# THE SURVEILLANCE, EPIDEMIOLOGY AND END RESULTS (SEER) PROGRAM AND PATHOLOGY

## REGISTRIES: BIG DATA, BIGGER PROBLEMS?

*Passara Jongkhajornpong, MD.*
*Clinical Epidemiology*

# THE SURVEILLANCE, EPIDEMIOLOGY AND END RESULTS (SEER) PROGRAM AND PATHOLOGY: TOWARDS STRENGTHENING THE CRITICAL RELATIONSHIP

Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME. The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology: Toward Strengthening the Critical Relationship. Am J Surg Pathol. 2016 Dec;40(12):e94-e102.

# The Surveillance, Epidemiology and End Results (SEER)

- Launched on January 1, 1973 by president Richard Nixon as part of the National Cancer Act
- Collects demographic, clinical and outcome information on all cancers diagnosed in representative geographic regions and subpopulations
- Included based on their ability to operate and maintain a high quality population-based cancer reporting system or Cancer Registry and to enhance the demographic and geographic diversity of data




**Figure 1. Map of Surveillance, Epidemiology, and End Results registries in the United States of America**
Geographic distribution of SEER cancer registries.

# SEER

- Population covered by SEER is representative of the general U.S. population in regard to measures of poverty and education
- Tends to have a higher proportion of foreign-born persons and urban dwellers and over-samples certain racial and ethnic minorities in order to improve diverse population representativeness
- Currently captures 400,000 cancer cases annually and stores cancer data on approximately 30% of the U.S. population.
- Pathology report is an important data; 80% of cases, pathology reports are obtained electronically in real-time from approximately 360 laboratories

# SEER

- Traditionally cancer registry staff members abstract standard data items
- Manually enter corresponding text into a data collection template
- Electronic pathology reporting by nearly 80% of laboratories has enable the use of natural language processing (NLP) software to automatically code data
- Reviewed by the registry staff; all data are checked, edited and incorporated into the annual registry database
- Submitted in a de-identified form to the National Cancer Institute (NCI)
- SEER submissions are checked for quality and completeness in the first November following the final reporting year, and released for public use and access in April of the subsequent year.

# SEER

- Funded by the Division of Cancer Control and Population Sciences (DCCPS) at the NCI
- Co-funding is provided for select SEER registries via the Centers for Disease Control and Prevention (CDC), National Program of Cancer Registries (NPCR), and participating state jurisdictions
- Coordinate with the North American Association of Central Cancer Registries (NAACCR) and NPCR to collect cancer data for the total U.S. population

# SEER Data

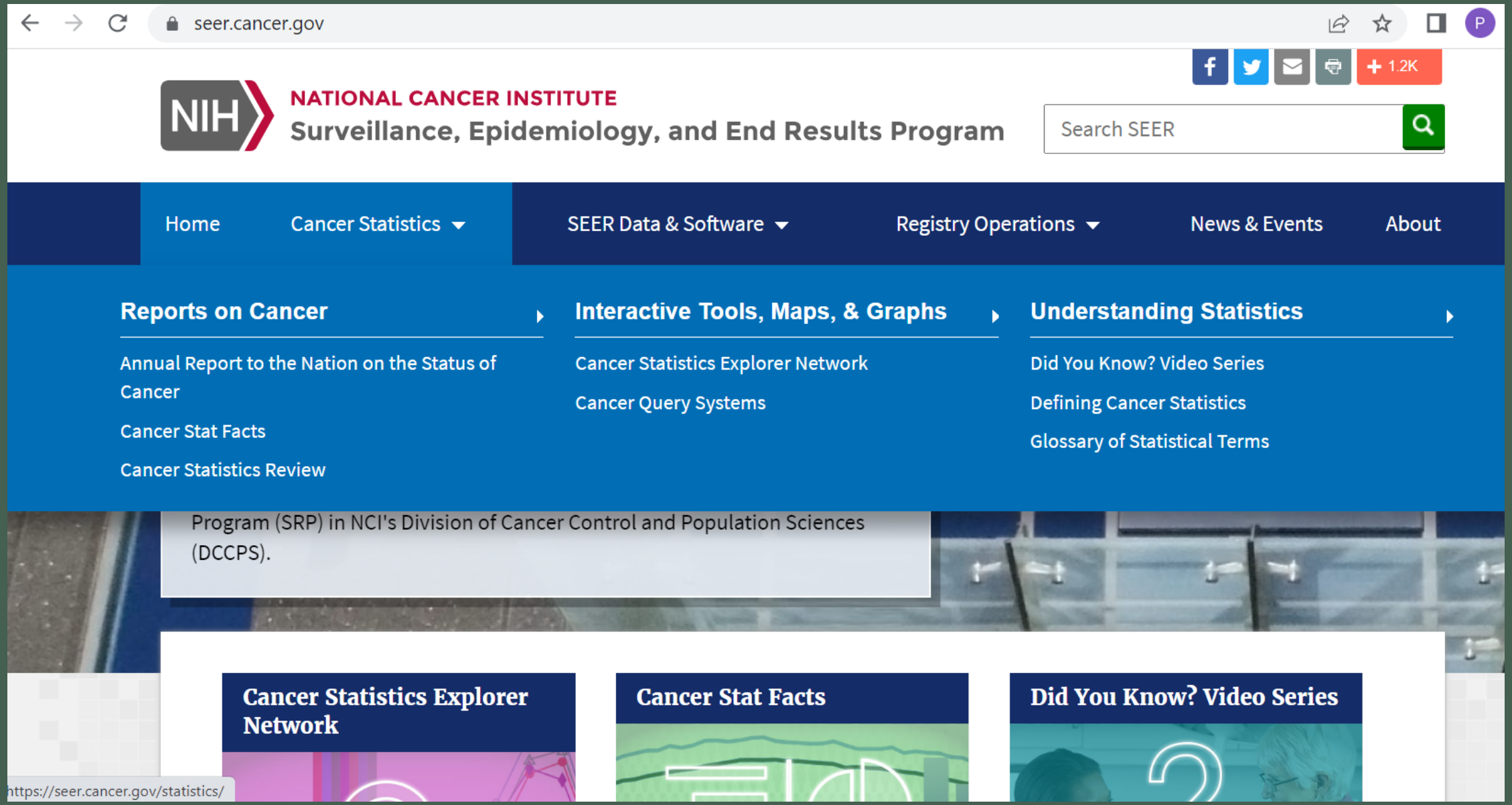
- All primary invasive cancers and some other diagnoses
- Demographic variables (age at diagnosis, gender, race/ethnicity, and county of residence)
- Surgical management and/or radiation therapy data relating to the first course of treatment
- Type of radiation therapy (neoadjuvant, adjuvant or intraoperative)
- Chemotherapy use (yes, no or unknown) may also be assessed
- Tumor data: anatomic site, laterality, size, and histopathological type based on the 2000 International Classification of Diseases for Oncology version 3 or ICD-O-3 (www.who.int/classifications/icd/adaptations/oncology)
- Tumor markers for some cancers, e.g., testis, breast, and prostate, tumor grade, extension/metastasis, site specific factors and staging based on version 7 of the American Joint Committee on Cancer (AJCC. www.cancerstaging.org)

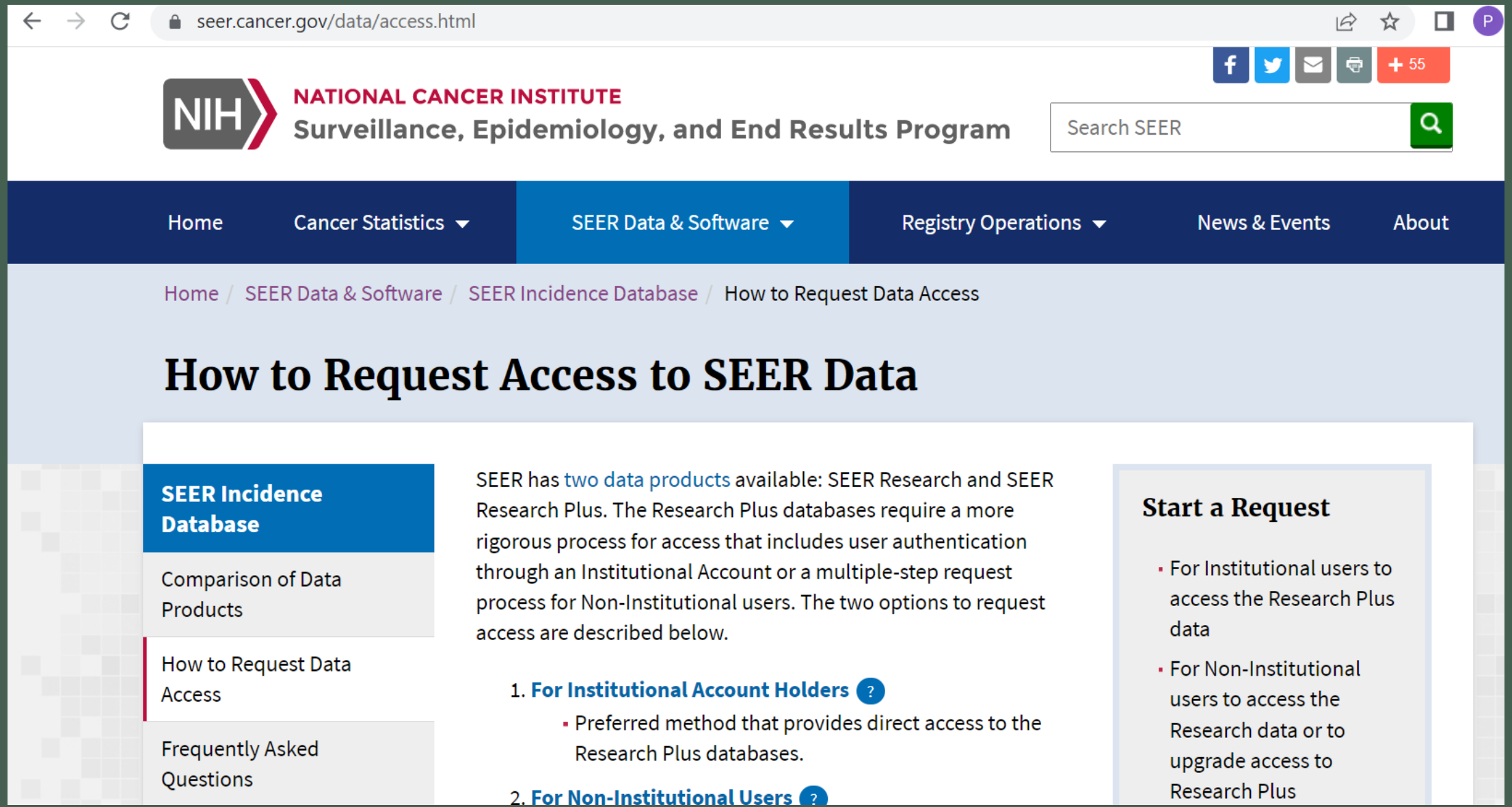
# SEER Data

- Vital status is confirmed by linkage to the National Death Index
- Follow up interval in SEER's original 7 Tumor Registries now exceeds 40 years
- Considered the gold standard for data quality amongst cancer registries in the US and globally
- Contractual agreements with regional registries and standards must be met before transmitting the data
- Virtual editing of the individual data submissions and consolidated abstracts are routinely performed ranging from 10% to 100% of all abstracts at the individual registries
- Quality program: ongoing education, training, and support for the regional registrars to prevent and correct errors in coding and identify missing data, and scheduled monitoring and evaluation of the data.

# SEER Data

- Laboratory use of standardized terminology and reporting templates
- Access to the SEER website (www.seer.cancer.gov) is unrestricted and information may be reproduced or copied without permission
- “Cancer Statistics Review (CSR)” option provides summaries of all cancers and site specific cancers in easy to understand text, graphs and figures
- “SEER data & software” option has information on data sets and software to analyze


seer.cancer.gov
+ 1.2K
NIH
NATIONAL CANCER INSTITUTE
Surveillance, Epidemiology, and End Results Program
Search SEER
Home
Cancer Statistics
SEER Data & Software
Registry Operations
News & Events
About
Reports on Cancer
Annual Report to the Nation on the Status of Cancer
Cancer Stat Facts
Cancer Statistics Review
Interactive Tools, Maps, & Graphs
Cancer Statistics Explorer Network
Cancer Query Systems
Understanding Statistics
Did You Know? Video Series
Defining Cancer Statistics
Glossary of Statistical Terms
Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS).
Cancer Statistics Explorer Network
Cancer Stat Facts
Did You Know? Video Series
https://seer.cancer.gov/statistics/


seer.cancer.gov/data/access.html
NATIONAL CANCER INSTITUTE
Surveillance, Epidemiology, and End Results Program
Search SEER
Home
Cancer Statistics
SEER Data & Software
Registry Operations
News & Events
About
Home / SEER Data & Software / SEER Incidence Database / How to Request Data Access
How to Request Data Access to SEER Data
SEER Incidence Database
Comparison of Data Products
How to Request Data Access
Frequently Asked Questions
SEER has two data products available: SEER Research and SEER Research Plus. The Research Plus databases require a more rigorous process for access that includes user authentication through an Institutional Account or a multiple-step request process for Non-Institutional users. The two options to request access are described below.
1. For Institutional Account Holders
Preferred method that provides direct access to the Research Plus databases.
2. For Non-Institutional Users
Start a Request
For Institutional users to access the Research Plus data
For Non-Institutional users to access the Research data or to upgrade access to Research Plus

# SEER Options

- *SEER Bio-specimen Pilot Programs:*

  Availability of pathology materials e.g., immunohistochemical (IHC) testing and next generation sequencing

  By 2010, the estimated number of cancer tissues was 141,241 esp. lung, colon/rectum, breast and prostate

  Tissue microarrays of some cancers e.g., breast, ovary, and colon/rectum are also available

- *SEER Data Analyses:*

  Consists of 18 regional cancer registries provides annual frequency distributions, incidence, prevalence and mortality rates over time on all cancers and site specific cancers

- SEER data resources are extensively analyzed by researchers worldwide and provide critical insights about cancer and the practice of oncology in the U.S

**Table 1**

Surveillance, Epidemiology and End Results Program: Unique Analyses and Critical Insights

| Analyses | | Critical Insights | |
|---|---|---|---|
| 1 | Population-based cancer rates | 1 | Absolute risk of cancer occurrence |
| 2 | Rare cancer rates | 2 | Precise and comprehensive description |
| 3 | Cancers rates in minority groups | 3 | Health disparity assessments |
| 4 | Birth-cohort effect | 4 | Risk factor exposure assessments |
| 5 | Calendar-period effect | 5 | Benefits/harms of screening |
| 6 | Risk modelling | 6 | Individual's risk of a particular cancer |
| 7 | Oncology practice and biomarker utilization | 7 | Therapeutic targets and diagnostic, predictive and prognostic markers. |

# Research uses of SEER data

- Typically observational and examines the distribution of cancer in populations and groups and how demographic, clinicopathological and treatment variables affect cancer burden
- Some studies link SEER data sets to others (composite data sets) analyzed to identify possible risk factors and generate hypotheses to be tested
- Absolute Risk of a Cancer, Rare Cancers and Cancers in Minority Groups
- Cancer Risk Prediction Modelling: “Breast Cancer Risk Assessment Tool” or “Gail Model” for estimating risk of developing breast cancer

# SEER: Strengths & Limitations

- ***Strengths:***
- representativeness and generalizability to the U.S. population, lengthy period of data collection, large numbers of cases, and collection of cancer specific outcomes.
- ***Limitations*:**
- Incompleteness of individual-level data collected on specific cancer risk and treatment, and inaccuracies and incompleteness of the data collected from the source registries
- Not collect data on risk-reducing procedures (e.g. risk-reducing salpingo-oophorectomy) or organ removal for non-cancer indications (e.g. hysterectomy); it is possible to adjust for such factors based on other information sources

# SEER Limitations

- Tumor recurrence data are currently not collected, therefore progression free survival, correlates of local, regional and distant control, and the effectiveness of salvage therapy cannot be assessed.
- Specific details on the type, dose and duration of chemo- and radiation therapy, and the use of other oral pharmaceuticals are not collected.
- Information gaps in treatment and follow up occur when individuals migrate into and out of SEER and non-SEER regions and could bias conclusions about a cancer behavior, particularly if re-location and outcome are linked.

# SEER Program Limitations

- Due to advances in tumor classification, SEER pathology terminology is ever changing and this creates secular issues regarding the comparability of SEER data that are captured over 40 years. Recent SEER pathology diagnoses are expected to be more precise than those from nearly 40 years ago.
- The accuracy and reproducibility of specific histotypes and cancer grades in the SEER data are mostly unknown as they have not been studied to any extent.

- Based on the few studies published, histotyping accuracy and agreement for some cancers e.g., Hodgkin's diseases may be very good, but agreement in histotyping and grading for others e.g., ovarian carcinoma are variable and organ site/cancer dependent.

Accuracy/reproducibility of select SEER pathology diagnoses in comparison to a standard review

| Organ site and/or cancer | Pathology standard | Accuracy/reproducibility |
|---|---|---|
| Kidney/renal cell carcinoma[35]<br>Clear cell carcinoma<br>Papillary carcinoma<br>Chromophobe carcinoma | 1 pathologist | Agreement=78.2% (kappa=0.55: Moderate)<br>Sensitivity/Specificity=79.1/88.1%<br>Sensitivity/Specificity=73.5/97.5%<br>Sensitivity/Specificity=72.4/97.6% |
| Lung[36]<br>Squamous cell carcinoma<br>Small cell carcinoma<br>Large cell carcinoma<br>Adenocarcinoma | 2 pathologists | Agreement=91%<br>Sensitivity/Specificity=70.9/96.2%<br>Agreement=98%.<br>Sensitivity/Specificity=94.1/98.8%<br>Agreement=87.9%.<br>Sensitivity/Specificity=21.9/93.7%<br>Agreement=82.9%.<br>Sensitivity/Specificity=80.8/84.4% |
| Hodgkin's disease[37] | 1 pathologist | Agreement=68% (kappa=0.66: Substantial) |
| Ovarian carcinoma[38]<br>*3 tier grade*<br>Serous carcinoma<br>– (2 tier grade)<br>Mucinous carcinoma<br>Endometrioid carcinoma<br>Clear cell carcinoma | 1 pathologist | Agreement=57% (kappa=0.21: Fair)<br>– Agreement=64% (kappa=0.10: Slight)<br>Agreement=44% (kappa=0.26: Fair)<br>Agreement=46% (kappa=0.26: Fair)<br>Agreement=24% (kappa=0.00: Chance) |

# SEER future and opportunities for pathology

- Expanded collection of biomarkers and treatment, custom annotation, linkage with other data bases for complete capture of relevant information, harmonization of coding systems over time, and expanded bio-specimen resources
- Electronic record linkage with pharmacy and commercial biomarker laboratory databases offers the prospect of more detailed and complete treatment and biomarker
- Use of NLP may minimize missing data and misclassification during data abstraction.
- Access to registry cancer bio-specimens for cancer research.

National Library of Medicine
National Center for Biotechnology Information

passara.jon@mahidol...

PubMed.gov

"Surveillance, Epidemiology and End Results"

Search

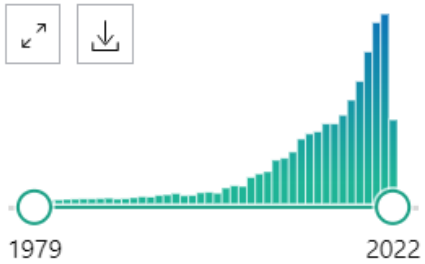
Advanced Create alert Create RSS User Guide

Save Email Send to Sorted by: Most recent Display options

MY NCBI FILTERS

All (12,203)
Review (222)

RESULTS BY YEAR




TEXT AVAILABILITY

- Abstract
- Free full text
- Full text

12,203 results

Page 1 of 1,221

1 Cite Share

Immune-related conditions and cancer-specific mortality among older adults with cancer in the United States.

Wang JH, Derkach A, Pfeiffer RM, Engels EA.

Int J Cancer. 2022 May 27. doi: 10.1002/ijc.34140. Online ahead of print.

PMID: 35633044

We evaluated 1,229,443 patients diagnosed with 20 common cancer types (age 67-99, years 1993-2013) using **Surveillance Epidemiology** and **End Results**-Medicare data. With Medicare claims, we ascertained immune-related medical conditions diagnosed before ca ...

2 Cite Share

Development of a Deep Learning Model for Malignant Small Bowel Tumors Survival: A SEER-Based Study.

Yin M, Lin J, Liu L, Gao J, Xu W, Yu C, Qu S, Liu X, Qian L, Xu C, Zhu J.

Diagnostics (Basel). 2022 May 17;12(5):1247. doi: 10.3390/diagnostics12051247.

PMID: 35626403

Methods The demographic and clinical features of patients with SBTs were extracted from the

Meta-Analysis > J Clin Oncol. 2016 Mar 10;34(8):803-9. doi: 10.1200/JCO.2015.62.0294. Epub 2016 Jan 19.

# Inferring the Effects of Cancer Treatment: Divergent Results From Early Breast Cancer Trialists' Collaborative Group Meta-Analyses of Randomized Trials and Observational Data From SEER Registries

Katherine E Henson [1], Reshma Jagsi [1], David Cutter [1], Paul McGale [1], Carolyn Taylor [1], Sarah C Darby [2]

## Abstract

**Purpose:** To compare the effect of breast cancer radiotherapy as estimated from observational data with findings from randomized trials.

**Materials and methods:** Rate ratios were obtained for selected end points among 13,932 women randomly assigned to receive radiotherapy or not in trials contributing to recent meta-analyses by the Early Breast Cancer Trialists' Collaborative Group. Estimates of the same quantities were derived for 393,840 women registered with breast cancer in the US SEER registries between 1973 and 2008.

**Results:** In the randomized trials, radiotherapy after breast-conserving surgery reduced mortality from both breast cancer (rate ratio, 0.82; 95% CI, 0.75 to 0.90) and all causes (rate ratio, 0.92; 95% CI, 0.86 to 0.99). Reductions of similar magnitude were seen in the trials of radiotherapy after mastectomy in node-positive disease (rate ratios, breast cancer 0.84; 95% CI, 0.76 to 0.94; all causes, 0.89; 95% CI, 0.81 to 0.97). In the observational data, radiotherapy after breast-conserving surgery was associated with much larger mortality reductions (rate ratios, breast cancer, 0.64; 95% CI, 0.62 to 0.66; all causes, 0.63; 95% CI, 0.62 to 0.65), whereas radiotherapy after mastectomy in node-positive disease was associated with substantial increases in mortality (rate ratios, breast cancer, 1.34; 95% CI, 1.31 to 1.37; all causes, 1.23; 95% CI, 1.22 to 1.25). Detailed adjustment of the observational data for potential confounders did not reduce the divergence from the randomized data.

**Conclusion:** This study of mortality after radiotherapy for breast cancer found strikingly divergent results between the Early Breast Cancer Trialists' Collaborative Group meta-analyses of randomized data and the SEER observational data, even when efforts had been made to remove confounding and selection biases. Nonrandomized comparisons are liable to provide misleading estimates of treatment effects. Therefore, they need careful justification every time they are used.


## Retraction

During further analyses of the data published in "Inferring the Effects of Cancer Treatment: Divergent Results From Early Breast Cancer Trialists' Collaborative Group Meta-Analyses of Randomized Trials and Observational Data From SEER Registries" (J Clin Oncol 34:803-809, 2016), the authors discovered an error in one line of code in the computer program used to calculate the results presented in the right side of Tables 3 and A5. Specifically, the error assigned the code "unknown" to the additional variables available in the SEER data set from 1990 onward for the majority of women. These extra variables, therefore, were not adequately taken into account in the columns labeled "Registrations Between 1990 and 2008 With Additional Stratification" in the right side of Tables 3 and A5 in the article. Results in the other tables in the article are not affected by this error.

Corrected versions of Tables 3 and A5 are presented below. The most important changes are for women with node-positive disease who were given mastectomy. Among these women, the breast cancer death rate ratio for those who were irradiated compared to those who were not was given in the original Table 3 as 1.32 (95% CI, 1.28 to 1.36), but in the corrected version it is now 0.89 (95% CI, 0.86 to 0.93), while the death rate ratio for all causes, which was given in the original Table 3 as 1.18 (95% CI, 1.15 to 1.22), is now 0.85 (95% CI, 0.81 to 0.88). Similar changes occur in the corrected version of Table A5. Other results in the right side of these two tables and in the related footnotes have also changed, but by smaller amounts.

In the corrected analyses, important differences between the randomized Early Breast Cancer Trialists' Collaborative Group (EBCTCG) data and the observational SEER data still remain, even with the additional stratification that can be performed for women registered from 1990. For example, in footnote f of Table 3, mortality from breast cancer in women with one to three positive nodes who received mastectomy is significantly higher with radiotherapy in the observational SEER data, in direct contrast to the EBCTCG data in that subgroup. The differences are summarized elsewhere.[1]

Although the overall conclusion is still valid (ie, that the observational SEER data can be misleading regarding causal effects of treatment), the incorrect results from the analyses of the SEER data for 1990 to 2008 played a major role in the original article. To mitigate any confusion due to this unfortunate error, the authors have unanimously requested that the article be fully retracted and that the updated findings be published separately in the Correspondence section of the journal. The authors apologize to *Journal of Clinical Oncology* and to its readers and reviewers.

### REFERENCE

1. McGale P, Cutter D, Darby SC, et al: Can observational data replace randomized trials? J Clin Oncol 34:3355-3357, 2016

# REGISTRIES: BIG DATA, BIGGER PROBLEMS?

Rubinger L, Ekhtiari S, Gazendam A, Bhandari M. Injury. 2021 Dec 11:S0020-1383(21)01001-9.

# Patient Registries

- Data systems organized in a way that allows the prospective collection and utilization of observational and clinical data to flexibly assess specific outcomes for a population with a stated scientific, clinical or policy purpose.
- Sourced from:
  - patient-reported data
  - physician-reported data
  - medical chart abstraction
  - electronic medical records (EMRs)
  - administrative and organization databases

# Databases VS Registries

- Database is an electronic set of data that is neither systematized nor organized for the explicit use for answering health-related questions.
- Development & innovation of electronic data collection along with the staggering size of databases has led to an increase in the number of registries used for research, policy, and administrative purposes.
- All users of registry-based literature: understand the strengths, limitations, and future directions of this information source, along with how to assess the quality registry data

# Registry classification

- Based on the methods of data collection
- Disease-, procedure- or pathology-specific registries, administrative-, health-systems-based or combined/linked registries, or product registries.
- National cancer registries of Sweden, Denmark, Finland, Iceland, and Norway, that prospectively collect data on cancer diagnoses, treatment, and outcomes > capable of integrating research-specific data
- Product registries are used for post-marketing surveillance in procedures, device, and pharmaceutical trials to demonstrate effectiveness and safety of products in real-world settings.
- Some registries overlaping categories eg., large country-wide registries for total joint arthroplasty represent a procedure specific registry, but also collect robust data on the real-world effectiveness of orthopedic arthroplasty implants

# Registry level of evidence

- Represent high-quality, prospectively designed cohort studies aimed at investigating or addressing a certain problem, hypothesis, or clinical entity, thus situating them **level II and III** quality.
- The recent advent of registry-based RCTs blurs that boundary between registry-based **level I and II evidence.**
- Registry-based RCTs represent a potential to introduce major efficiencies when it comes to patient recruitment and follow-up.
- Advantageous as rapid consecutive patient enrollment, reduced per-patient cost of trial implementation, and represent pragmatic trials that can easily be generalized to the population.

# Strengths of registries

- Large volume of collected data; efficiently produced, curated, systematized into scientific and clinically based conclusions
- Save time and cost: one of the first such trials, dubbed the TASTE trial (Thrombus Aspiration during ST-Elevation myocardial infarction), purported to spend US $50 per patient, which was estimated at 2% the cost of a conventional RCT.
- Non-RCT based registries are expensive to maintain; eg., the Swedish Total Joint Arthroplasty Registry costs 35 million euros per year to maintain - offset by the systems level cost-savings drawn from the data
- The prospective nature also presents a unique opportunity for researchers and consumers of literature alike to draw conclusions not possible with smaller cohort or database studies.

# Strengths of registries

- Large catchment of population-based data helps to reduce concerns around study participation, and the various biases (i.e. volunteer bias, selection bias, etc.), adds to the external validity of the registry-based conclusions that can have a strong role in health systems planning
- Because of the linkage of registries to interconnected health data, especially in the context of robust EMRs, now possible to retrieve extensive clinical information of registry participants, thus generating thousands of data points for a single patient.
- Taken together, this large amount of clinical data can generate impressive conclusions, should that ‘big data’ be interpreted correctly.

# Limitations of registries

- The quality of data depends on the initial purpose for the registry, and the resources and methods available to the registry administrators to upkeep.
- The design of datasets, the methodology of data collection, and the accuracy of that data gathered by and for registries may vary across the different types and locations of registries.
- The quality can be plagued by missing or incomplete data.
- Low, or biased, or unadjudicated enrollment into the registries can bias conclusions drawn from registries-based data
- In RCTs, patients are actively followed at pre-determined intervals, while patient follow-up data is added to registries on a less structured and pragmatic basis, or passively.

# Limitations of registries

- Adverse event rates are typically lower in registries than RCTs
- Registries that choose hard clinical end points (e.g., death, or revision surgery) are less susceptible to ascertainment bias and underreporting of complications or adverse events due to diverse definitions.
- it is possible to adjudicate outcome events or audit data in registries to ensure certain standards data

# Registry quality assessment

1. Consider whether the research question is appropriate for registry data, and whether translating those clinical questions into measurable exposures and outcomes is efficient and practical
2. Data sources for registries as appropriate for the type of research study published (i.e. case control, cohort-based etc.)
3. Patient population included in the registry, and the methodology of collecting data, must be representative of the clinical entity; registry size, duration of data collection, an appropriate comparison group for the study population
4. Internal & external validity; generalizability, information bias, sampling and selection bias, loss to follow up, and an assessment of the total magnitude of bias must all be assessed.

- Currently, no well-defined widely accepted quality assessment tool exists

# Future directions

- To continue to broaden the data pools available to registries,
- A strong movement to create global registries, either by amalgamation, or by creation of novel registries
- The International Orthopaedic Multicentre Study in Fracture Care (INORMUS) provides an excellent example of a de novo global registry. With the primary objective of determining the mortality, re-operation and infection rates of musculoskeletal trauma patients within 30 days post-hospital admission, the INORMUS study continues on its process of enrolling 40,000 patients from low-to-middle income countries in Africa, Asia, and Latin America.

# Future directions

- The next generation registry will shift the database-centered thinking to a focus on integration and incorporation of layers of data, to ultimately move from surveillance to improve clinical care in real time and integration in a “big data” health information system.
- With the foundation of registry data, inclusive of the robust EMR data now being incorporated and natural language processing based data elements, convoluted ML methods provide a novel way to derive clinical simulations, models and inform decision-making across medicine.

# Conclusion

- Registries represent an important source of systematized data that can generate important conclusions for ultimately improving clinical care of patients.
- The strength of registries lies in the immense amount of dedicated data.
- Limitations for registries exist. The importance of understanding the variability of the quality between registries cannot be understated.
- The future of registries is quite exciting with increases in data collection, especially with continued adoption or EMR, and application of ML and AI-based algorithms to learn from the data in real time.


THANK YOU