OVERVIEW OF PREVALENCE STATISTICS AND THE SEER*STAT PREVALENCE SESSION

AGENDA

- An overview of cancer prevalence statistics
- Limited-duration prevalence using SEER*Stat
  - Demo SEER*Stat
- Estimating United States prevalence counts
  - Demo ProjPrev and Excel
- Other related methods and applications
  - COMPREV (completeness index method) to estimate complete prevalence
  - MIAMOD/PIAMOD- to estimate and project complete prevalence

PREVALENCE DEFINITION

- Cancer prevalence is the proportion (or number) of people diagnosed with cancer and alive at a given point in time t

$$\text{Prevalence}(t) = \frac{\#\ \text{Prevalent cases}\ (t)}{\text{Population}(t)}$$

- Includes new and pre-existing cases alive on a certain date
- Prevalence is reported as counts (numbers)

IMPORTANCE AND USES

- Quantifies the burden of cancer
- Provides estimates of the cancer survivorship
- Useful in planning, health care utilization and costs of care
- Major users in the US:
  - NCI Director and Congress to justify investments in cancer research
  - Policy Makers and Cancer Advocacy Groups
  - Cancer survivorship community
  - FDA Office of Orphan Drug Products Development
    - Cancer sites with less than 200,000 survivors

PREVALENCE FROM SURVEY DATA

- In the past prevalence was estimated from cross-sectional survey data by direct counting # prevalent cases (after considering missing responses)
- Surveys included specific questions about cancer survivorship which allowed estimation of different types of cancer prevalence
  - Whether survivors were in different phases of treatment/care: Care prevalence
  - Any person diagnosed with cancer in the household : % households with cancer survivors

PREVALENCE FROM SURVEY DATA

- Limitations:
  - Based on self-report (accuracy of clinical information)
  - Difficult to estimate rare cancers (sample size)
  - Under sample of people receiving care (in medical facilities)
- Source: National Health Interview Survey (NHIS)
  - Byrne, Kessler, Devesa (1992)
  - Hewitt, Breen, Devesa (1999)
  - Summary Health Statistics U.S. Adults: NHIS 1998

AFTER 2000 US PREVALENCE BEGAN TO BE ESTIMATED FROM CANCER REGISTRY DATA

- Registry data is well suited for estimation of cancer prevalence as it contains information on cancer incidence (diagnosis of cancer) and follow-up information on life status

👍 Clinical information from hospital records (not-self reported)

👍 Population covered not a random sample of the US population

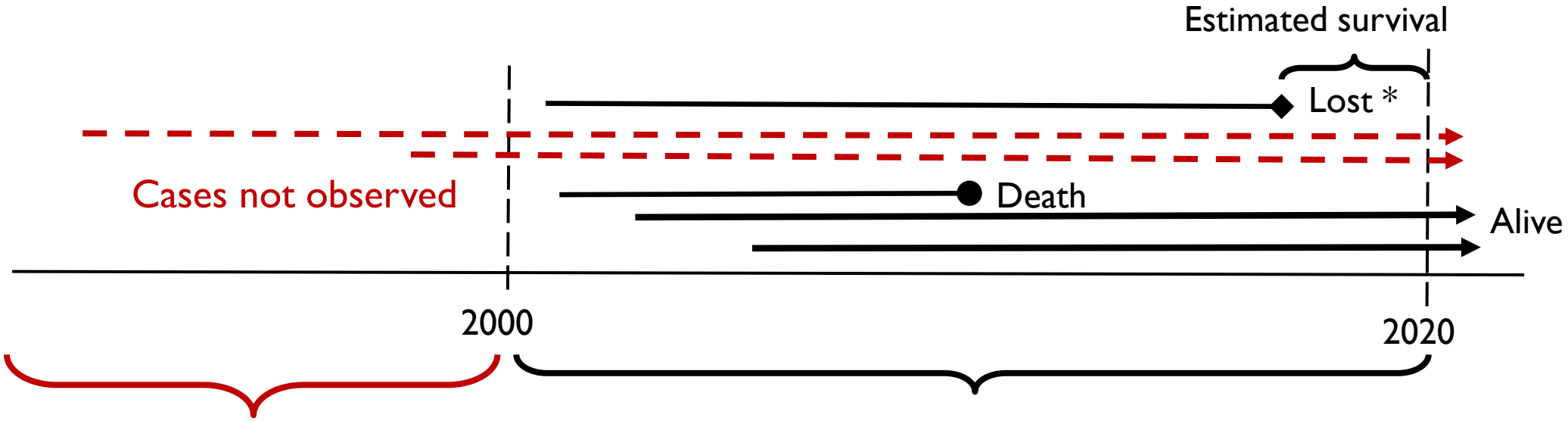
👎 Only life-status information and no other follow-up information (e.g. recurrence, in care, …)

COMPLETE PREVALENCE VS. LIMITED DURATION PREVALENCE

- **Complete prevalence**: living persons ever diagnosed with cancer
  - Overall measure of cancer burden and survivorship
  - Requires modeling (registries do not include survivors diagnosed prior to 1975)
- **Limited duration prevalence** - prevalence of persons diagnosed with cancer in the last x years.
  - E.g. 0-5 years, 5-10 years, etc.
  - For example, most cancer treatment generally takes place within the first 5 years from diagnosis[1]
  - Easier to estimate from cancer registry data

[1] *Pisani et al. (2002) Int J Cancer*

LIMITED DURATION VS. COMPLETE PREVALENCE METHODS



The Completeness Index Method

Complete Prevalence

COMPREV

Counting Method:

20-year limited duration (LD) prevalence at 1/1/2020

SEERStat

SEER*STAT AND THE COUNTING METHOD

- Provides prevalence by duration: 5-year, 10-year, etc.. Up to the maximum duration of the registry data
- Counts the number of people diagnosed with cancer in the registry that are alive at the prevalence date
- For people lost to follow-up estimate their probability of being alive at the prevalence date
- In SEER*Stat users need to define:
  - How persons with multiple tumors are counted
  - Variables or cohorts for the estimation of survival for those lost to follow-up

ESTIMATING PREVALENCE USING SEER DATA

- SEER-8 (1975-2020) →45-year LD prevalence (close to complete) (8.3% of the US)
- SEER-12 (1992-2020)→ 28-year LD prevalence (12.4% of the US)
- SEER-22 Limited-Field (2000-2020)→ 20-year LD prevalence (48.0% of the US)

PREVALENCE DATE

- Prevalence is calculated at a point in time, e.g. prevalence on 1/1/2020
- We usually calculate prevalence at the January 1st of the last year with data, e.g. 1/1/2020 (SEER*Stat can calculate prevalence at 7/1/year)
- Use the average of the mid-year populations in 2019 and 2020 as denominator
- It is not a rate because it does not measure occurrences of events in a time interval. It is reported as counts, or proportion, percent of people

TYPE OF FOLLOW-UP

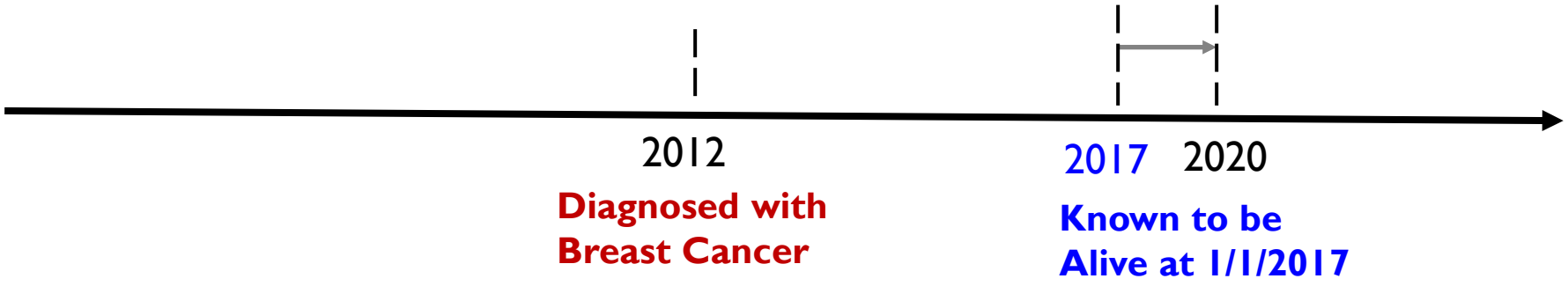
- **Reported alive survival**
- Linkage with mortality databases → Good death ascertainment
- Collects dates of last known to be alive from other linkages (SSA, drivers license, voter registration…)
- Used for most SEER registries and some NPCR registries

- **Presumed alive survival**
- Linkage with mortality databases → Good death ascertainment
- Does not capture date last known alive for alive patients
  - All patients alive are presumed alive on the prevalence date
- Most NPCR registries

The user does not have to worry about choosing a particular method for a particular registry. **This is pre-coded in SEER*Stat**

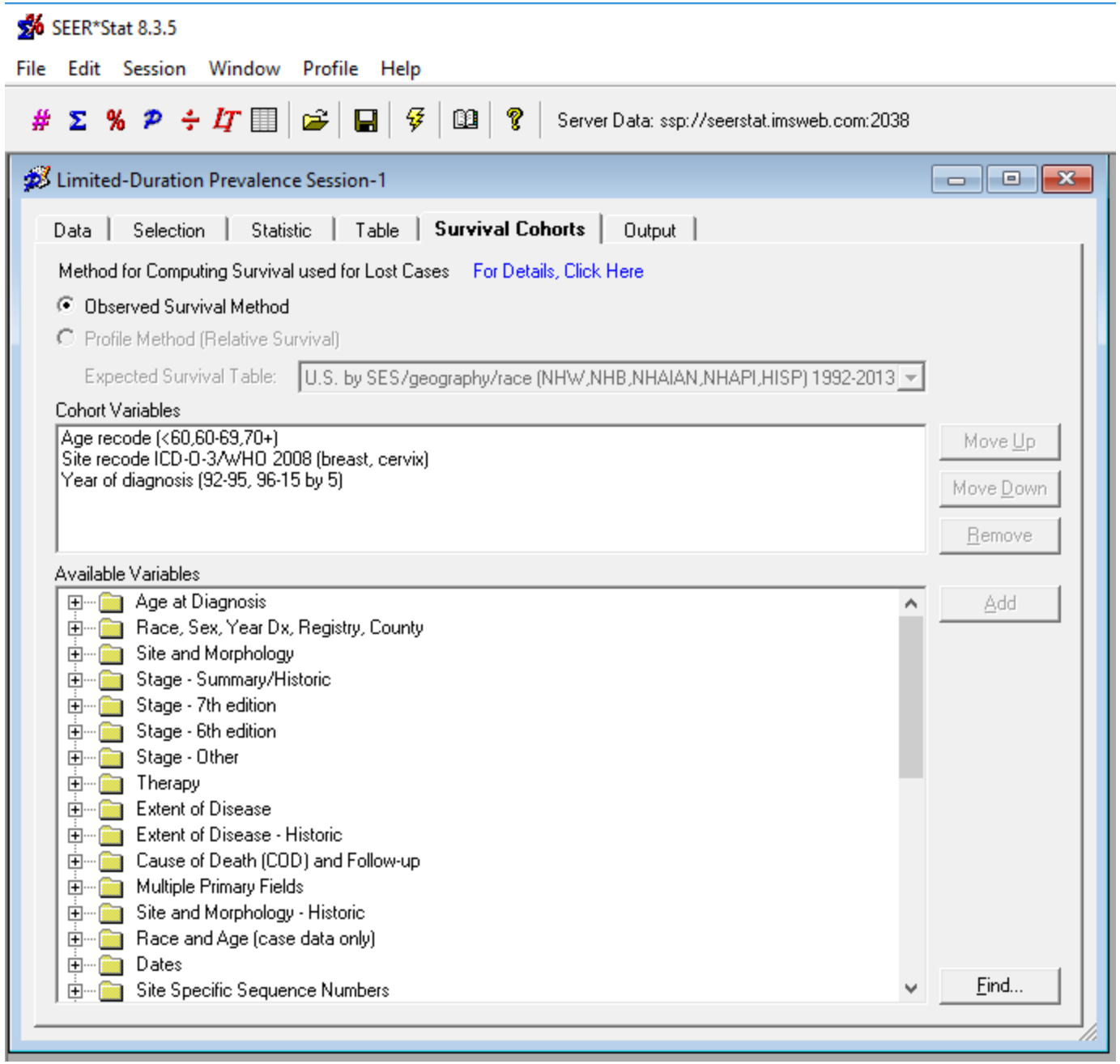
ESTIMATE SURVIVAL FOR PEOPLE LOST TO FOLLOW-UP

- Example: women diagnosed with breast cancer 1/1/2012, alive at 1/1/2017 (date of last contact) (small percent of cases)
- Calculate prevalence at 1/1/2020
- SEER*Stat estimates the conditional probability of surviving 3 years given this person survived 5 years.
- Uses a cohort of similar patients defined by the user



SURVIVAL COHORTS

- Include variables that:
  - Affect survival e.g. age, year of diagnosis, sex, cancer site, etc.
  - Used to display prevalence statistics
  - Include all possible cases and No overlapping groups
- SEER*Stat includes **system-supplied variables**



IMPACT OF DIFFERENT SURVIVAL COHORTS ON 24-YEAR MELANOMA PREVALENCE ESTIMATE

- **In general, estimates do not differ much by different cohort definitions**
- Example: 24-year LD prevalence for Melanoma all races at 1/1/2016 (SEER-12 1992-2016)
  - If no cohort is specified survival is estimated from all cases: 117,133
  - Age (<60, 60-69,70+), sex (males vs females) and race (White/unknown, Black, Other, Unknown): 117,558
  - Difference of 425 cases (less than 0.5%)

COUNTING PERSONS WITH MULTIPLE TUMORS

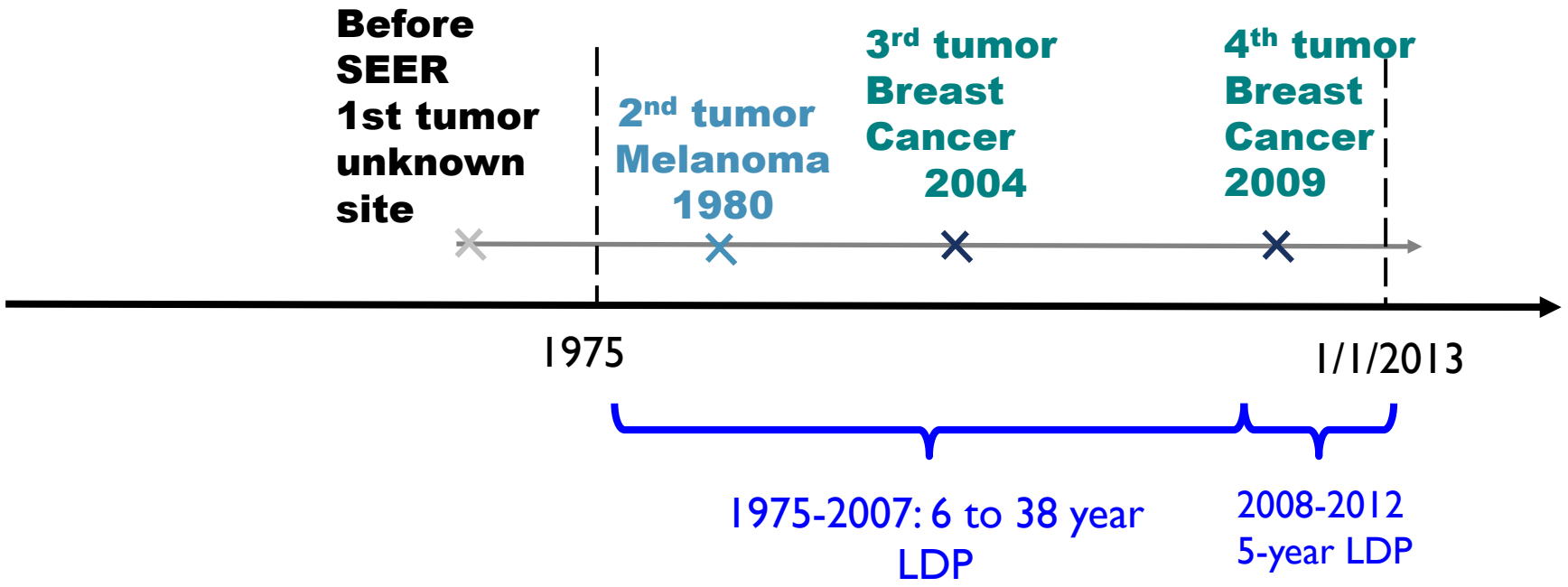
- Prevalence statistics usually represent people and not tumors
- How can we count persons diagnosed with multiple tumors?
- Most used methods:
  - "First Malignant Primary Only (Non-reported Assumed Malignant)": previously used
  - "All Tumors Matching Selection Criteria / One Tumor per Statistic": currently used

COUNTING PERSONS WITH MULTIPLE TUMORS

- First Malignant Primary only (Non-reported Assumed Malignant)
  - This method includes only the first malignant tumor.
  - Advantage: less ambiguity, the prevalence of all sites is the sum of the prevalence by site.
  - Disadvantage: an underestimate of prevalence by cancer site
  - Was the default for reporting of prevalence statistics for many years.
- **All Tumors Matching Selection Criteria is the default in SEER*Stat**
  - Counts the first malignant tumor per prevalence statistic
  - Used in the NCI's Cancer Statistics, SEER*Explorer and Report to the Nation

PREVALENCE CALCULATION FOR A PERSON WITH MULTIPLE TUMORS EXAMPLE

- Women diagnosed with a cancer (unknown) in 1970, with melanoma in 1980 and breast cancer in 2004 and 2009



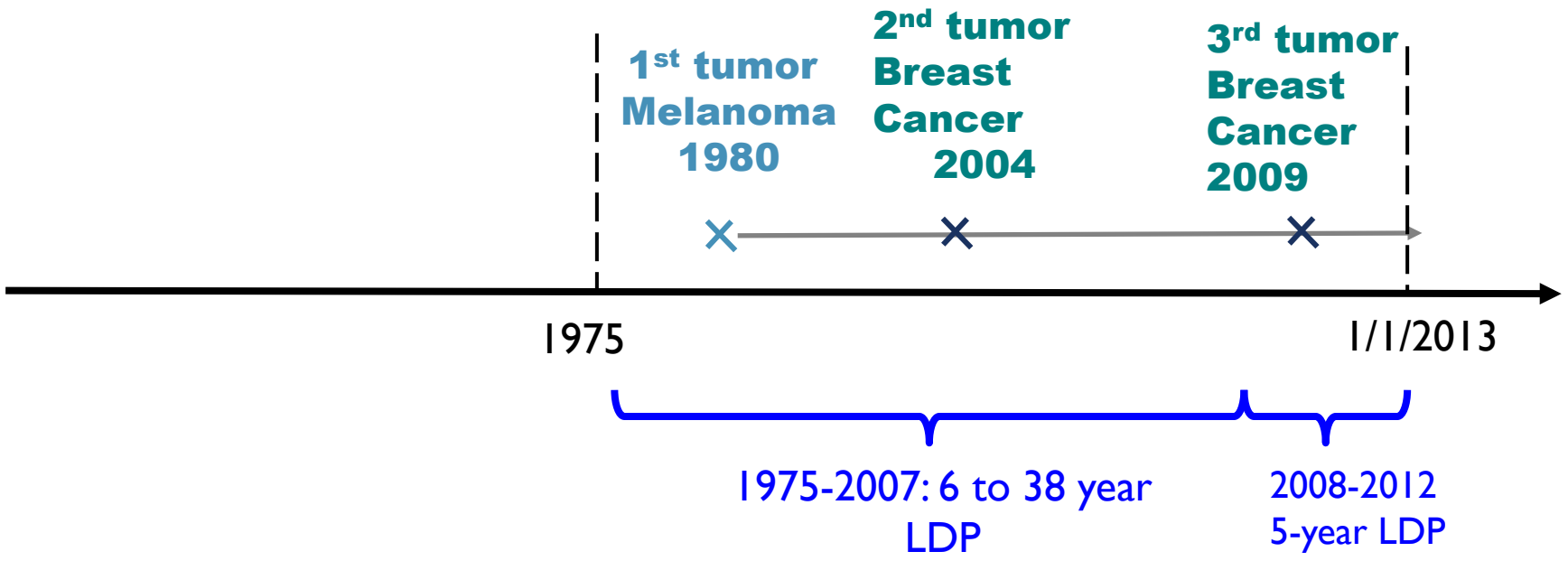
EXAMPLE: WOMEN DIAGNOSED WITH A CANCER IN 1970, WITH MELANOMA IN 1980 AND BREAST CANCER IN 2004 AND 2009

	Tumors contributing to statistics			
	Tumors selected	38-year LD Prevalence		
		All sites	Melanoma	Breast
First Malignant Primary Only	None	None	None	None
All Tumors Matching Selection Criteria	2nd, 3rd and 4th	2nd M (1980)	2nd M (1980)	3rd B (2004)

* One tumor per statistic

PREVALENCE CALCULATION FOR A PERSON WITH MULTIPLE TUMORS EXAMPLE

- Women diagnosed with melanoma in 1980 (1st) and breast cancer in 2004 (2nd) and 2009 (3rd)



EXAMPLE: WOMEN DIAGNOSED WITH MELANOMA IN 1980 (1ST) AND BREAST CANCER IN 2004 (2ND) AND 2009 (3RD)

	Tumors contributing to statistics			
	Tumors selected	38-year LD Prevalence		
		All sites	Melanoma	Breast
First Malignant Primary Only	1st M	1st M	1st M	None
All Tumors Matching Selection Criteria	1st, 2nd, and 3rd	1st M	1st M	2nd B

* One tumor per statistic

EXAMPLE: WOMEN DIAGNOSED WITH MELANOMA IN 1980 (1ST) AND BREAST CANCER IN 2004 (2ND) AND 2009 (3RD)

- One run to estimate both 0-5-year and 6-38 LD prevalence statistics
- Try yourself

	Tumors selected in the run	5-year LD Prevalence			6-37 year LD Prevalence		
		All sites	Melanoma	Breast	All sites	Melanom	Breast
First Primary Only							
All Tumors Matching Selection Criteria*							

* One tumor per statistic

EXAMPLE: WOMEN DIAGNOSED WITH A CANCER IN 1970, WITH MELANOMA IN 1980 AND BREAST CANCER IN 2004 AND 2009

- One run to estimate both 0-5-year and 6-38 LD prevalence statistics

	Tumors selected in the run	5-year LD Prevalence			6-38 year LD Prevalence		
		All sites	Melanoma	Breast	All sites	Melanom	Breast
First Primary Only	1st M	None	None	None	1st M	1st M	None
All Tumors Matching Selection Criteria	1st, 2nd, 3rd	3rd B	None	3rd B	1st M	1st M	2nd B

* One tumor per statistic

SEER*STAT DEMO

- Exercise 1: 28-year LDP for all sites combined by sex and age using SEER 12 database
- Exercise 2a & 2b: 5-year LDP for female breast cancer, by time since diagnosis using SEER 22 (excl IL and MA)
- SEER*Stat and ProjPrev
  - Exercise 3: 16-year LDP by summary stage (at diagnosis) for patients diagnosed with Lung and Bronchus cancers, by race, sex, age projected to the US

ESTIMATING US PREVALENCE

- Usually, people are interested in the number of survivors in the US, in a state or in a health administrative area
- Registries may not cover the whole area

ESTIMATING US PREVALENCE FROM SEER DATA

- For a given sex, race, and cancer site:
  1. Multiply SEER age-specific prevalence by the corresponding US age-specific population:

$$Counts_{US}(age, race, sex) = Prev_{SEER}(age, race, sex) \times Pop_{US}(age, race, sex)$$

  2. Sum US age-specific prevalence counts over all ages and all races

$$Counts_{US}(sex) = \sum_{Age} \sum_{Race} Counts_{US}(age, race)$$

ESTIMATING US PREVALENCE USING SEER DATA

- ProjPrev software: projects prevalence proportions to different populations and the output can be used in COMPREV
- Excel can also be used
- The method assumes that sex, race, and age prevalence proportions in US and SEER are similar and does not control for other differences between SEER and US such as socio-economic status, etc…

HANDLING THE UNKNOWN RACE

- SEER data may have missing/unknown race for some cases
- Populations do not have unknown race
- How do we handle race Unknown when estimating the # of cancer survivors in the U.S. for all races combined?
- We estimate prevalence of whites and unknown race combined and multiplied by the US white population

PREVALENCE INTERPRETATION

- Prevalence is a function of cancer incidence, cancer survival, general life expectancy and age structure of the population.
- While the interpretation of survival and incidence statistics is straightforward, understanding prevalence is complex.
- Low incidence → good news
- High survival → good news
- High prevalence can be bad or good news → It depends on the reasons why prevalence is high

PREVALENCE INTERPRETATION

- Higher prevalence can occur because:
  - Higher survival (better treatments)→ good news
  - Higher incidence (higher risk factors) → bad news
  - Older population → Nothing we can control
- Cancer prevalence alone:
  - Not a suitable statistic for epidemiologic studies or when examining health disparities
  - A health services research measure used to quantify the cancer burden for planning, resource allocation, and assessing the survivorship community

ESTIMATING US PREVALENCE DEMO

PROJPREV

HTTPS://SURVEILLANCE.CANCER.GOV/PROJPREV/

EXCEL

CANCER PREVALENCE IS NOT "ONE SIZE FITS ALL"

- We present some definitions – there are many other definitions
  - People affected by cancer (e.g. Families of cancer patients)
  - People in different phases of care
  - By time since diagnosis (can provide insight on phases of care)
  - People diagnosed as children or as adolescents or young adults
  - Cancer survivors living with metastatic disease
  - People in recurrence or who will eventually recur (recurrence prevalence)

PREVALENCE RESOURCES

- To calculate by yourself:
  - SEER*Stat (https://seer.cancer.gov/seerstat/)
  - COMPREV (https://surveillance.cancer.gov/comprev/)
- Pre-calculated prevalence statistics by cancer site and sex are published in
  - SEER*Explorer (https://seer.cancer.gov/explorer/)
  - CanQues (http://surveillance.cancer.gov/prevalence/canques.html)
  - Office of Cancer Survivorship website (http://cancercontrol.cancer.gov/ocs/prevalence/index.html)
  - American Cancer Society's: Cancer Treatment & Survivorship Facts & Figures (2022)
- Methods: http://surveillance.cancer.gov/prevalence/

METHODS TO ESTIMATE COMPLETE PREVALENCE

- COMPREV
- MIAMOD/PIAMOD

COMPLETENESS INDEX METHOD (COMPREV)

- Completeness index method is implemented in the COMPREV software
- Estimates complete prevalence by "completing" limited duration prevalence
- It uses incidence and survival models to estimate cases diagnosed prior to start of cancer registration and the probabilities that they are alive at the prevalence date

Complete Prevalence (ComPrev) Software

Version 3.0.31 (Beta) released August 2022

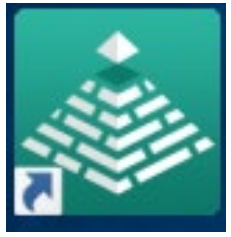
We are pleased to announce a beta version of a completely new Comprev software application. This new version has been completely updated with a new interface and online help. This release is labeled “Beta”, which means we continue to fix bugs and add new features to the program. At this time, there are no known calculation errors in the program.

The ComPrev software calculates complete prevalence based on limited-duration prevalence statistics. Complete prevalence represents the proportion of people alive on a certain day who previously had a diagnosis of the disease, regardless of how long ago the cancer was diagnosed. The software contains incidence and survival models estimated using SEER cancer data for a combination of cancer sites, sex and races. These models are used to calculate the completeness index (Capocaccia & De Angelis, 1997; Merrill et al., 2000). The completeness index represents the percent of "completeness" of limited duration prevalence. For example, 90% indicates that limited duration prevalence represents 90% of complete prevalence. Complete prevalence is calculated by dividing limited duration prevalence by the completeness index as a proportion. Limited-duration prevalence statistics can be extracted from SEER*Stat software and imported into ComPrev. The software provides standard errors for complete prevalence estimates using methods from Gigli et al., 2006.

Tutorials

- A collection of tutorials for Comprev can be found on the Tutorials page of the online help.

https://surveillance.cancer.gov/comprev/

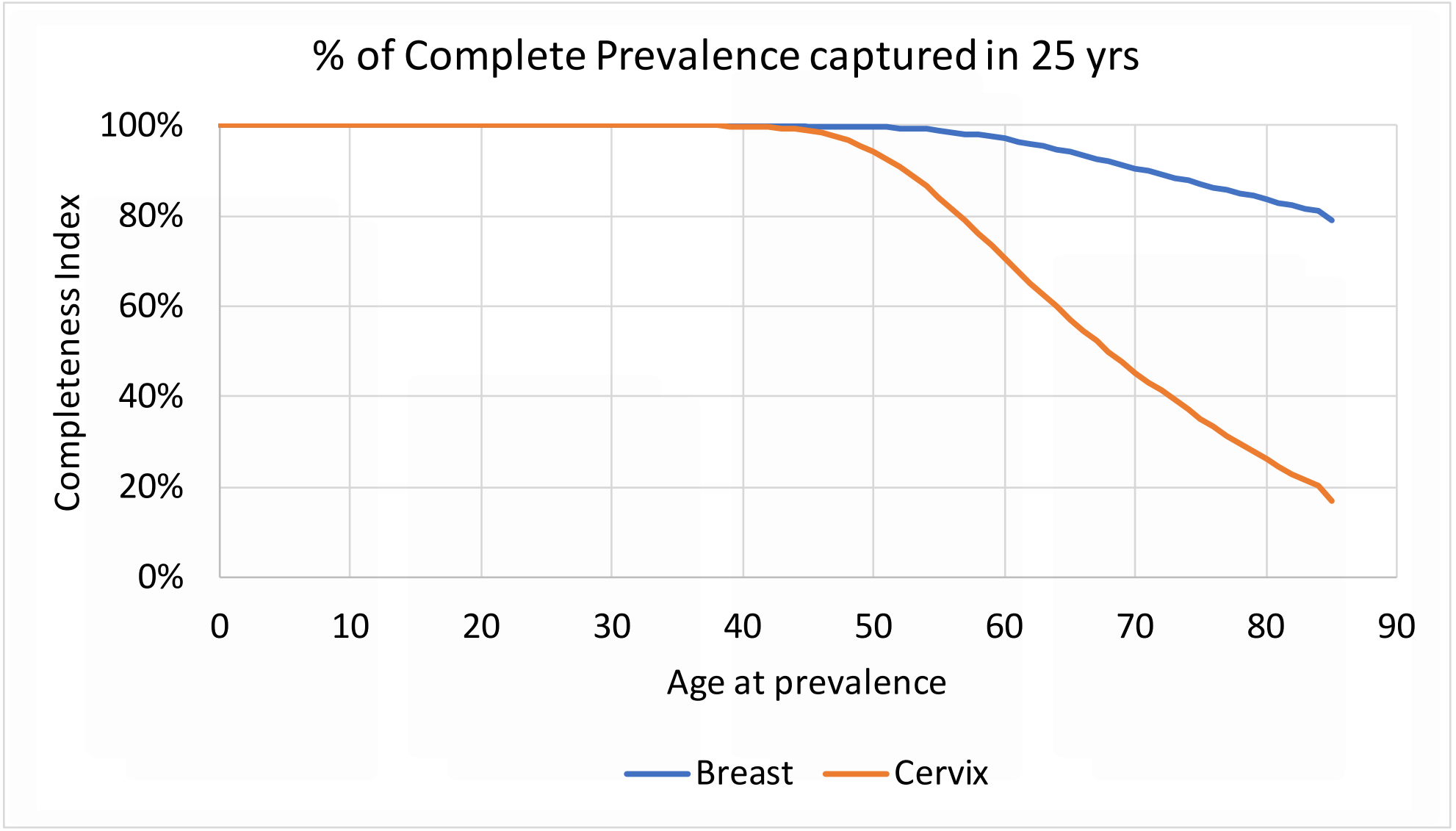


COMPLETE PREVALENCE AT AGE 70 AT 1/1/2020 USING DATA FROM 2000-2019

- The COMPREV software includes incidence and survival models for major cancer sites by sex, previously estimated using the historical SEER-9 registries
- Requires limited duration prevalence by age as input
- The method is based on the prevalence equation

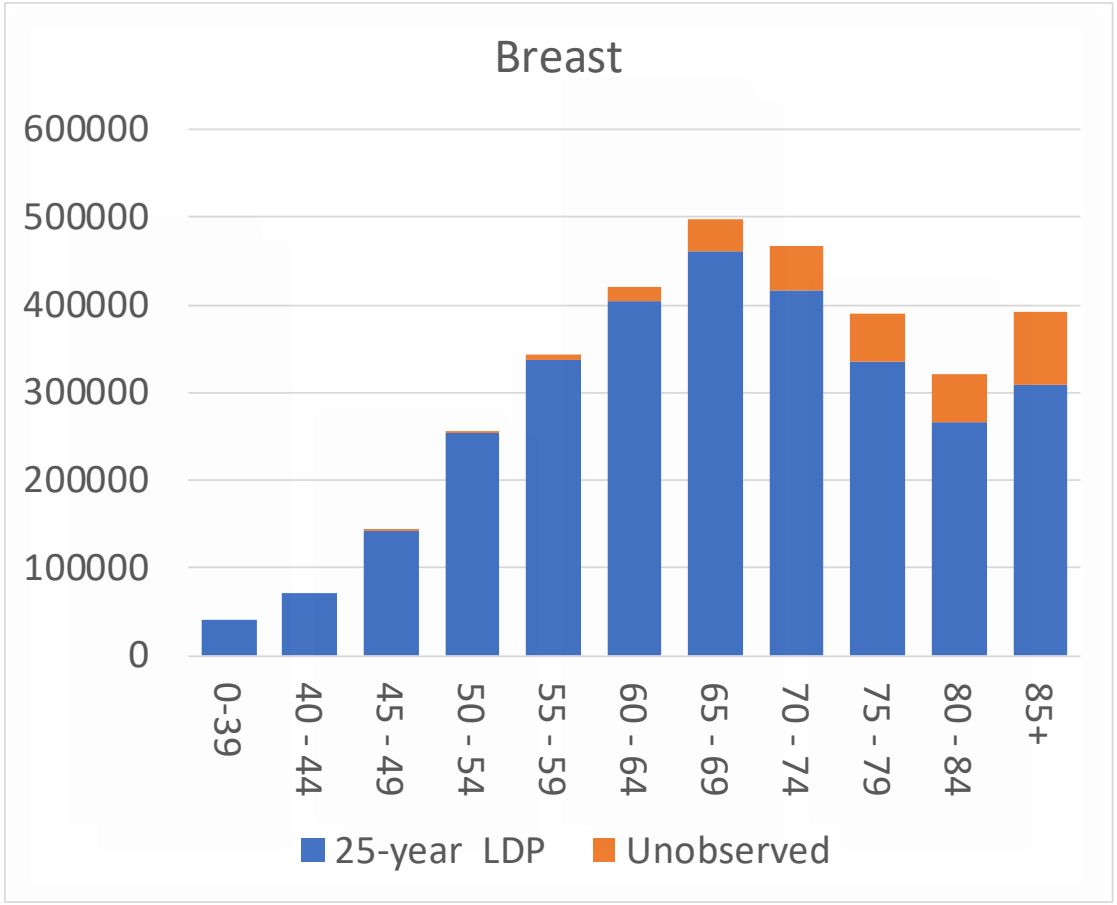
$$\text{Prev}(\text{current age} = 70) = \sum_{ageDX=0}^{70} \text{I}(ageDX)\ \text{Surv}(70 - ageDX)$$

Completeness Index

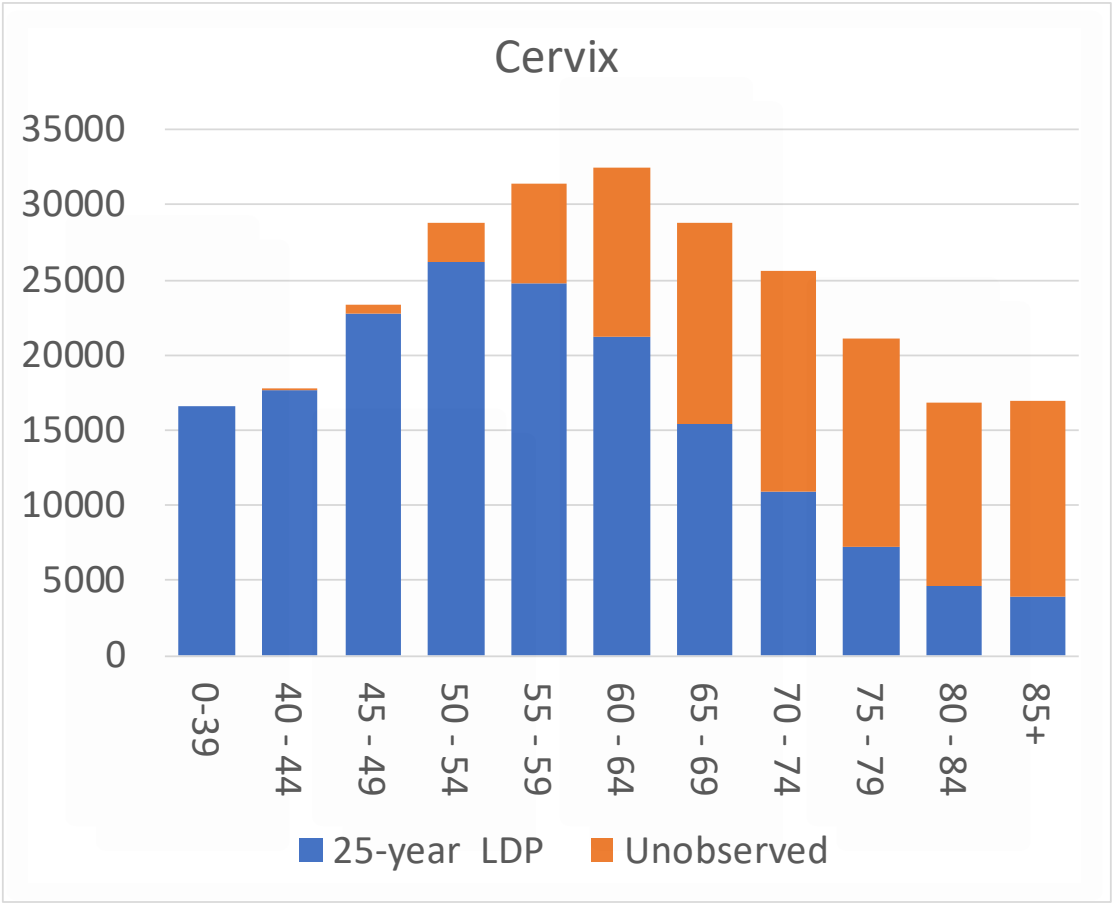


25-year Limited Duration and Complete Prevalence

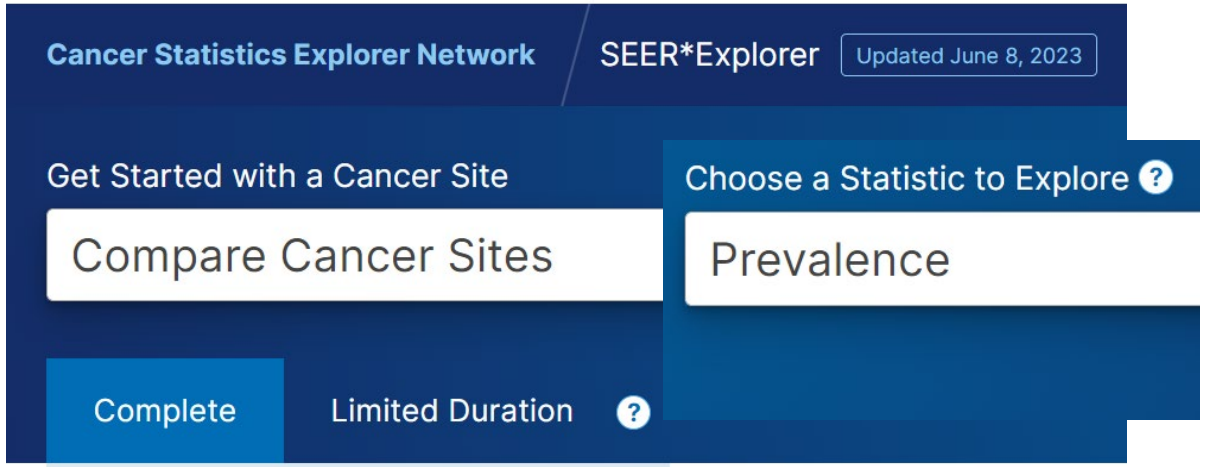
25-year = 3.0 M (90%)
Complete= 3.3 M



25-year = 171,000 (66%)
Complete = 260,000



Age at prevalence



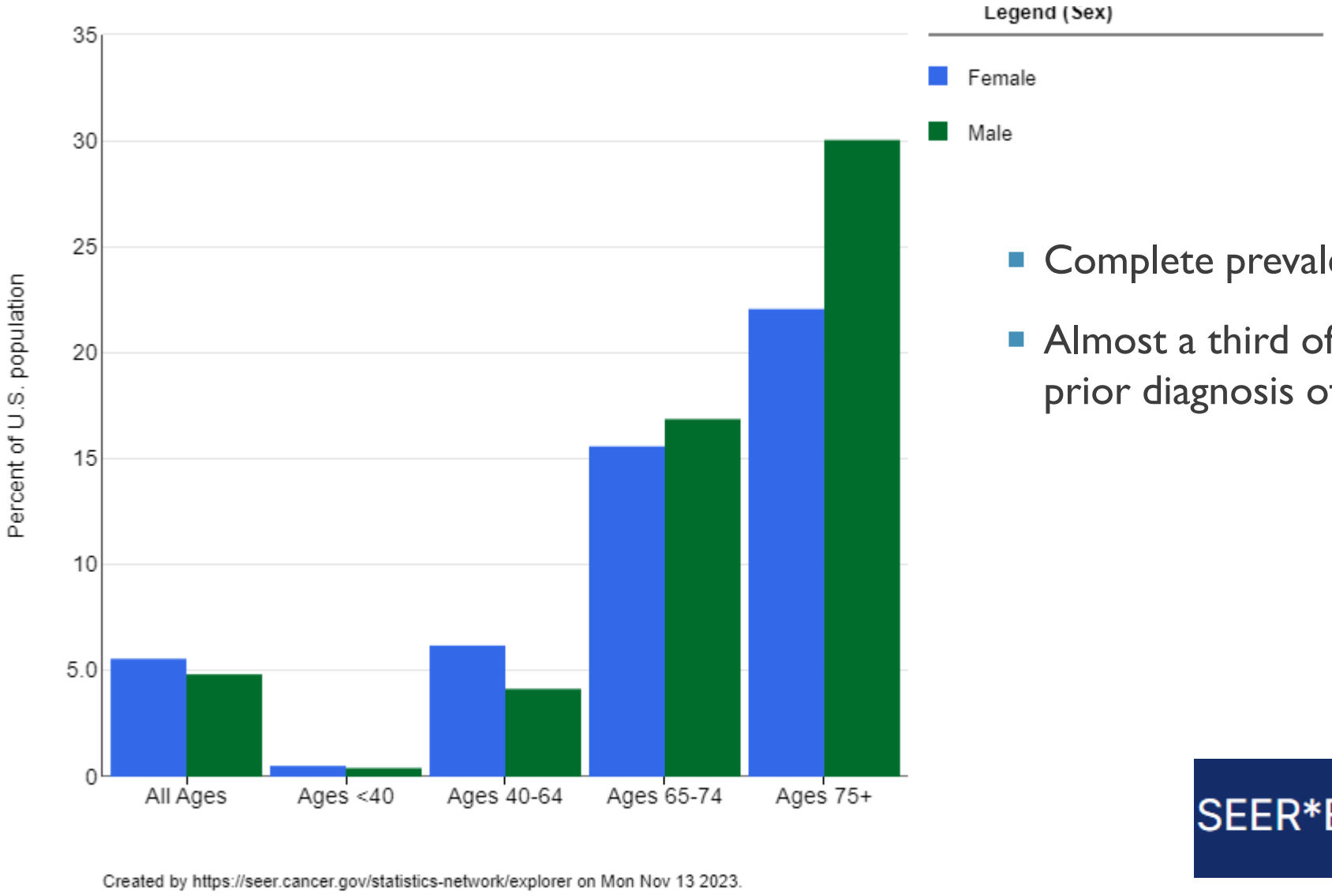
https://seer.cancer.gov/explorer/

Cancer sites diagnosed at a younger ages and good prognosis→Limited duration prevalence is less complete

Cancer Site	Complete Prevalence	% of US pop.	28-year LD Prevalence	% of complete prevalence
All Cancer Sites Combined	17,113,494	5.2	15,855,954	93%
Female Breast	3,886,830		3,617,942	93%
Prostate	3,343,976		3,323,494	99%
Melanoma of the Skin	1,413,976		1,244,357	88%
Colon and Rectum	1,388,422		1,314,740	95%
Thyroid	951,193		794,543	84%
Corpus and Uterus, NOS	845,825		751,277	89%
Non-Hodgkin Lymphoma	788,781		750,602	95%
Bladder (Invasive & In Situ)	725,549		679,097	94%
Kidney and Renal Pelvis	628,355		582,734	93%
Lung and Bronchus	603,989		569,877	94%
Leukemia	490,875		456,481	93%
Oral Cavity and Pharynx	424,284		386,342	91%
Cervix Uteri	296,981		203,897	69%
Testis	291,294		218,471	75%
Ovary	236,511		197,092	83%
Hodgkin Lymphoma	223,512		165,856	74%
Brain and Other Nervous System	180,047		143,932	80%
Myeloma	170,405		168,234	99%
Stomach	127,211		120,578	95%

All Cancer Sites Combined
People Alive with Cancer (U.S. Prevalence) on January 1, 2020

By Age at Prevalence and Sex



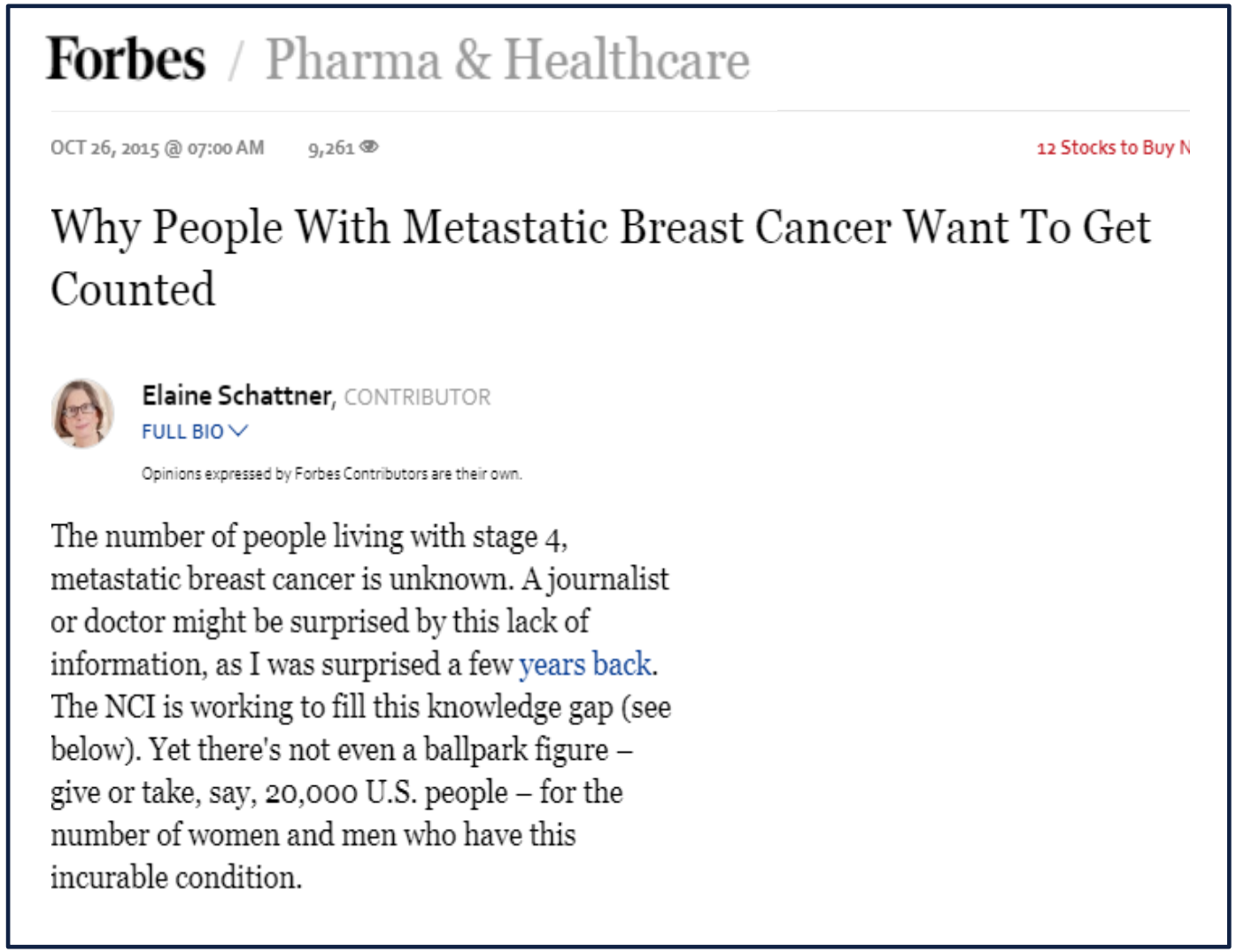
- Complete prevalence (SEER*Stat + COMPREV)
- Almost a third of US men ages 75 and older had a prior diagnosis of cancer

SEER*Explorer Updated July 31, 2023

OTHER METHODS FOR ESTIMATING PREVALENCE

- Back Calculation/Transition Rate Methods:
  - MIAMOD uses cancer mortality and survival (back-calculation)
  - PIAMOD uses cancer incidence and survival
- Because they model incidence and survival they allow for future projections
- https://www.iss.it/en/-/miamod-piamod

STUDY 1: ESTIMATING THE NUMBER OF WOMEN LIVING WITH METASTATIC BREAST CANCER



Forbes / Pharma & Healthcare

OCT 26, 2015 @ 07:00 AM 9,261 12 Stocks to Buy N

Why People With Metastatic Breast Cancer Want To Get Counted

Elaine Schattner, CONTRIBUTOR
FULL BIO

Opinions expressed by Forbes Contributors are their own.

The number of people living with stage 4, metastatic breast cancer is unknown. A journalist or doctor might be surprised by this lack of information, as I was surprised a few years back. The NCI is working to fill this knowledge gap (see below). Yet there's not even a ballpark figure – give or take, say, 20,000 U.S. people – for the number of women and men who have this incurable condition.

- We knew the number of women alive and initially diagnosed with stage 4 metastatic breast cancer
- What we didn't know was the number diagnosed with early stage breast cancer who progressed to metastatic breast cancer
- Modeling to fill the gaps in the data

BURDEN MEASURE: METASTATIC CANCER SURVIVORSHIP

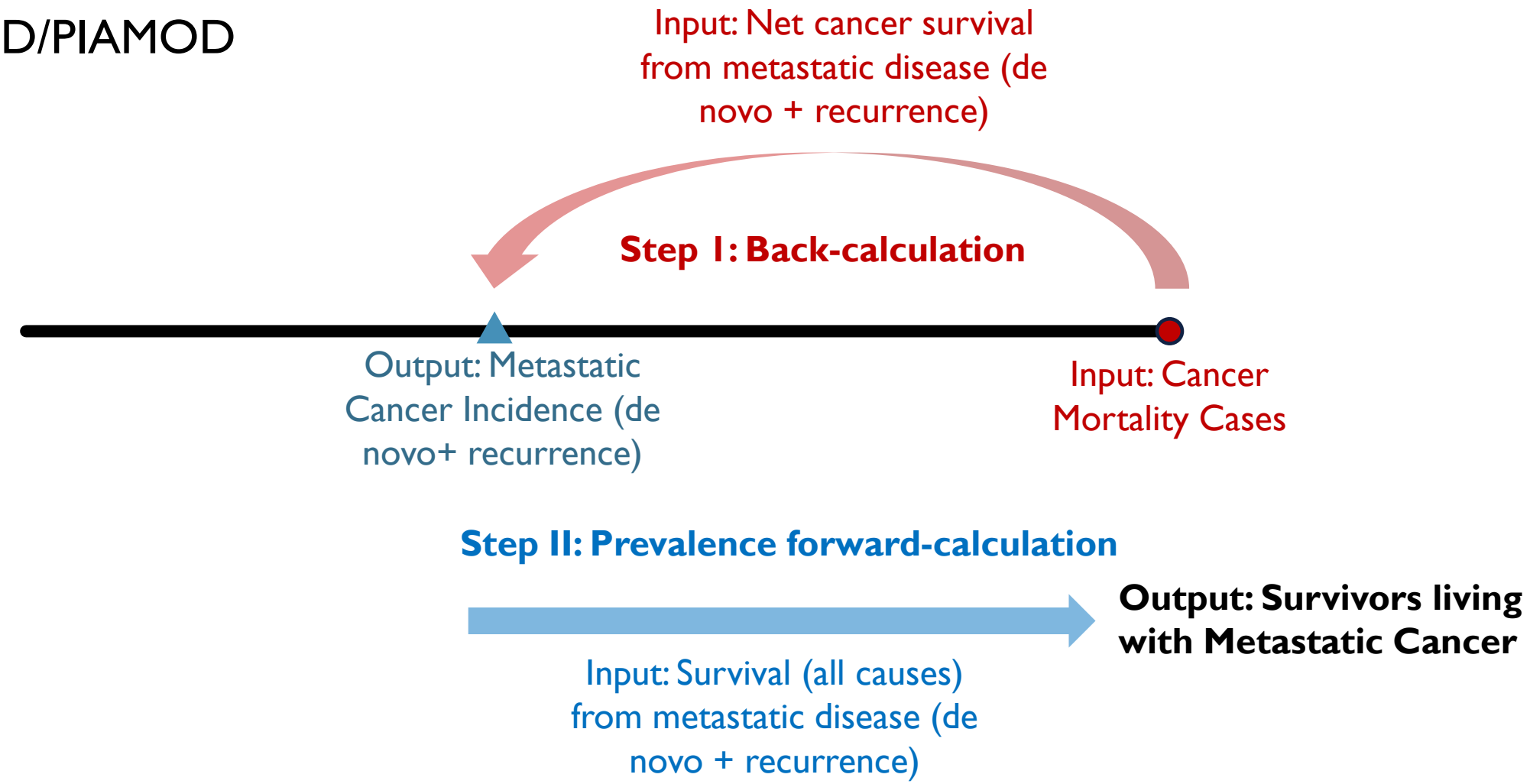
- The 2021 NCI Office of Cancer Survivorship workshop was convened to identify gaps and opportunities pertaining to survivors living with metastatic disease, an often-overlooked population
- One gap/opportunity identified was
  - Development of new strategies to **estimate the number of individuals living with metastatic cancers including those who are diagnosed with or advanced to metastatic cancer**

*Mollica et al. Survivorship for individuals living with advanced and metastatic cancers: NCI meeting report. J Natl Cancer Inst. 2021;114(4):489–495.

Estimation of the Number of Individuals Living With Metastatic Cancer in the United States

Lisa Gallicchio, PhD,* Theresa P. Devasia, PhD, Emily Tonorezos, MD, MPH, Michelle A. Mollica, PhD, MPH, RN, OCN, Angela Mariotto, PhD

Method: MIAMOD/PIAMOD



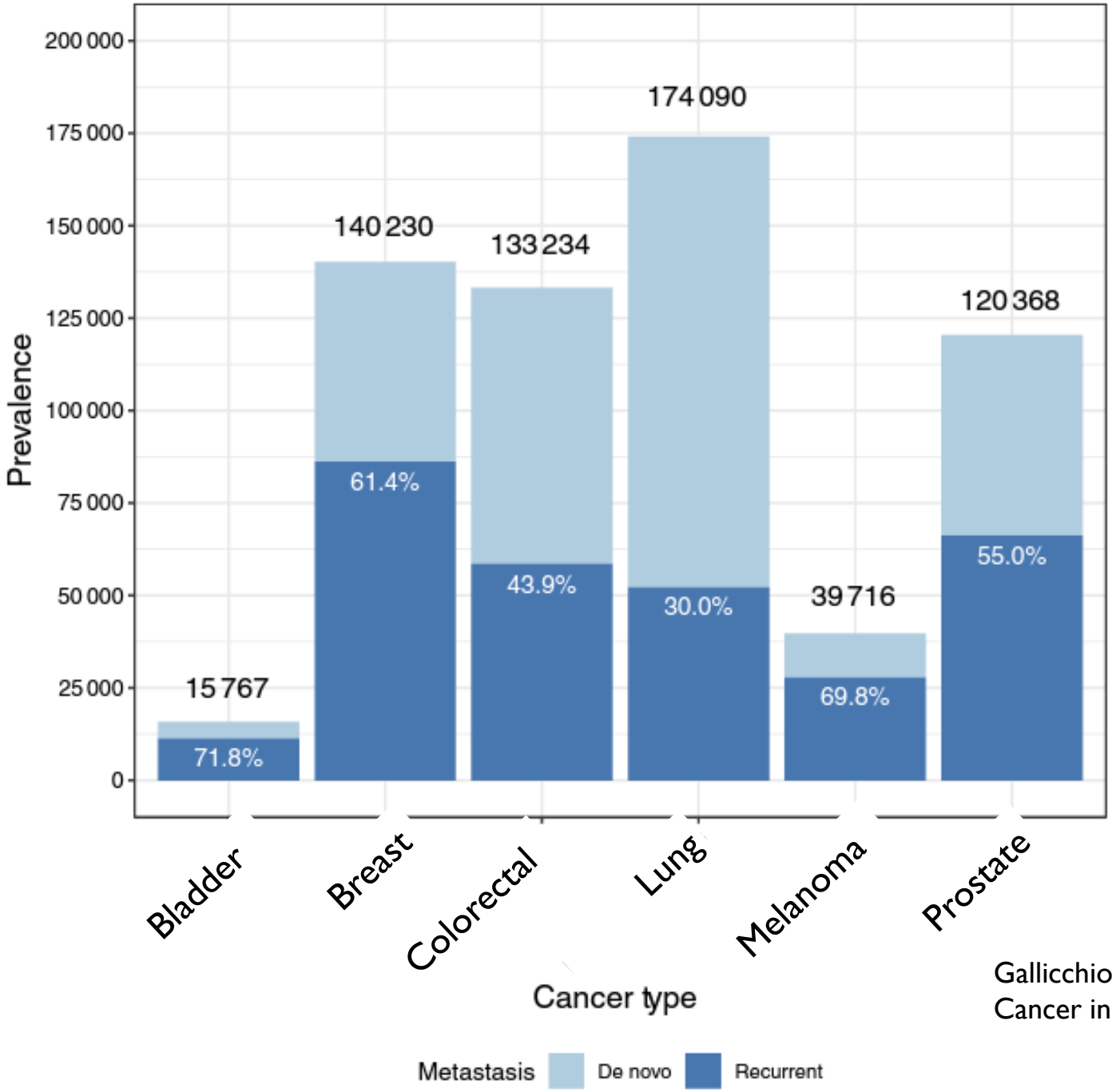
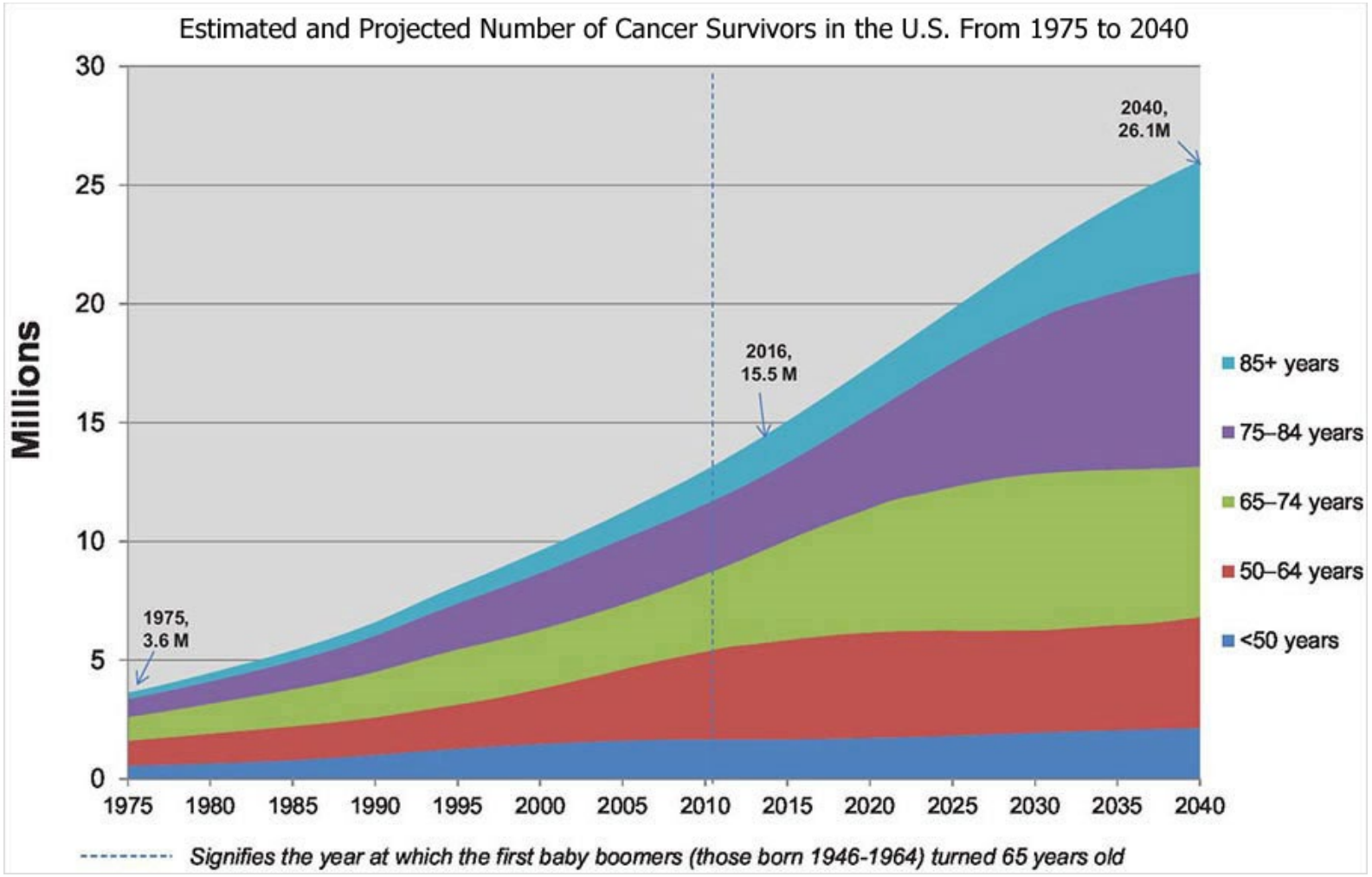


Figure2. Number of metastatic cancer survivors (January1,2018). Percents who were initially diagnosed with metastasis (de novo)or who experienced recurrent metastatic disease after being initially diagnosed with early stage disease.

Gallicchio et al. Estimation of the Number of Individuals Living With Metastatic Cancer in the US. JNCI, 2022, 1476–1483.

NATIONAL CANCER INSTITUTE
Division of Cancer Control & Population Sciences
Office of Cancer Survivorship

https://cancercontrol.cancer.gov/ocs/statistics/index.html



- Incidence and survival data were used in PIAMOD to project prevalence in future years

THANK YOU

TYPE OF FOLLOW-UP

- Reported alive survival – registry has good death ascertainment and good follow-up for alive patients (high percentage of alive patients have date of last contact after the study cut-off date) – generally obtained through additional linkages (SSA, drivers license, voter registration…)
  - Some patients will be alive, but have a date of last contract prior to the database study cutoff date (unknown if alive on prevalence date)
  - Used for all registries in SEER databases and some registries in NAACCR databases
- Presumed alive survival – registry has good death ascertainment, but does not have good capture of date last known alive for alive patients
  - All patients will either die prior to the prevalence data (not counted in prevalence) or be presumed alive on the prevalence date (counted in prevalence)
  - Used for all registries in NPCR databases and some registries in NAACCR databases

REPORTED VS PRESUMED ALIVE PREVALENCE

Limited-Duration Prevalence Session-1 Matrix-1

Page: Standard Prevalence

	Estimated Prevalence Percent	Estimated Prevalence Count	Population at Prevalence Date	Known Alive	Lost	Lost Estimated Alive	Dead Prior to Prevalence Date
Presumed alive state	1.35829%	9,985.0	735,113.5	9,985	0	0.0	3,952
Reported alive state	1.46134%	576,317.0	39,437,536.5	535,681	47,007	40,636.0	222,388

January 1, 2019, 5-Year Limited Duration Prevalence. Populations were estimated by averaging 2018 and 2019 populations.

There are no Lost cases when using Presumed Alive Survival

ESTIMATING US COMPLETE PREVALENCE FROM SEER DATA

- Using SEER data we can estimate long duration prevalence
- The method accounts for differences in age, sex and race between SEER and US. However it does not account for other differences such as socio-economic status, etc.
- Recently NAACCR CiNA data was used to estimate 5-year LD cancer prevalence for blood cancers for the U.S., states, and for the individual Leukemia & Lymphoma Society (LLS) Chapters (patient outreach/service delivery areas).
  - Requires population-based incidence and follow-up/death ascertainment (fit for use for survival statistics)

ESTIMATING US 5-YEARS LD PREVALENCE USING MORE COMPLETE GEOGRAPHIC COVERAGE

- Recently NAACCR CiNA data was used to estimate 5-year LD cancer prevalence for blood cancers for the U.S., states, and for the individual Leukemia & Lymphoma Society (LLS) Chapters (patient outreach/service delivery areas).
- It used data from registries that contained follow-up/death ascertainment fit for use for survival statistics
- Counting method was used
  - For registries meeting SEER follow-up standards survival estimates were used to adjust for loss to follow-up
  - For other registries, it was assumed that all deaths were ascertained through the study cutoff date and remaining persons were presumed to be alive, which may slightly overestimate prevalence