National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

SEER Registries: Population-based infrastructure to support cancer research

Board of Scientific Advisors

December 1, 2015

Lynne Penberthy MD, MPH

Presentation Objectives

- Describe
  - SEER program
  - Challenges to cancer surveillance
  - New initiatives to address the challenges
  - Projects to expand SEER's capacity to support research
- Receive suggestions or comments from board members on strategic priorities for SEER

Surveillance Epidemiology and End Results (SEER)

The SEER Program is a national resource:

- Funded by NCI to support research on the diagnosis, treatment and outcomes of cancer since 1973
- Population-based registries covering 30% of the US population
- 400,000+ incident (newly diagnosed) cases reported annually

SEER Program

- Most commonly used data to represent trends over time
- > 4,000 downloads of SEER public use file annually
- 17,000 publications using SEER data since 1975
- ~40,000 manuscripts referencing SEER data
- 112 research grants ($87 million) funded in 2011-12 where SEER data was critical to the grant

SEER Program

- Only population-based system in the US that includes a broad set of clinical variables
- Variable selection driven by guidelines and standards
  - Current - 32 predictive & prognostic biomarkers
  - In Process - Guideline review to identify relevant new variables to be collected
    - EGFR/ALK lung cancer
    - BRAF/MSI Colon cancer

SEER Program

Data Completeness and Rigorous Quality Control

- Ongoing expansion of real time electronic pathology report feeds (360+ labs)
- Intensive visual editing of key data at the central registry level for accuracy across multiple reporting sources
- Optimizing methods to assure complete and accurate data through re-abstraction and focused review

SEER Program

Integration with NCI Cancer Centers

- SEER registries at NCI Cancer Centers
- SEER PI meetings focus on integration of cancer center research with SEER registries
- Leverage cancer center expertise in informatics
- Formal component of cancer center informatics cores (e.g., Fred Hutch, KY, New Mexico)

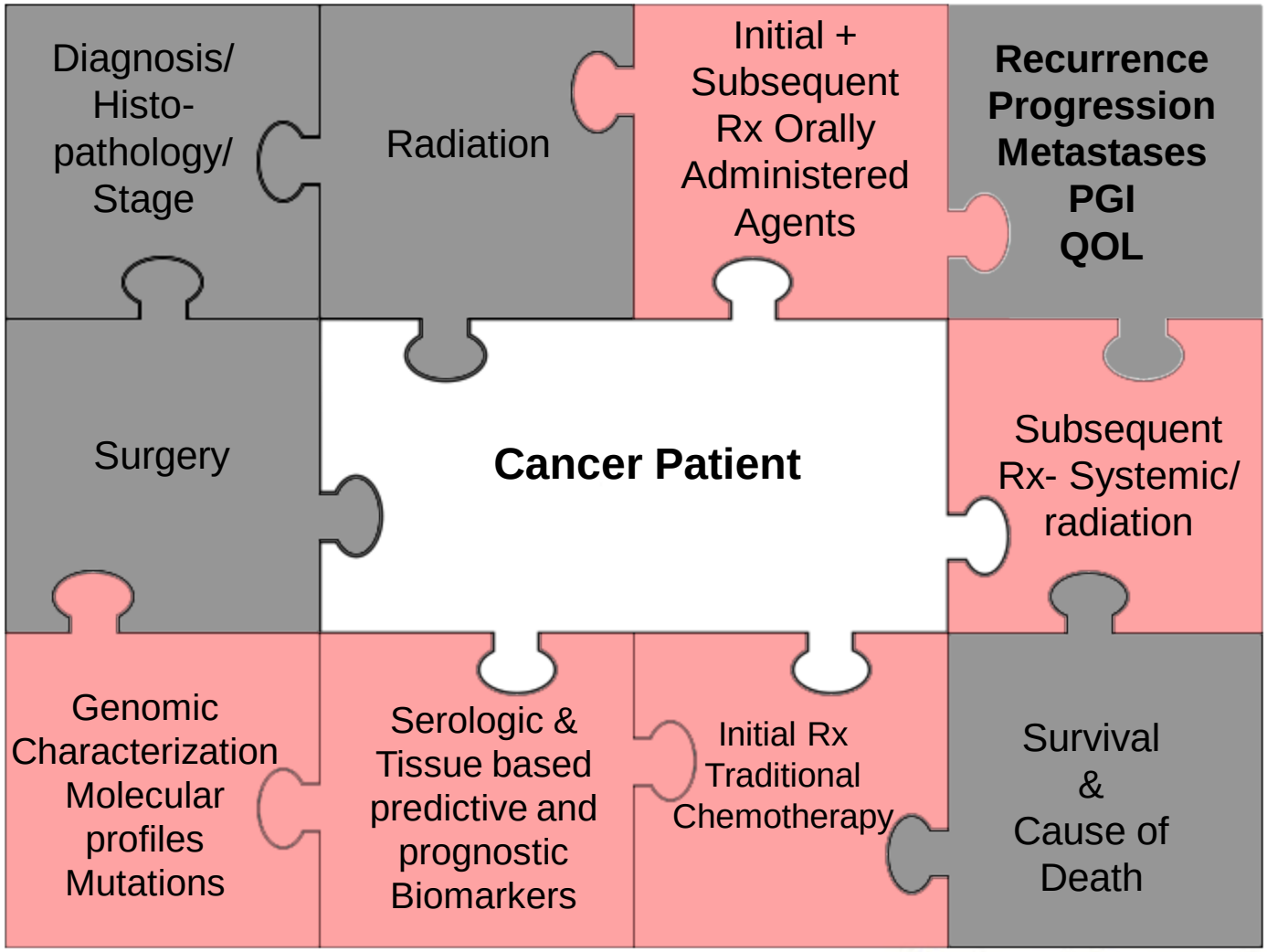
Cancer Surveillance Challenges

Data collection

- Complexity of cancer care
- Expanding data characterizing each cancer (precision medicine)
- Current manual processes for abstraction and data capture

Putting the puzzle together for each cancer patient: Data Collection

Diagnosis → Death



National Cancer Institute

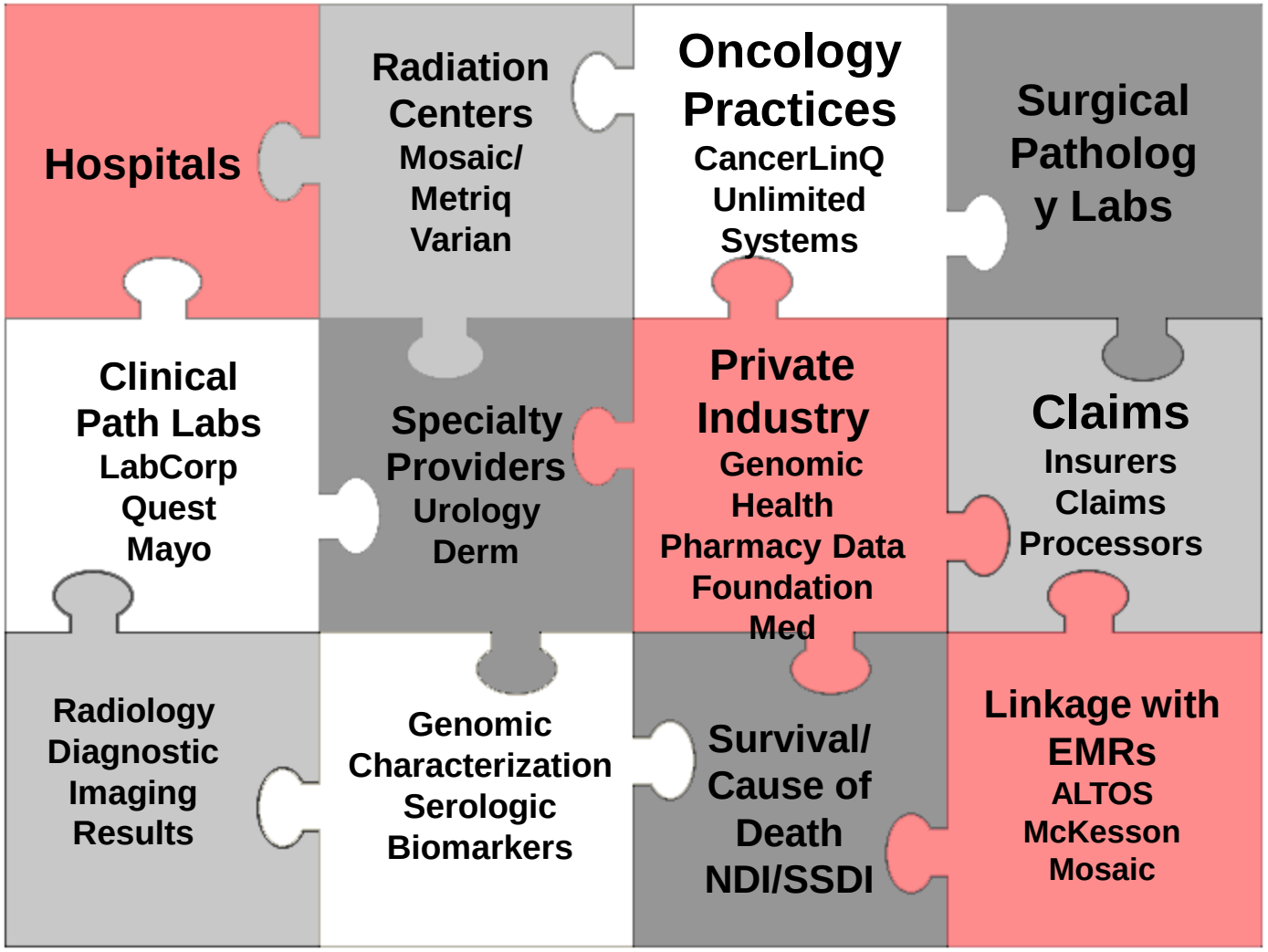
Cancer Surveillance Challenges

Data Sources

- Dispersion of cancer diagnosis and treatment across multiple health care providers/locations (no longer only hospital-based)
- Requires accessing information outside traditional registration sources
    - Pathology labs
    - Physician offices
    - Pharmacies
    - Freestanding integrated specialty practices

Putting the puzzle together for each cancer patient: Data Sources

Diagnosis → Death



Strategic Priorities for the SEER Program

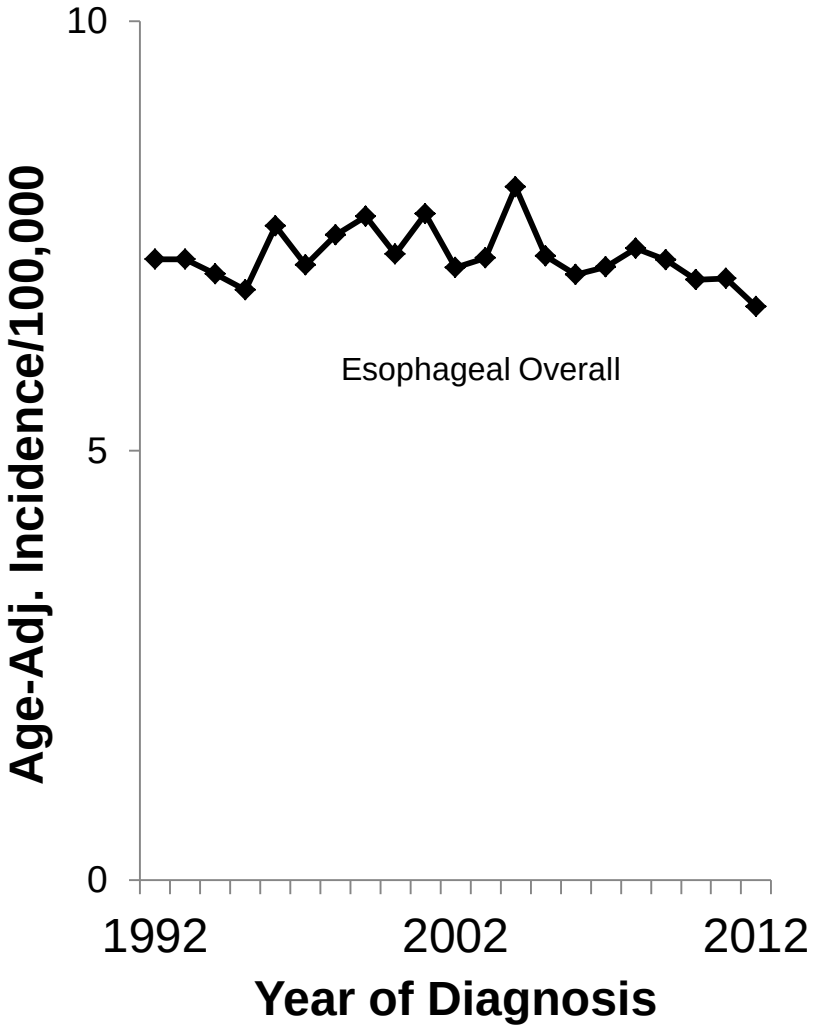
1. Represent data in more clinically relevant categories
2. Automate and directly capture data via
   - Linkages
   - Auto-processing of data (Natural Language Processing)
3. Expand outcomes data collection
4. Expand the capacity of SEER to support cancer research

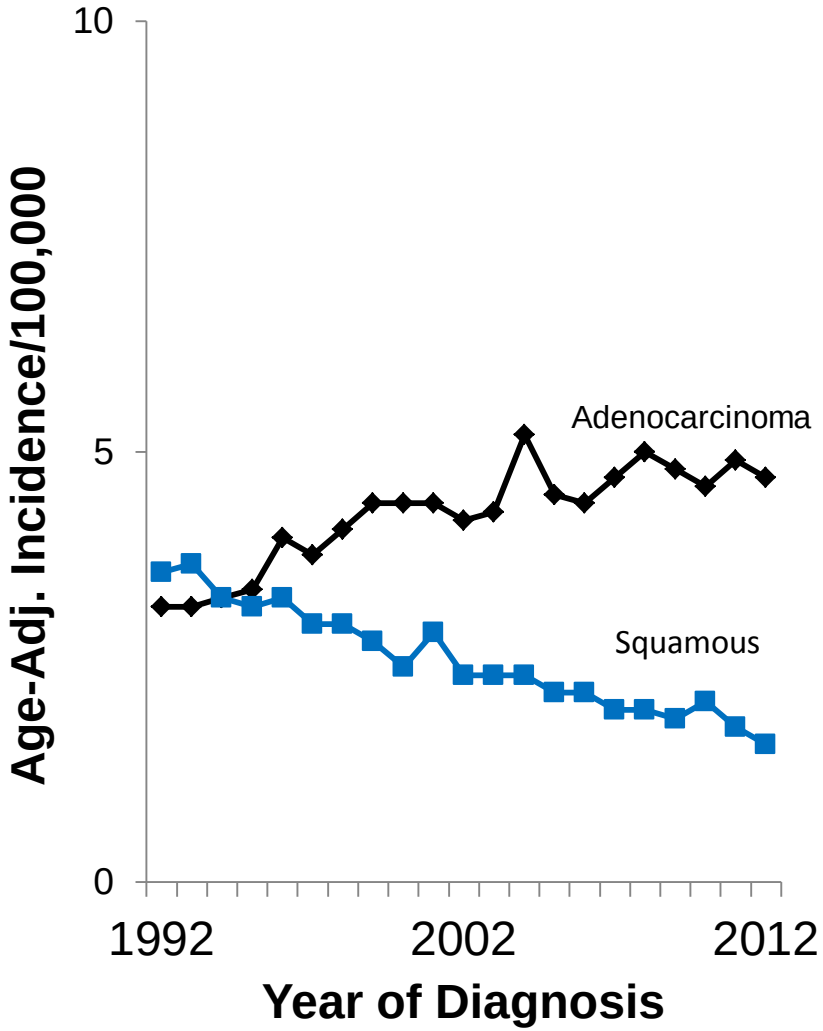
Represent data in more clinically relevant categories

Problem: Statistics by organ site do not represent cancer as it is currently understood and treated

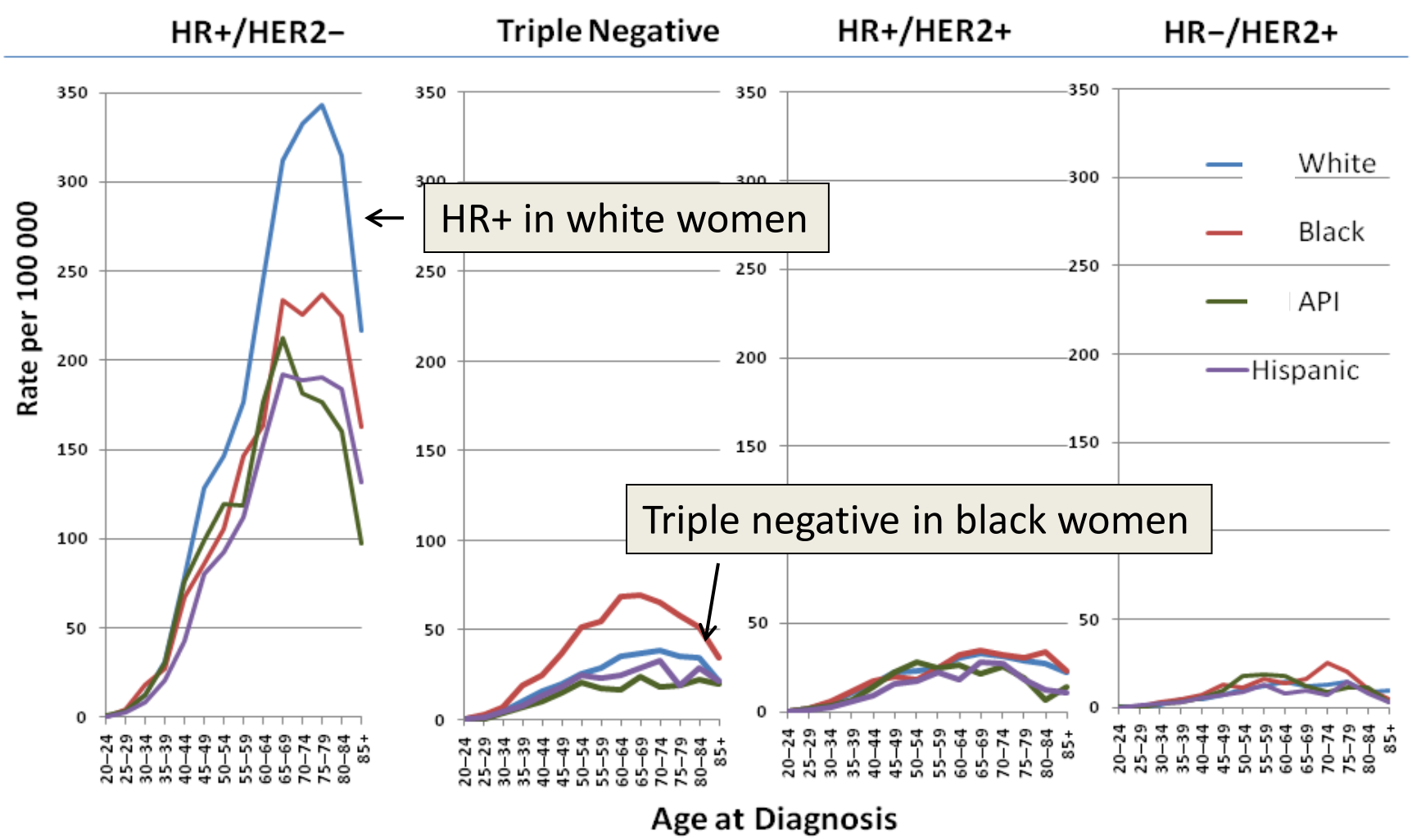
Solution: Present statistics by clinically relevant categories, e.g., histology, molecular characterization

Example: Reporting data in more clinically relevant categories - Esophageal Cancer in men - overall and by histologic subtype





Example: Reporting data in more clinically relevant categories - breast cancer incidence by subtype & race/ethnicity (2010)



Howlader et al. JNCI J Natl Cancer Inst April 28, 2014 [Epub ahead of print]

Automate and directly capture data via linkages - treatment

Problem: Lack of complete and detailed treatment

Solutions:

- Link with existing data for pharmacy-provided oral drugs
- Capture and process standardized claims for infusion therapy

Treatment Linkages: Oral agents

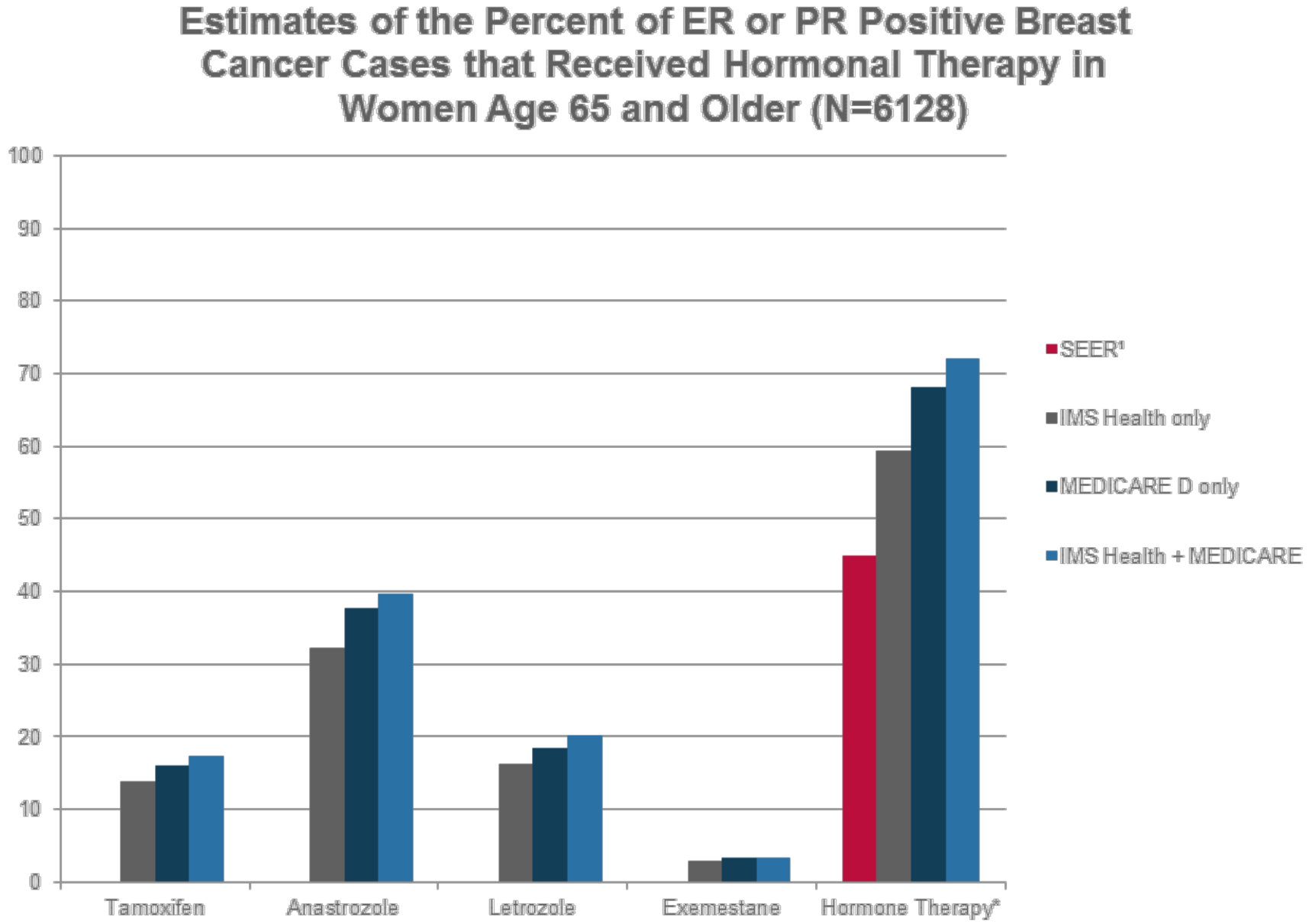
- 25%+ systemic Rx and growing
- No population based information (CTs data only)
- Capturing pharmacy data offers potential for
  - Supplementing treatment
  - Monitoring disparities in use and nonadherence
  - Identifying adverse events

Treatment Linkages: Oral agents solutions?

IMSHealth linkage

- 70% of pharmacy transactions
- Pilot study linking with 4 registries
- Comparing completeness with Medicare Part D, Patterns of Care, special studies for breast, colon, CML, MM

Preliminary results from IMSHealth Pilot Linkage



Treatment Linkages: Oral agents solutions?

Large pharmacy chain central repository linkages

- in discussion with Walgreens, CVS

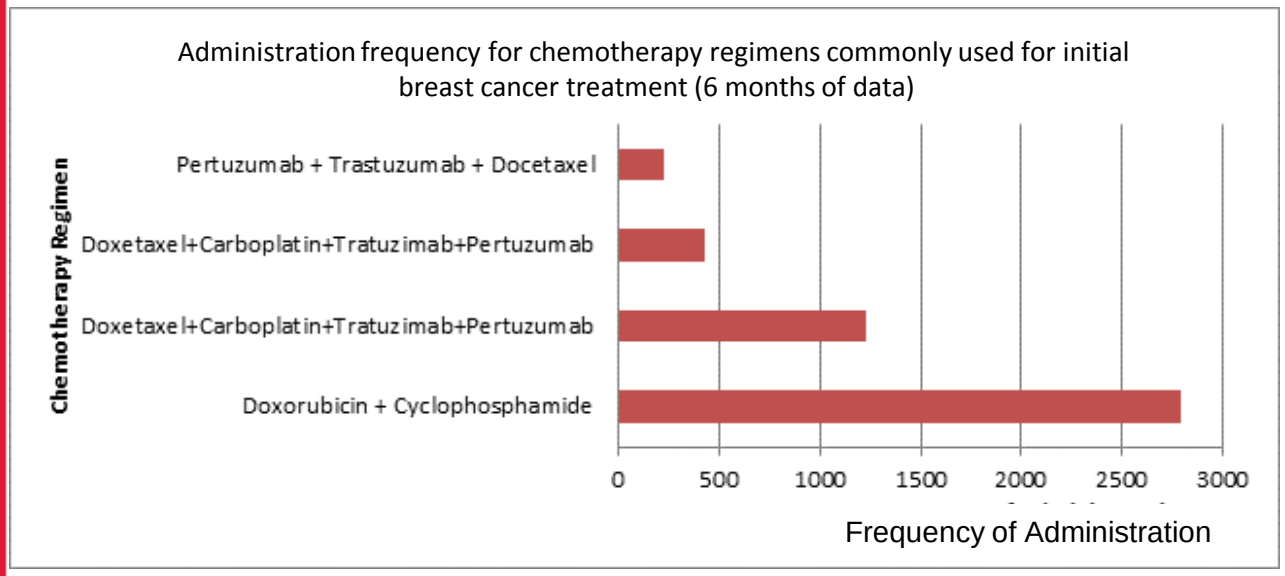
Link with pharmacy “switchers” (processors)

- Relay Health represents 75%+ pharmacy transactions
- Emdeon Data holds 20% of transactions

Treatment Linkages: Claims data for infusion Rx

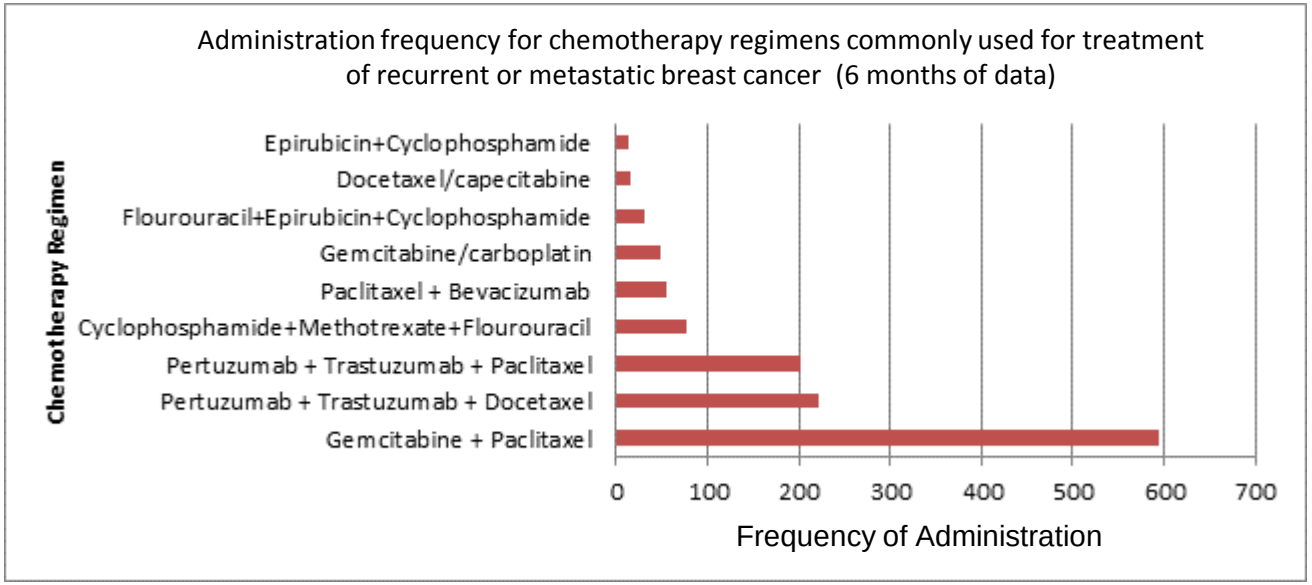
- Value of claims for treatment
  - Standardized format and nomenclature from all providers
  - High degree of accuracy and detail based on HCPCs
- Medicare
- Central claims processors (oncology)
  - Represent patient populations for all payers
  - A single central processor for 25-45% of oncologists within 7 SEER registries
  - Pilot in GA 12 oncology practices

Preliminary Data from 6 Months Claims in 4 Georgia Oncology Practices: Common Regimens for Treatment of Initial and Recurrent Breast Cancer



Common Regimens
Initial treatment of Breast Cancer
4 regimens - 4,676 administrations

Common Regimens
Treatment of Recurrent or Metastatic Breast Cancer
9 regimens -1,262 administrations



Automate and directly capture data via linkages – clinical data

Problem: Inability of registries to access relevant clinical test results

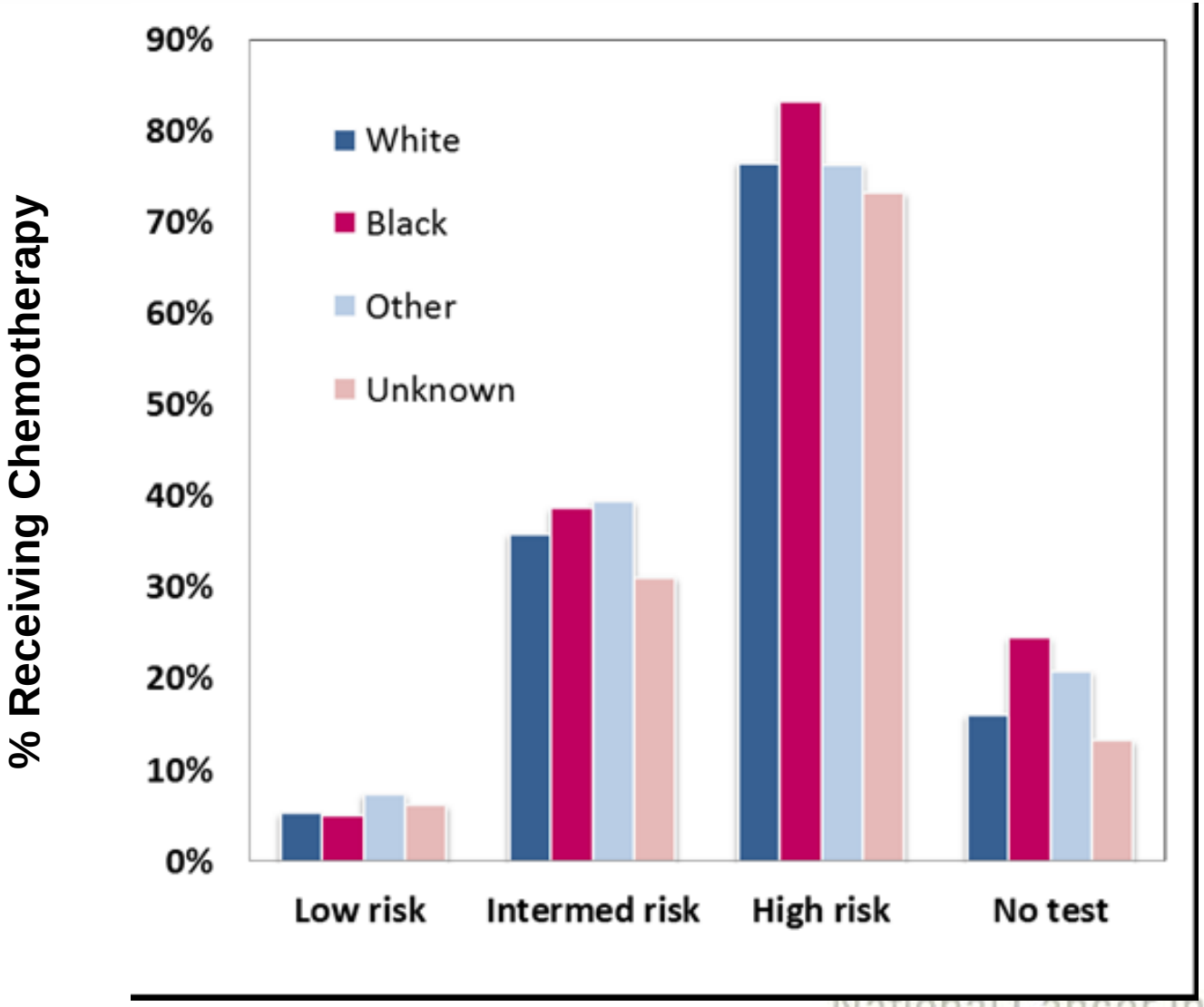
Solution: Develop partnerships with commercial entities who perform tests for direct data feed

Clinical data: Example of linkages with commercial partners

Oncotype DX: Linkage with GHI data 2004-2013

- Added 40% of test results to existing data from hospital reported results
- Largely test results sent directly to physician practices

Percent of Women receiving ChemoRx by Oncotype DX Risk Category and Race in SEER data (2010-2012)



Oncotype DX Risk Score Category & Chemotherapy among tested and non-tested women

Clinical data: Linkages with commercial partners, next steps

- Molecular signatures – Oncotype DX, Genome DX, Myriad (DCIS, Prostate)
- Foundation Medicine
- Linkage with BRCA panels for breast and ovarian (CA and GA)

Automate and directly capture data via auto-processing and natural language processing

Problem: Key data is in unstructured text

Solution: Leverage existing capacity across academic and commercial enterprise for natural language processing and data extraction

Automate and directly capture data via auto-processing and natural language processing

- Focus on unstructured electronic pathology reports
  - > 80% of cases with ≥ 1 report
  - Real time reporting
  - Multiple pathology reports per patient
    - Subsequent tissue based test results
    - Subsequent biopsies

Automate and directly capture data via auto-processing and natural language processing

Expand data not collected or available only in unstructured text

- Guideline indicated biomarkers
- Rapid inclusion of emerging biomarkers
- Monitor dissemination over time
- Metastatic disease (biopsy of recurrent lesion)

Leverage lessons learned to other unstructured text-radiologic imaging dictations

Expand Outcomes Data Collection

Problem: Survival no longer the only important outcome for cancer patients

Solutions:

- Leverage multiple data sources for disease status
- Engage the patient

Expand Outcomes Data Collection

Focus on better understanding the course of disease among > 15 million cancer survivors

Capturing Disease Progression/Recurrence

- Complex diagnostic patterns require multiple approaches varying by cancer site (e.g., NLP and serologic biomarkers)

Collecting Patient-Generated Health Information

- Working with partners to test solutions, e.g., patient portals, direct patient reporting, and patient-generated data sources

Expanding the capacity of SEER to support cancer research

SEER-Linked Virtual Bio-Repository

What is it?

- A ***virtual*** repository of SEER-based tissue with annotation
- Tool for researchers to search de-identified abstracts and linked e path reports to select a set of relevant specimens
- Ultimate aims
  - Annotation and search capacity of abstracts + e path reports for all SEER cases with tissue
  - Centralization of requests for specimens and custom annotation
  - Capacity for investigators to custom select relevant cases for their research

SEER-Linked Virtual Bio-Repository: Benefits

- Population based – permitting comparison of subsets
- Available across a broad spectrum of health care facilities/pathology labs (not just academic centers)
- Access to rare cancers and exceptional outcomes
- Linked long term outcomes
- Existing annotation with clinical and demographic data
- Potential for custom annotation
- Renewable with > 400,000 incident cases annually

SEER-Linked Virtual Bio-Repository Pilot

7 registries funded for pilot of pancreas and breast 9/15

- Focus on “exceptional” survivors
  - 431 early stage node negative breast cancer (< 2 yr survival)
  - 224 pancreatic adenoca long term survivors (> 5 yr survival)

- Purpose
  - Assess best practices across multiple registries
  - Estimate costs of supporting a SEER wide system
  - Assess availability of specimens
  - Understand human subjects/consent as requirements vary by registry and prepare for common rule changes

Virtual Pooled Registry with NAACCR and NPCR

What is it?

- A ***virtual*** national cancer registry
- Tool for researchers to automatically link patients with ***all US cancer registries***
- Ultimate aims
  - Automated linkage via Honest Broker
  - Centralized IRB
  - Return of patient information on cancers, survival, cause of death, treatment etc.

Virtual Pooled Registry

Who would benefit?

NCI with potential cost savings and enhanced efficiency of current linkage processes

- Cohort studies
- Follow up for Clinical Trials

FDA

- Post-marketing surveillance

Cancer registries

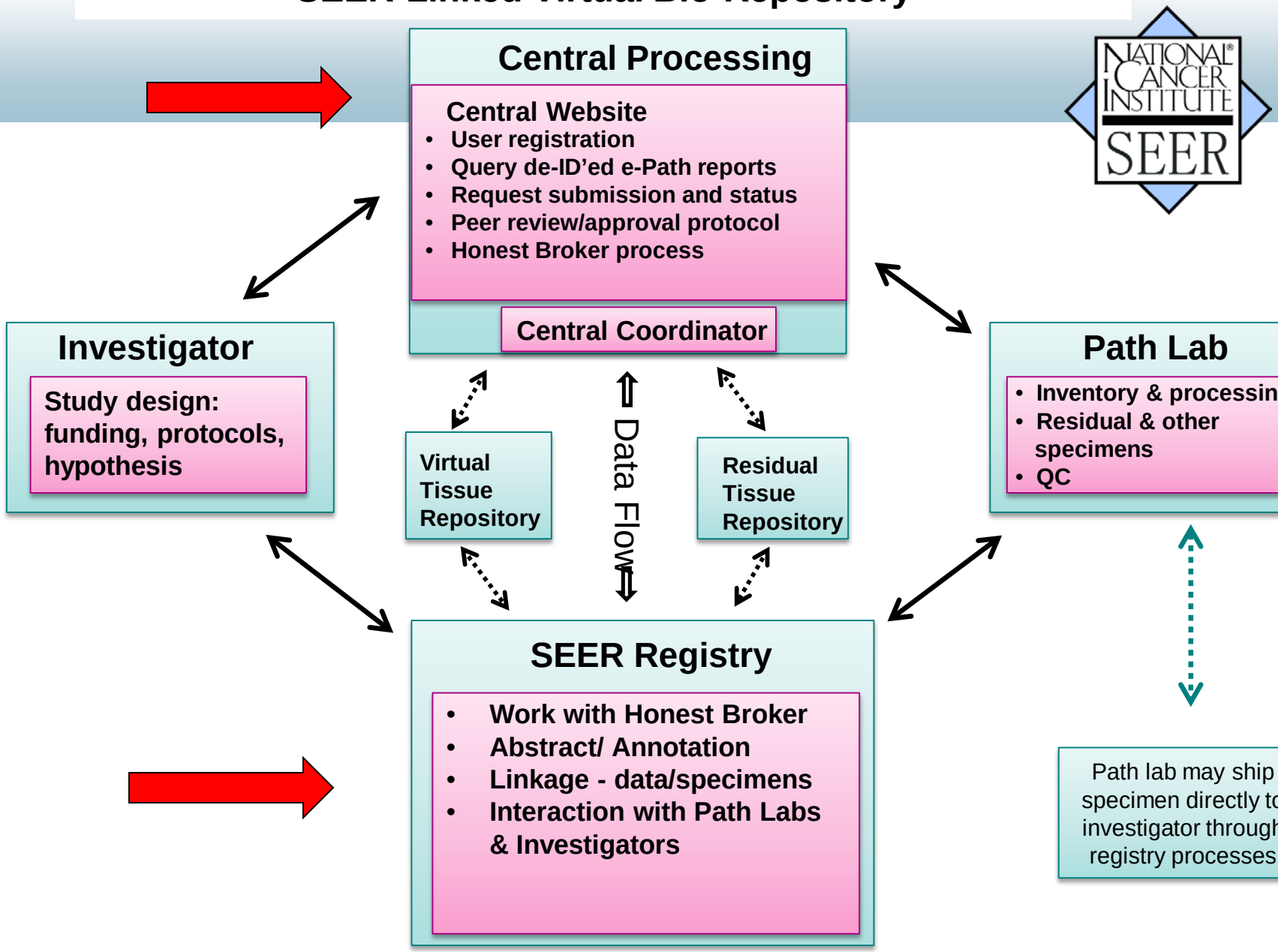
- De-duplication of cases
- Accurate assessment of multiple primary incidence

Strategic Priorities for the SEER Program

1. Represent data in more clinically relevant categories
2. Automate and directly capture data via
   - Linkages
   - Auto-processing of data (Natural Language Processing)
3. Expand outcomes data collection
4. Expand the capacity of SEER to support cancer research

Extra Slides

SEER-Linked Virtual Bio-Repository



Preliminary Data: Georgia Claims Pilot (6 mos)

SEER*DMS GCCS 17.0

Data Search

Actions

Apply | Reset | Save

SQL Search

SQL | Results | Filters

16 items.

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

Note this includes Only DX codes with Counts > 10,000