Understanding SEER-Medicare data

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Population Health Share Resources (PHSR)
University of Colorado Comphrensive Cancer Center
May 2023

Outline

- The basics of both data sources
- Key data fields available in both data sources
- Data structure and additional files (e.g., non-cancer controls, other linkages)
- Examples of research questions for which SEER-Medicare data are useful
- Examples of research questions for which SEER-Medicare data are not useful
- Examples of projects (cohort selection, sample sizes, findings)
- Limitations and alternatives data sources
- PHSR pre- and post-award SEER-Medicare services

The basics

- SEER-Medicare is a linkage of the Surveillance, Epidemiology, and End Results (SEER) data with Medicare claims
- SEER collects cancer incidence data from cancer registries, but not all registries. It covers about 47.9% of the U.S. population
- Medicare claims are **billing records** based on what the Centers for Medicare and Medicaid (CMS) pays
- It does not include all people covered under Medicare; only those who enroll in fee-for-service Medicare (aka Traditional Medicare or Original Medicare)
- To date, although this will change, "claims" for those in Medicare Advantage (aka Part C or Medicare HMO) are not included – currently about 50% of the Medicare population

The basics

- Registry data are great to determine a person's diagnosis date, stage (if relevant), tumor characteristics, race/ethnicity, vital statistics
- Treatment information is limited to initial treatment with some updates
- Not longitudinal – we can't tell what happened before or after a cancer diagnosis
- Claims are longitudinal but have very little clinical information other than what is required to submit a claim
- SEER-Medicare complement each other well – with many caveats

Registries that are part of SEER

- States that contribute data in dark blue

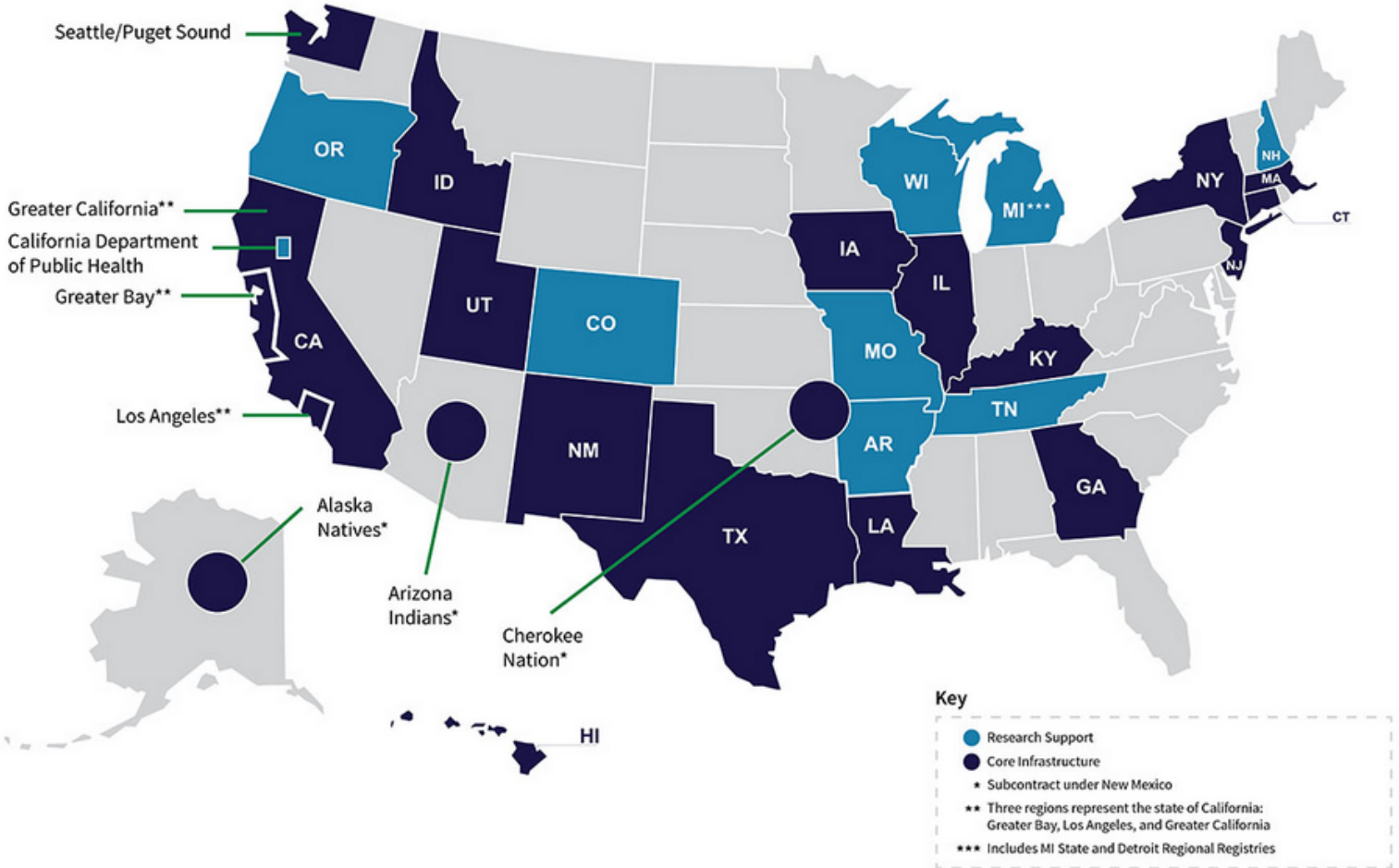
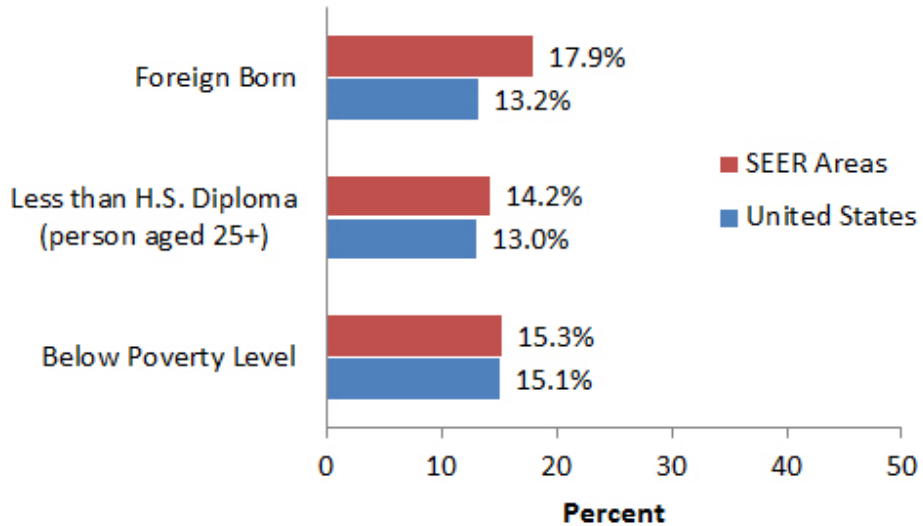


Figure: `https://seer.cancer.gov/registries/`

SEER vs the US

- Some demographic comparisons



* The data source for these is the 2012-2016 American Community Survey. SEER areas included in this figure are the States of Connecticut, Hawaii, Idaho, Iowa, Kentucky, Louisiana, Massachusetts, New Mexico, New York, Utah, Wisconsin; multi-county areas of Atlanta, rural Georgi remaining counties of Georgia, San Francisco-Oakland, Seattle-Puget Sound, San Jose-Monterey, Los Angeles county, remaining counties of California; and American Indians/Alaska Natives in Arizona, Alaska and Cherokee Nation.

Figure: `https://seer.cancer.gov/registries/`

SEER fields and advantages

- Very clear documentation of variables and variable codes/values
- Gold standard for cancer incidence
- SEER data (1975-2020) can be accessed for research
- Some fields require more justification to be released
- Several types of requests: Research, Research Limited, Research Plus

SEER fields

- Example of documentation

November 2022 Data Submission Research Plus Data Items						
	Item # refers to the NAACCR item number - see http://datadictionary.naaccr.org/default.aspx?Version=22					
	CS= Collaborative Staging					
	SSF = Site-specific Factor					
	Limited-Field refers to SEER 22 databases that have a limited number of variables					
	* Indicates that a field is available in the SEER 22 (excl IL and MA) Limited-field database					
	Columns B, C, and D list availability of the variables in Research and other databases					
	Column E indicates if a field is available at the individual case level via a Case Listing session					
Name	**Research**	**Research Limited-Field**	**Research Plus Limited-Field**	**Available in Case Listing**	**NAACCR Item #**	**Description**
Age recode with <1 year olds	Yes	Yes	Yes	Yes		The age recode variable is based on Age at I determined by the age groupings in the pop years, 5-9 years, ..., 85+ years). Will be in Race and Age (case data only) in r See data description: https://seer.cancer.gov/data-software/doc
Race recode (White, Black, Other)	Yes	Yes	Yes	Yes		Race recode is based on the race variables used to link to the populations for white, bl https://seer.cancer.gov/seerstat/variables
Sex	Yes	Yes	Yes	Yes	220	Includes 1=Male and 2=Female from Sex [N populations for males and females when ca
Year of diagnosis	Yes	Yes	Yes	Yes	390	Year of Diagnosis: values are 1973-2014 bu There are no unknown values on the file.
SEER registry	No	No	Yes	No	40	This field show the SEER registries which co diagnosis for that registry. This data item va https://seer.cancer.gov/data-software/doc
Louisiana 2005 - 1st vs 2nd half of year	No	No	Yes	No		This field is used to separate Louisiana cas to link to different population estimates us https://seer.cancer.gov/data/hurricane.ht
State-county	No	No	Yes	No		State and county at diagnosis. Can be used
						This data item identifies whether or not the limit analyses of AI/AN race to areas served

Figure: `https://seer.cancer.gov/data-software/documentation/`

Medicare claims

- People become eligible for Medicare after turning 65 or if they are disable or have End Stage Renal Disease
- Each year, people can choose between two types of Medicare: Traditional Medicare or Medicare Advantage
- Medicare Advantage is managed by private insurance companies – the insurance company is paid a capitated payment for each enrolleee
- Traditional Medicare is managed by CMS. Although Traditional Medicare is also called fee-for-services Medicare, CMS has different payment modalities
- Since 1986, hospitalization payments are based on Diagnostic Related Groups (DRGs) (aka prospective payments), not fee-for-service
- ("Value-based payment" has come to mean the opposite of fee-for-service payments; DRG falls into value-based payment models)

Medicare claims

- In 2007, 19% of the Medicare population were enrolled in Medicare Advantage
- In 2022, 48% of the Medicare population were enrolled in Medicare Advantage
- Favorable selection into Medicare Advantage was initially a problem (healthier people enrolling into MA or switching to TM when sick)
- MA penetration varies by state and region – more common in urban areas
- WY 9%, ND 9%, CO 52 %, FL 56%
- Medicare Advantage "claims" are now available, SEER-Medicare is working on linking these data to SEER

Medicare claims data

- Contrary to registry data, **claims data are difficult to use**
- The key to understanding Traditional Medicare claims data is that they are billing records, which were not designed to be used for research
- One has to understand how a particular health service is paid, and how the Traditional Medicare program is organized
- It's also important to understand how **Medicare policy changes** over time
- Claims data are pre-processed before they are released
- The **information available is related to the information present in claim forms**

Claims

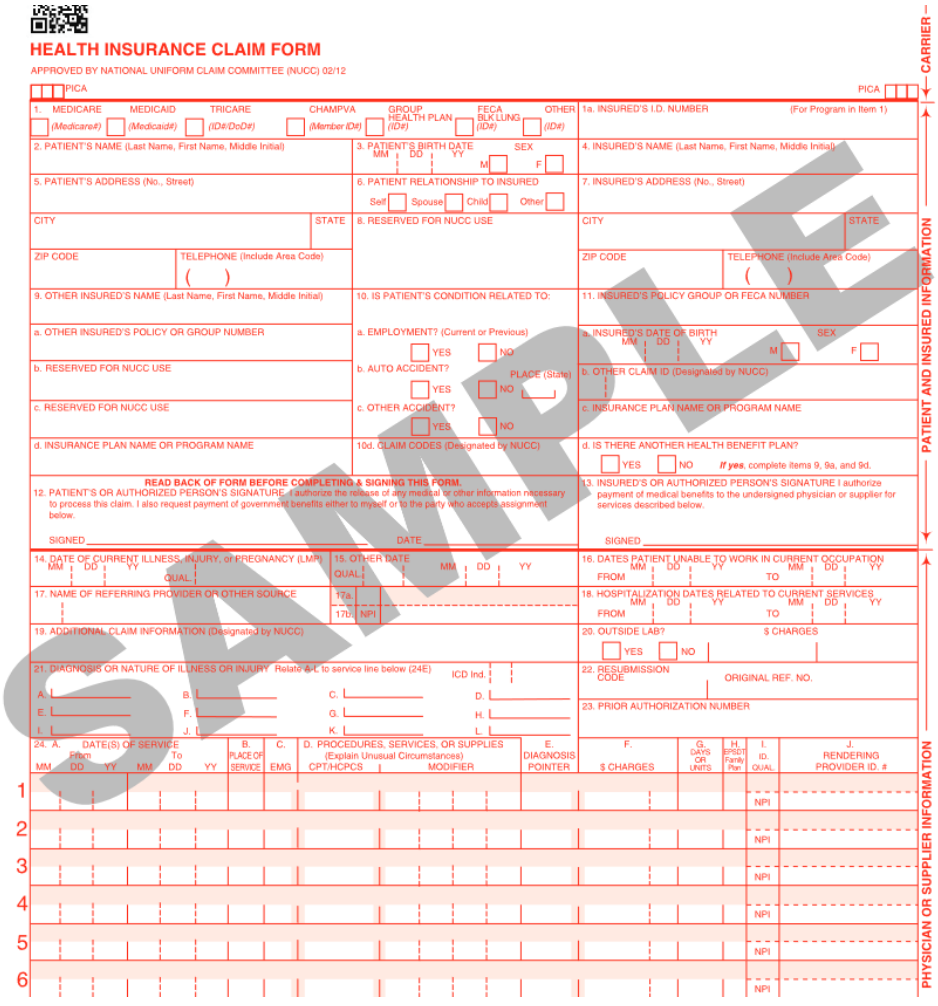


Figure: `https://www.cms.gov/Medicare/CMS-Forms/CMS-Forms/downloads/cms1500.pdf`

Data structure

- The structure of the data is linked to the structure of Traditional Medicare
- **Inpatient data**: Data for each hospitalization or stay in a Skilled Nursing Facility (SNF). Each row is a patient stay. Aliases: MEDPAR, Part A, Hospital insurance
- **Outpatient data**: Data based on doctors visits that do not require hospitalizations. Aliases: Part B, supplemental medical insurance
- **Carrier files or physician files**: Doctor visits (could be in practices or in hospitals, depending on provider organization). Aliases: Part B, supplemental medical insurance
- **Part D**: Prescription data for those enrolled in Part D (managed by private insurance). Must have part A and B to enroll

Data structure

- **Master Beneficiary Summary File (MBSF)** (aka, denominator file)
- Documentation for the Patient Entitlement and Diagnosis Summary File (PEDSF) (SEER-Medicare specific)
- Information about each beneficiary (demographics) and coverage
- State/County FIPS code, Zip code (only first 3 digits)
- Monthly enrollment codes (Parts A, B, and D, or MA)
- Most TM beneficiaries are enrolled in Parts A and B, not all are enrolled in Part D
- Most TM have some form of **supplemental insurance** like Medigap
- **Claims from supplemental insurance not included in SEER-Mediciare**

Example

Suppose you want to study women enrolled in TM diagnosed with breast cancer who are hospitalized for a surgery and afterwards receive intravenous chemotherapy plus Abemaciclib or other prescription drugs

- Details about the month (no day available) and year of diagnosis and other tumor characteristics (stage, HR+, HER2) come from the registry
- Details on the hospitalization and surgery (mastectomy, lumpectomy) are available in MEDPAR, but some information on anesthesiologist of surgeon charges could be in carrier files
- Data on IV chemo are available in the outpatient or carrier files
- Data on Abemaciclib is available in Part D – **if a person has Part D**

Example

- Even in the simple scenario, we need to think about inclusion and exclusions rules
- We would need to restrict the study to women enrolled in Parts A, B, and D at the time of diagnosis
- We would need to ensure that data with missings are removed
- Part D plans are complicated, so there is always uncertainty about whether the particular Part D plan for a person covers a medication or not
- If expenditures are of interest, we would need to investigate physician fees associated with the surgery (based on dates)
- Because we do not have information on supplemental insurance, not possible to use SEER-Medicare to study **cost sharing**

Most relevant data fields

- **Diagnostic codes**: ICD-9 and ICD-10 available in all claim files
- **Diagnostic Related Group (DRG)**: In inpatient/MEDPAR, there are also DRG codes, which are based on ICD codes (e.g., DRG 582 MASTECTOMY FOR MALIGNANCY WITH CC/MCC)
- **Current Procedural Terminology (CPT)/Healthcare Common Procedure Coding System (HCPCS)**. Procedures codes (e.g., 96401-96417, injection and IV infusion chemotherapy and other highly complex drug or biologic agent)
- **National Drug Code** (NDC) in Part D files

Data fields

- **Service dates** for all items (admission, discharge, etc)
- No exact timing (hours)
- **Source of care** (hospital, office, etc) and National Provider Identifier (NPI) for providers
- Information of providers can be linked to other sources
- Charges, payments available – charges are "list prices," so not informative
- It is possible to calculate Medicare costs (i.e., Medicare expenditures)
- (**Careful with duals**, those enrolled in Medicaid and Medicare. No Medicaid claims in SEER-Medicare)

Longitudinal data

- Data available before and after getting cancer
- Possible to code, for example, comorbid conditions X years before a cancer diagnosis
- NCI mantains the NCI Comordibidy Index (based on Charlson and Klabunde commordibity indices)
- They are dummy variables coding commorbid conditions plus a summary index
- SAS macros available (`https://healthcaredelivery.cancer.gov/seermedicare/considerations/calculation.html`)

NCI comorbidity index

Prevalence of comorbidity conditions by method (current with CPT codes, revised with and without CPT-4 codes) for breast cancer patients diagnosed 1997-2007 N=105,465

Comorbid Conditions	Frequency (Prevalence*)			Percent Change* (Revised-Original) / Original	
	Original (with CPT-4)	Revised with CPT-4	Revised without CPT-4	Revised with CPT-4	Revised without CPT-4
Conditions with CPT-4 codes					
Moderate/Severe Liver Disease	99 (0%)	114 (0%)	114 (0%)	15%	15%
Cerebrovascular Disease (CVD)	5113 (5%)	5784 (5%)	5784 (5%)	13%	13%
Peripheral Vascular Disease (surgical)†	137(0%)	133 (0%)	133(0%)	-3%	-3%
Conditions with code revisions					
Renal disease	1924 (2%)	2249 (2%)	2249 (2%)	17%	17%
Peripheral Vascular Disease (diagnosis) †	3155 (3%)	7137 (7%)	7137 (7%)	126%	126%
Paralysis (Hemiplegia or Paraplegia)	437 (0%)	608 (1%)	608 (1%)	39%	39%
Dementia	1582 (2%)	3075 (3%)	3075 (3%)	94%	94%
Mild liver disease	365 (0%)	446 (0%)	446 (0%)	22%	22%
Congestive heart failure (CHF)	7705 (7%)	8393 (8%)	8393 (8%)	9%	9%
Chronic Obstructive Pulmonary Disease (COPD)	10913 (10%)	11453 (11%)	11453 (11%)	5%	5%
Diabetes with complications	3377 (3%)	4946 (5%)	4946 (5%)	46%	46%
Diabetes	18581 (18%)	18746 (18%)	18746 (18%)	1%	1%

Figure: `https://healthcaredelivery.cancer.gov/seermedicare/considerations/comorbidity.html`

Related data

- 5% of non-cancer controls available
- SEER-Medicare has been linked to other datasets:
  - Census file data by zip code
  - Chronic conditions flags
  - Home Health Agency data (HHA)
  - Hospital characteristics
  - Minimum Data Set (MDS) – nursing homes
  - Outcome and Assessment Information Set (OASIS) – home health
  - MD-PPAS (Medicare Data on Provider Practice and Specialty) (remember, it depends on who bills!)
  - Hospital Mergers and Acquisitions File
- Medicare Advantage data should be available soon(ish)

Data issues

- People may switch to Medicare Advantage – so need to restrict to continuously enrolled in Traditional Medicare
- Conditions must be diagnosed and coded in claims, but claims are billing records. A patient may have hypertension or diabetes, but it could be undercoded in claims
- Many papers written on theses issues (comparing Medicare claims data with surverys like the MCBS for example – self-reported data)
- Difficult to know when a comorbid condition started
- No laboratory information, no test results, no pathologic reports
- Must be careful with trends because payment policy changes

Data issues

- Not possible to know every detail of a hospitalization because MEDPAR files are a summary of each stay
- For example, there is not detail on drugs administered during a hospitalization
- Outpatient, carrier files, and part D files are more detailed but not possible to distinguish "rule out" tests. **Best strategy is to rely on validate algorithms** like those in the CCW
- A patient could have received treatments not paid by Medicare

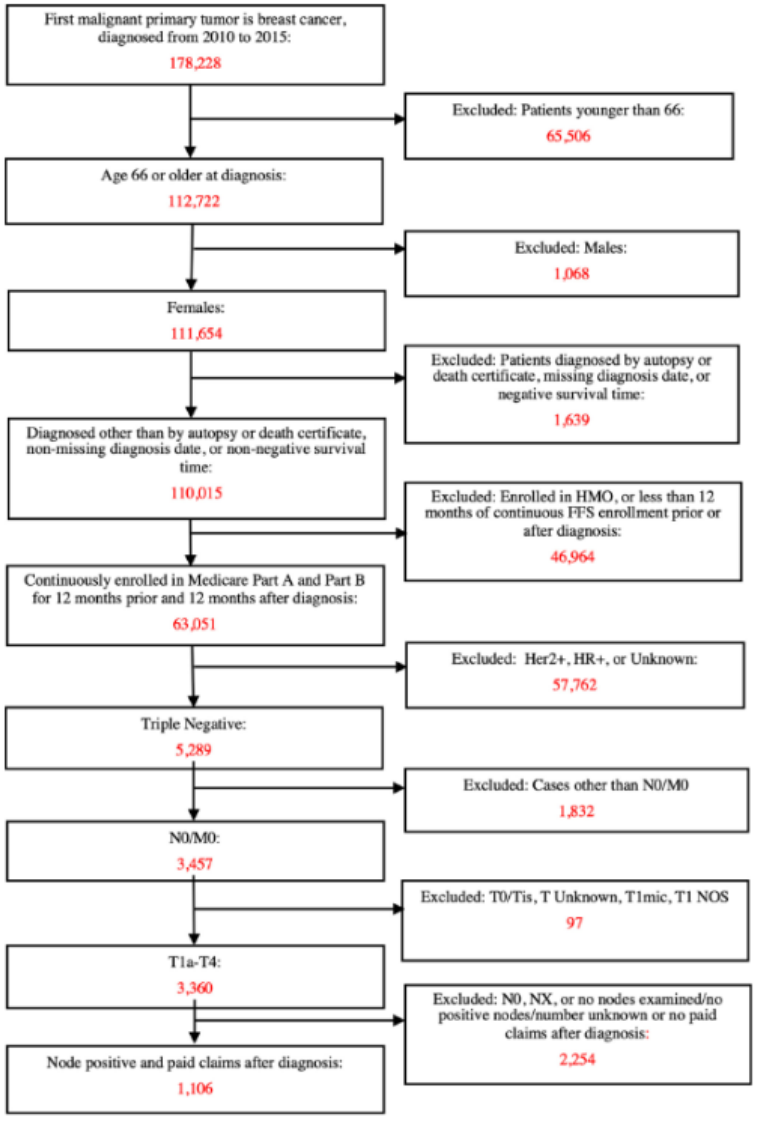
PHSR example

Roy, S., Lakritz, S., Schreiber, A. R., Molina, E., Kabos, P., Wood, M., ... & Diamond, J. R. (2023). C**linical outcomes of adjuvant taxane plus anthracycline versus taxane-based chemotherapy regimens in older adults with node-positive, triple-negative breast cancer: A SEER–Medicare study**. *European Journal of Cancer*, 185, 69-82.

Abstract ***Background:*** Triple-negative breast cancer (TNBC) is a subtype of breast cancer associated with an aggressive clinical course. Adjuvant chemotherapy reduces the risk of recurrence and improves survival in patients with node-positive TNBC. The benefit of anthracycline plus taxane (ATAX) regimens compared with non-anthracycline-containing, taxane-based regimens (TAX) in older women with node-positive TNBC is not well characterised. ***Methods:*** Using the Surveillance, Epidemiology, and End Results—Medicare database, we identified 1106 women with node-positive TNBC diagnosed at age 66 years and older between 2010 and 2015. We compared patient clinical characteristics according to adjuvant chemotherapy regimen (chemotherapy versus no chemotherapy and ATAX versus TAX). Logistic regression was performed to estimate the odds ratios (OR) and 95% confidence intervals (CIs). Kaplan—Meier survival curves were generated to estimate 3-year overall survival (OS) and cancer-specific survival (CSS). Cox proportional hazard models were used to analyse OS and CSS while controlling for patient and tumour characteristics. ***Results:*** Of the 1106 patients in our cohort, 767 (69.3%) received adjuvant chemotherapy with ATAX (364/767, 47.5%), TAX (297/767, 39%) or other regimens (106/767, 13.8%).

Figure: `https://www.sciencedirect.com/science/article/pii/S0959804923001028`

Patient selection



Patient selection

- Using the PEDSF file, identified women with an ICD-9 or ICD-10 code indicating primary tumor in the breast, diagnosed from 2010-2015
- Continuously enrolled in Medicare Part A and Part B for at least 12 months prior and 12 months after their diagnosis
- Medicare enrollment was determined using the MEDPAR file (MEDPAR contains information from beneficiary summary file)
- Sample was further limited to those who were triple negative (used ER status, PR status, Derived HER2 and Breast Subtype fields in PEDSF), T1a-T4 (used Derived AJCC fields in PEDSF), node positive (used regional lymph node field in PDSF), and had at least one paid claim after diagnosis (a claim in at least MEDPAR, Outpat, NCH, DME files)
- We applied the 2014 version SAS macro program to calculate the Charlson Index for the comorbid conditions.
- We were also interested in specific cardiac conditions that we wanted to flag and control for in the analysis.

What is SEER-Medicare not good for?

- It's not a good source of data for epidemiological studies as it does not contain all relevant populations (SEER is better but more limited)
- Not a good source to understand **all treatments performed at facilities** as it only has information on the Traditional Medicare population further restricted to participating SEER registries
- Limited clinical information
- **Observational data**. Many sources of confounding that are difficult to control for but it's possible use methods for causal inference (difference-in-difference, regression discontinuity, instrumental variables)

PHSR

- We can help you assess the feasibility of a project – we also have acess and knowledge about alternative data sources
- Our analysts have years of experience analysing SEER-Medicare and similar data. The learning curve for working with claims is steep
- We can also perform analyses or can refer you to other cores if not
- **We will purchase additional data so please contact us if you have a project idea**
- Data must be purchased for a specific project. Each paper and each revision is subject to approval from IMS, the contractors managing data releases for SEER-Medicare
- It is possible to request data for multiple related projects

Contact PHSR

- Co-directors: Marcelo Perraillon (marcelo.perraillon@cuanschutz.edu) and Jamie Studts (qualitative component)
- PHSR manager and analyst: Elizabeth Molina (elizabeth.molina@cuanschutz.edu)
- Web: `https://medschool.cuanschutz.edu/colorado-cancer-center/research/shared-resources/population-health`