	16.83	15.26	11.07
	2013	16.75	16.79	16.49	11.96
	2014	16.59	16.84	17.44	12.65
	2015	17.60	16.64	16.89	12.25
	2016	15.71	15.79	18.39	13.34
	2017 - 2019	NA	NA	0	27.46

M. Multiple myeloma

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	43.80	44.08	46.98	46.06
	Male	56.20	55.92	53.02	53.94
Age at diagnosis	0-19	0.03	0.02	0.00	0.00
	20-34	0.50	0.46	0.39	0.35
	35-49	6.52	6.40	5.50	5.38
	50-64	30.00	30.02	30.75	29.61
	65-79	43.25	44.12	45.91	46.95
	80+	19.70	18.98	17.45	17.71
Race	Asian	5.73	2.62	1.95	1.95
	Black	20.37	20.62	17.66	17.97
	White	73.26	75.70	68.62	67.7
	Other	0.64	1.05	11.77	12.37
Region	Midwest	10.00	21.08	12.90	12.76
	Northeast	16.67	20.00	28.40	28.00
	South	25.12	39.99	37.82	38.41
	West	48.21	18.93	19.29	19.22
	Other	NA	NA	1.58	1.61
Metastatic at diagnosis	Yes	95.58	95.41	100.00	100.00
	No	4.42	4.59	NA	NA
Stage	Stage I	NA	NA	34.58	34.22
	Stage II	NA	NA	33.03	33.06
	Stage III	NA	NA	32.38	32.73
Year of diagnosis	2011	15.51	15.67	12.79	9.14
	2012	16.19	16.09	15.03	10.74
	2013	16.46	16.67	16.53	11.81
	2014	16.69	17.04	16.82	12.02
	2015	17.42	17.38	18.90	13.50
	2016	17.73	17.15	19.94	14.24
	2017 - 2019	NA	NA	0	28.56

N. Advanced NSCLC

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	47.28	46.56	47.42	47.40
	Male	52.72	53.44	52.58	52.60
Age at diagnosis	0-19	0.01	0.01	0.01	0.01
	20-34	0.17	0.14	0.20	0.19
	35-49	3.12	3.18	3.69	3.59
	50-64	27.80	28.75	30.63	30.47
	65-79	50.26	50.6	51.01	51.18
	80+	18.65	17.32	14.47	14.56
Race	Asian	7.13	2.96	2.72	2.82
	Black	11.84	11.30	9.31	9.34
	White	80.52	84.96	78.11	77.72
	Other	0.51	0.78	9.87	10.11
Region	Midwest	10.97	24.06	15.27	15.19
	Northeast	17.18	19.72	27.43	27.46
	South	27.93	39.43	39.46	39.72
	West	43.91	16.78	17.16	16.98
	Other	NA	NA	0.68	0.64
Metastatic at diagnosis	Yes	52.03	49.75	64.18	64.43
	No	47.97	50.25	35.82	35.57
Stage	Stage I	24.81	NA	8.50	9.08
	Stage II	7.46	NA	5.49	5.39
	Stage III	19.01	NA	21.81	21.10
	Stage IV	48.73	NA	64.18	64.43
	Stage 0	NA	NA	0.01	0.01
Year of diagnosis	2011	16.51	16.33	12.36	8.63
	2012	16.59	16.50	15.20	10.61
	2013	16.61	16.64	17.17	11.99
	2014	16.72	16.80	18.17	12.69
	2015	16.82	16.93	18.91	13.21
	2016	16.74	16.80	18.20	12.71
	2017 - 2019	NA	NA	0	24.67
	Prior to 2011	NA	NA	0	5.47

O. Ovarian carcinoma

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	100.00	100.00	100.00	100.00
	Male	NA	0	NA	NA
Age at diagnosis	0-19	1.30	1.26	0.23	0.17
	20-34	3.56	3.45	1.98	1.94
	35-49	13.15	12.55	11.59	11.26
	50-64	35.13	34.86	36.89	36.14
	65-79	32.62	34.15	38.91	39.46
	80+	14.25	13.74)	10.40	11.01
Race	Asian	8.73	4.26	2.67	2.53
	Black	8.95	9.18	6.20	6.16
	White	81.62	85.39	78.84	78.51
	Other	0.69	1.17	12.29	12.80
Region	Midwest	9.15	21.86	13.37	12.68
	Northeast	16.39	20.23	22.00	21.52
	South	19.74	35.53	41.30	42.13
	West	54.72	22.38	21.63	21.69
	Other	NA	NA	1.72	1.98
Metastatic at diagnosis	Yes	60.80	57.35	21.38	22.84
	No	39.20	42.65	78.62	77.16
Stage	Stage I	23.50	NA	20.13	19.63
	Stage II	8.73	NA	8.91	9.02
	Stage III	39.02	NA	48.69	47.65
	Stage IV	28.75	NA	22.26	23.70
Year of diagnosis	2011	16.40	16.61	14.81	10.90
	2012	16.58	16.67	16.70	12.30
	2013	16.58	16.68	16.66	12.27
	2014	16.87	16.78	17.17	12.64
	2015	17.12	17.00	16.70	12.30
	2016	16.44	16.26	17.97	13.24
	2017 - 2019	NA	NA	0	26.35

P. Metastatic pancreatic carcinoma

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	48.94	48.13	46.88	46.36
	Male	51.06	51.87	53.12	53.64
Age at diagnosis	0-19	0.13	0.12	0.00	0.00
	20-34	0.68	0.60	0.28	0.26
	35-49	4.55	4.72	3.94	3.87
	50-64	27.81	28.79	32.67	31.72
	65-79	43.03	44.34	50.77	51.49
	80+	23.80	21.44	12.34	12.66
Race	Asian	7.70	3.38	1.82	1.82
	Black	12.12	12.60	9.02	9.34
	White	79.51	83.05	79.29	77.81
	Other	0.67	0.97	9.86	11.03
Region	Midwest	9.88	22.11	12.18	12.05
	Northeast	17.03	20.15	30.45	29.58
	South	22.64	37.18	39.90	40.85
	West	50.45	20.56	16.53	16.66
	Other	NA	NA	0.93	0.85
Metastatic at diagnosis	Metastatic	54.43	53.34	64.99	71.74
	Not Metastatic	45.57	46.66	35.01	28.26
Stage	Stage 0	0	NA	NA	NA
	Stage I	12.23	NA	4.11	3.57
	Stage II	25.64	NA	23.32	18.10
	Stage III	9.05	NA	7.59	6.58
	Stage IV	53.08	NA	64.99	71.74
Year of diagnosis	2011	NA	0	0.57	0.34
	2012	NA	0	2.22	1.34
	2013	NA	0	7.44	4.50
	2014	32.84	32.62	27.57	16.67
	2015	33.76	33.55	31.19	18.85
	2016	33.4	33.83	31.01	18.75
	2017 - 2019	NA	NA	0	38.85
	Prior to 2011	NA	NA	0	0.70

Q. Metastatic prostate cancer

		SEER	NPCR	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Male	100.00	100.00	100.00	100.00
	Female	NA	0	NA	NA
Age at diagnosis	0-19	0	0	0.00	0.00
	20-34	0.01	0.01	0.00	0.01
	35-49	2.22	2.19	1.64	1.90
	50-64	38.73	39.22	26.79	34.08
	65-79	49.77	50.22	51.70	50.23
	80+	9.27	8.36	19.88	13.78
Race	Asian	5.09	2.32	1.85	1.66
	Black	16.43	16.76	11.70	10.65
	White	78.04	79.88	74.54	75.88
	Other	0.43	1.04	11.91	11.81
Region	Midwest	9.55	21.95	13.50	12.92
	Northeast	17.70	20.21	20.30	21.24
	South	24.40	37.99	43.47	43.58
	West	48.34	19.84	21.24	21.00
	Other	NA	NA	1.48	1.25
Metastatic at diagnosis	Yes	7.69	7.56	74.24	59.70
	No	92.31	92.44	25.76	40.30
Stage	Stage I	6.26	NA	0.62	1.32
	Stage II	72.41	NA	5.51	9.25
	Stage III	9.69	NA	3.48	5.17
	Stage IV	11.63	NA	90.38	84.25
Year of diagnosis	2011	NA	0	7.87	3.65
	2012	NA	0	8.14	3.78
	2013	25.16	24.83	21.00	9.74
	2014	23.62	23.94	21.22	9.84
	2015	25.11	25.43	20.45	9.49
	2016	26.11	25.80	21.32	9.89
	2017 - 2019	NA	NA	0	16.59
	Prior to 2011	NA	NA	0	37.03

R. Metastatic renal-cell carcinoma (RCC)

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	36.25	36.88	30.69	31.40
	Male	63.75	63.12	69.31	68.60
Age at diagnosis	0-19	1.16	1.05	0.00	0.04
	20-34	2.01	1.85	0.67	0.77
	35-49	12.58	11.86	7.63	9.08
	50-64	36.85	36.03	39.11	40.75
	65-79	36.42	37.76	43.15	41.19
	80+	10.98	11.45	9.44	8.17
Race	Asian	5.31	2.31	1.83	1.63
	Black	12.68	12.11	7.90	7.56
	White	81.02	84.15	76.23	76.81
	Other	0.99	1.44	14.04	14.00
Region	Midwest	9.81	22.96	16.56	15.64
	Northeast	15.17	18.44	20.20	20.46
	South	25.44	38.77	39.38	40.30
	West	49.58	19.82	22.66	22.41
	Other	NA	NA	1.21	1.19
Metastatic at diagnosis	Yes	14.61	14.68	60.21	53.60
	No	85.39	85.32	39.79	46.40
Stage	Stage I	62.56	NA	NA	NA
	Stage II	7.81	NA	NA	NA
	Stage III	13.72%	NA	NA	NA
	Stage IV	15.91	NA	62.24	55.12
	Unspecified Non-metastatic	NA	NA	37.76	44.88
Year of diagnosis	2011	15.42	15.44	14.12	8.70
	2012	16.03	15.97	15.01	9.25
	2013	16.43	16.37	16.94	10.43
	2014	16.89	16.95	18.14	11.18
	2015	17.65	17.52	18.01	11.09
	2016	17.58	17.75	17.79	10.95
	2017 - 2019	NA	NA	0	18.53
	Prior to 2011	NA	NA	0	19.87

S. Small-cell lung cancer (SCLC).

		SEER, %	NPCR, %	Flatiron Health, %	
				2011-2016	2011-2019
Sex	Female	50.55	50.94	51.61	52.48
	Male	49.45	49.06	48.39	47.52
Age at diagnosis	0-19	0.01	0	0.00	0.00
	20-34	0.07	0.06	0.05	0.10
	35-49	2.59	2.75	2.76	2.60
	50-64	33.05	33.92	35.62	34.63
	65-79	51.29	51.26	51.69	52.57
	80+	13.00	12.00	9.88	10.10
Race	Asian	3.65	1.39	0.93	0.96
	Black	8.98	7.86	6.08	6.23
	White	86.76	89.92	84.00	83.19
	Other	0.62	0.83	8.99	9.61
Region	Midwest	13.23	26.71	18.11	17.48
	Northeast	14.94	17.35	24.35	23.53
	South	35.67	41.98	43.76	44.52
	West	36.16	13.96	13.42	14.11
	Other	NA	NA	0.36	0.36%
Metastatic at diagnosis	Metastatic	74.60	72.14	71.75	70.86
	Not Metastatic	25.40	27.86	28.25	29.14
Stage	Stage I	4.17	NA	4.07	4.27
	Stage II	3.04	NA	4.44	4.42
	Stage III	22.38	NA	19.75	20.45
	Stage IV	70.40	NA	71.75	70.86
Year of diagnosis	2013	25.12	25.08	23.34	14.77
	2014	25.75	25.07	24.90	15.76
	2015	25.33	25.39	25.31	16.01
	2016	23.80	24.45	26.46	16.74
	2017 - 2019	NA	NA	0	36.72

APPENDIX II: APRIL 2023 UPDATE

Differences in age distribution between the three databases were previously noted in this study, such that a lower proportion of patients over 80 years of age at diagnosis was seen in the Flatiron Health database as compared with SEER and NPCR sources. Of note, birth year data for elderly patients in this analysis was subject to a standard algorithmic transformation in order to mitigate their higher risk of patient re-identification. In December 2022, Flatiron Health instituted an improvement to this algorithm that reduced the extent of birth year transformation and, hence, calculated age at diagnosis. The improved algorithm enables the reporting of the true birth year for a greater share (>90%) of patients than in the original algorithm. Using the improved algorithm with patient records available in the Flatiron Health databases for analysis as of May 31, 2019 and December 31, 2016 results in changes to the distributions previously reported in patients' age at initial diagnosis. These changes are incorporated into Tables 2 through 20 and A3.A through A3.S herein (previously reported distributions are available in older versions of this paper).

For most disease-specific databases, the updated birth years resulted in an increase of patient records showing a diagnosis at 80 years or greater. Most updated distributions match their counterpart SEER and/or NPCR distributions to within 5% for each age category. The greatest changes appear among patients with malignant pleural mesothelioma (MPM), with, for example, 25.1% of the MPM patients available for analysis with an initial diagnosis from January 2011 to December 2016 showing an age of 80 years or greater at diagnosis, versus 14.65% prior to the update. These findings suggest that similarity in age distribution between Flatiron Health, SEER, and NPCR is greater than initially reported, as differences seen in the original dataset were largely related to the prior approach to algorithmic masking of birth year for elderly patients.