SEER Registries: Current Research Opportunities and Future Directions

Nadia Howlader, MS
Surveillance Research Program, NCI

California Association of Regional Cancer Registries Conference
Sacramento, CA
December 7, 2015
Outline

• SEER Mission
• Clinically Relevant Data Presentation and Analysis
• Challenges & Opportunities with Cancer Registries
• Describe Selected New Initiatives to Enhance Cancer Registries
• Conclusions
SEER Mission
Surveillance, Epidemiology, and End Results (SEER)

- SEER mission to support research
  - Diagnosis, treatment, and outcomes of cancer since 1973
  - Provide baseline data on U.S. cancer incidence and survival trends
- Population-based registries representing 30% of the U.S. population
- Over 450,000 incident cases reported annually
Clinically Relevant Data Presentation and Analysis
Female Breast Cancer

- **Delay-Adjusted Incidence**
  - Incidence stabilizing and rates converging among black and white women

- **Mortality**
  - Mortality decreasing but diverging rates in black and white women
Breast Cancer Incidence by Molecular Subtypes 2010, SEER.

HR⁺/HER2⁻  Triple-negative  HR⁺/HER2⁺  HR⁻/HER2⁺

HR⁺ in white women

Triple negative in black women

Nadia Howlader et al. JNCI J Natl Cancer Inst 2014;jnci.dju055
First Population-based Estimates of Breast Cancer Molecular Subtypes: California Registry

Christina A. Clarke et al. JNCI J Natl Cancer Inst 2012;104:1094-1101
Non-Hodgkin Lymphoma (NHL) Mortality: US General Population

How much do specific NHL subtypes contribute to this trend in mortality?

Data source: National Center for Health Statistics (NCHS)
NHL Mortality Trends by Subtypes, SEER

A. Overall NHL Death

B. NHL Death By Tumor Subtypes

Howlader et al. Cancer Epidemiol Biomarkers Prev; 2015
Introduction of R-CHOP treatment

Mortality

Howlader et al. Cancer Epidemiol Biomarkers Prev; 2015

National Cancer Institute
NHL Trends by HIV Status, SEER

A. NHL, HIV+

1996 when HAART was introduced

Incidence

Mortality

B. NHL, HIV-

Incidence

Mortality
Challenges & Opportunities with Cancer Registries
Fill in clinical gaps

Challenges & Opportunities

Current SEER

Diagnosis
First round of treatment/surgery
Recurrence
Subsequent treatments/surgery
Death (Survival)


Source: Adapted from California Health Care Foundation Report 2014

National Cancer Institute
Challenges & Opportunities

• Overall data capture process is mostly **manual**
• 65%–80% patient-specific data items come from **unstructured texts**¹
  - E-path reports or clinical notes
  - Natural Language Processing (NLP) is mature enough for text mining to obtain rich clinical information
• Due to aging of population, cancer cases expected to grow in the future
  - Information on these cases need to be abstracted with flat resources

¹ Hiatt et al. JNCI 2015

National Cancer Institute
New Initiatives
New Initiatives

• Enhance and improve SEER
  o Linkages (3 studies)
  o NLP
  o Modeling with innovative statistical methods
• Grants portfolio to support cancer surveillance research
• Strategic projects to enhance SEER’s ability to support research
  o SEER-linked Virtual Bio-Repository (VTR)
  o Virtual Pooled Registry (VPR)
1. Linkage with Pharmacy Data: Oral Drugs
SEER and IMS Health Linkage Study

• IMS Health pharmacy data 70% of pharmacy transactions
• Pilot study linking with selected SEER registries
  o Greater California
  o Los Angeles
  o Louisiana
  o Seattle
  o San Jose-Monterey
  o San Francisco-Oakland
• Comparing completeness with
  o Medicare
  o Patterns of Care
  o Investigating treatment completeness for 4 cancer sites (breast, colon, CML, multiple myeloma)
Preliminary Results from IMS Health Pilot

Estimates of the Percent of ER or PR Positive Breast Cancer Cases that Received Hormonal Therapy in Women Age 65 and Older (N=6128)

National Cancer Institute
2. Linkage to Enhance Infusion-based Chemotherapy in SEER

• Central claims processors (oncology)

• Represent populations of patients from providers includes all insurers

• Pilot with Georgia cancer registry
  o 150+ distinct oncology providers in process for detailed treatment and case finding
  o Capture ~41% of all providers in Georgia (but reaching out to bigger facilities with other claims processors)
  o >1 million test claims processed
Preliminary Data from 6 Months Claims in 4 Georgia Oncology Practices: Common Regimens for Treatment of Initial and Recurrent Breast Cancer

**Common Regimens for Initial Breast Cancer Treatment**
- **Pertuzumab + Trastuzumab + Docetaxel**: 4,676 administrations
- **Doxetaxel + Carboplatin + Trastuzumab + Pertuzumab**: Frequency of Administration
- **Doxetaxel + Carboplatin + Trastuzumab + Pertuzumab**: Frequency of Administration
- **Doxorubicin + Cyclophosphamide**: Frequency of Administration

**Common Regimens for Recurrent or Metastatic Breast Cancer**
- **Epirubicin + Cyclophosphamide**: Frequency of Administration
- **Docetaxel + Capecitabine**: Frequency of Administration
- **Flourouracil + Epirubicin + Cyclophosphamide**: Frequency of Administration
- **Gemcitabine + Carboplatin**: Frequency of Administration
- **Paclitaxel + Bevacizumab**: Frequency of Administration
- **Cyclophosphamide + Methotrexate + Flourouracil**: Frequency of Administration
- **Pertuzumab + Trastuzumab + Paclitaxel**: Frequency of Administration
- **Pertuzumab + Trastuzumab + Docetaxel**: Frequency of Administration
- **Gemcitabine + Paclitaxel**: Frequency of Administration
3. Linkage with Commercial Partners: Oncotype DX

- The 21-Gene Recurrence Score® Assay (Oncotype DX®) is validated to predict recurrence risk and chemotherapy benefit in estrogen receptor-positive (ER+) breast cancer.

- SEER collected this information for 2010+ breast cancer cases (but felt the information was incomplete):
  - Largely test results sent directly to physician practices.

- Oncotype DX: Linkage with Genomic Health (data 2004-2012):
  - *Added 40% test results* to existing data from hospital reported results.
  - Currently analyzing differences between reported and non-reported patients and results.
Chemotherapy Use by Oncotype DX Risk Category & Race. SEER, 2010-2012

% Receiving Chemotherapy

- White
- Black
- Other
- Unknown

Risk Categories:
- Low risk
- Intermed risk
- High risk
- No test

National Cancer Institute
Addressing Challenges with Newly Collected Data Items: Missing data

• Registries tend to collect all data items on all cancer patients

• However, initial year collection of a new data item may have incomplete information
  o e.g., ER status, HER2 status

• Use statistical modeling (e.g., multiple imputation) to fill-in missing values
Distribution of Unknown ER Status Changing

Observed Data

- Positive
- Negative
- Missing

Ignoring Missing ER

- Positive
- Negative

National Cancer Institute
Use of Imputed Population-based Cancer Registry Data as a Method of Accounting for Missing Information: Application to Estrogen Receptor Status for Breast Cancer

Nadia Howlader*, Anne–Michelle Noone, Mandi Yu and Kathleen A. Cronin

Abstract

The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program provides a rich source of data stratified according to tumor biomarkers that play an important role in cancer surveillance research. These data are useful for analyzing trends in cancer incidence and survival. These tumor markers, however, are often prone to missing observations. To address the problem of missing data, the authors employed sequential regression multivariate imputation for breast cancer variables, with a particular focus on estrogen receptor status, using data from 13 SEER registries covering the period 1992–2007. In this paper, they present an approach to accounting for missing information through the creation of imputed data sets that can be analyzed using existing software.
NLP Pilot

• SRP working with multiple commercial and academic partners to evaluate and compare NLP capabilities of each

• 2568 de-identified path reports (5 registries)
  o Represent 70-80 pathology labs
  o Two sets (a training and validation) of path reports are manually annotated to establish a “gold standard”
  o Data items to be extracted for breast and lung cancer cases: primary cancer site, histology, grade, tumor size, biomarkers, etc
  o Compare automated extraction to the “gold standard”
SEER-Linked VTR

• VTR pilot is a project to determine if SEER registries can identify valuable biospecimens that would support contemporary cancer research questions

• Pilot to
  o Determine ethical and privacy requirements of the laboratory and define best practices
  o Quality control process
  o To locate biospecimens in pathology laboratories ability to perform custom annotation
  o To determine per subject costs to identify biospecimens and perform custom annotation
Other Projects/Initiatives

- SEER Virtual Pooled Registry
- Capturing outcomes other than survival
- Patient Generated Health Information
  - Pilot studies Avon & Met breast alliance/ACS/NCI/NAACCR: (focus groups)
  - Pilot to assess predictors of financial toxicity using public data files
Conclusions
Conclusions

• Many new strategic initiatives underway to enhance the surveillance infrastructure

• Result in collection of more clinically relevant data elements
  o Leading up to diagnosis (e.g., risk factors)
  o Between diagnosis and death (e.g., recurrence, detailed treatment, biomarkers)

• Position cancer registries to support next generation cancer research
Thank you!
Contact Information

Nadia Howlader, MS
Mathematical Statistician
Surveillance Informatics Branch
Surveillance Research Program, NCI
E-mail: howladern@mail.nih.gov
Extra Slides
Evaluation of NLP tools

- Use similar approach/methods for all NLP tools if possible and feasible
  - The same set of path reports divided in:
    - Training/development set (provided to NLP entities)
    - Validation set (kept in house at IMS)
  - Both sets are manually annotated to establish a “gold standard”
    - By cancer registrars (2 Westat CRs)
    - 10% of reports to be abstracted by both CRs
Pilot Competition for NLP

- Linguamatics/ I2E
- ASCO CancerLinQ/SAP
- IBM Watson
- Health Language Analytics
- DOE
- Fred Hutchinson CCC NLP
- Kentucky CR NLP
### Preliminary Data from GA Claims Pilot (6+ Months Claims)

**Note this includes Only DX codes with counts >10,000**

<table>
<thead>
<tr>
<th>count</th>
<th>diagnosis_code</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>103054</td>
<td>1749</td>
<td>Malg neo breast (female), unspec</td>
</tr>
<tr>
<td>46358</td>
<td>1629</td>
<td>Malg neo bronchus &amp; lung, unspec</td>
</tr>
<tr>
<td>36836</td>
<td>185</td>
<td>Malignant neoplasm of prostate</td>
</tr>
<tr>
<td>34871</td>
<td>1744</td>
<td>Malg neo upper-outer quadrant</td>
</tr>
<tr>
<td>30275</td>
<td>20300</td>
<td>Multiple myeloma w/out mention of remission</td>
</tr>
<tr>
<td>27950</td>
<td>1539</td>
<td>Malg neo colon, unspec</td>
</tr>
<tr>
<td>27777</td>
<td>2859</td>
<td>Anaemia, unspec</td>
</tr>
<tr>
<td>27483</td>
<td>2809</td>
<td>Iron deficiency anaemia, unspec</td>
</tr>
<tr>
<td>25383</td>
<td>1623</td>
<td>Malg neo upper lobe, bronchus or lung</td>
</tr>
<tr>
<td>18323</td>
<td>1541</td>
<td>Malg neo rectum</td>
</tr>
<tr>
<td>17491</td>
<td>20280</td>
<td>Oth lymphomas unspec site, extranodal &amp; solid organ sites</td>
</tr>
<tr>
<td>17449</td>
<td>20410</td>
<td>Chronic lymphoid leukaemia w/out mention of remission</td>
</tr>
<tr>
<td>14308</td>
<td>2875</td>
<td>Thrombocytopenia, unspec</td>
</tr>
<tr>
<td>12589</td>
<td>1830</td>
<td>Malg neo ovary</td>
</tr>
<tr>
<td>11688</td>
<td>1985</td>
<td>2nd malg neo bone &amp; bone marrow</td>
</tr>
<tr>
<td>11303</td>
<td>1625</td>
<td>Malg neo lower lobe, bronchus or lung</td>
</tr>
</tbody>
</table>