SEER DMS Face to Face Meeting
VPR update
September 26, 2018

NIH NATIONAL CANCER INSTITUTE

Background: The Need

- There is no nationwide registry that could be used to link with a cohort or clinical trial population
  - The current infrastructure consists of 50+ central (state and regional) registries
  - Linking for one cohort (Adventist Health) took approximately 3 years and required filling out 47 different IRB applications

- The National Cancer Institute and other Federal organizations support linkages in follow-up to many studies including: cohorts, clinical trials and other epidemiologic research
  - DCCPS alone provides support for follow up of >1.1 million participants in cohort studies
    - Conservative cost for follow up estimated to be $2.2 to $8.8 million per year
  - Other divisions support cohort studies, follow up of clinical trial patients etc.

Thus the need for a "Virtual Pooled Registry"

- A ***virtual*** national cancer registry that:
  - Permits linkages of patients (cohorts, clinical trials, other research studies) to ALL registries across the US
  - While maintaining patient identifiers behind registry firewalls
  - But permitting access to appropriately approved investigators (https://www.naaccr.org/about-vpr-cls/#Background)
- Ultimate aims are to develop a system with:
  - Automated linkage via an Honest Broker through a central website
  - A Centralized and/or templated IRB
    - Eliminating 50+ IRB applications and reviews
    - CIRB will be for minimal risk human subjects research
  - Rapid return of patient information on cancers, survival, cause of death, treatment etc. to the investigator

Pilot Linkages: NCI's US Radiologic Technologists

DCEG Cohort of 146K radiologic technologists (Rad Tech)

- Data collection methods include surveys every 10 years (4 total) to capture:
  - self-report of any cancer
  - medical record review of subset of cancers to assess the risk of radiation exposure on (specific) cancer risk

Linkage with 45 registries resulted in:

- Increased Case-Ascertainment
  - Rad Tech surveys (2003-2005) and (2012-2014) identified 11,396* self-reported cancers
  - VPR linkage with 45 registries: **(N=24,235) – more than double the self report**
- Greater Completeness of Cancer Data
  - Rad Tech Usual Method: 6 data Items collected only for subset from medical records
  - Data added from Registry Linkage: ***40+ data items on ALL cancer sites***
- Cost Efficiencies
  - Rad Tech time/resources for latest cohort survey: ~$1.28M cost to NCI
  - Total estimated charges to link with 45 registries: $58,000

*Number includes most recent surveys - detailed analysis in process for exact match of time intervals

VPR next steps

- Finalize analysis of Rad Tech cohort cost efficiencies
- Working with CDC to develop IAA to support VPR efforts to ensure that registries can support these activities in all 50 states
- Central IRB (CIRB) contract for minimal risk linkage studies in place 2019
  - Work with registries to utilize templated IRB
  - Negotiate acceptance of CIRB (22 of 45 willing to accept CIRB currently)
- Finalize enhancements to linkage software (Match*Pro)
  - Available for use for other linkages
- Our next pilot will be with the Pediatric Cohort Study from St Jude

Summary of Workshop in July

NIH NATIONAL CANCER INSTITUTE

Met with multiple registry PIs

Purpose:

- Identify potential barriers of the VPR scaled system
- Develop possible solutions to enable scaling
- Estimate costs for scaled VPR and discuss potential mechanisms to secure funding for registries to support a scaled VPR

Challenges

Opposing requirements for registries and investigators

- Investigators- must permit re use of data as part of NIH Data Sharing Policy
- Registries – no re-release of data (including linked data)
- What to do?

National Cancer Institute

Proposed solutions

- No re-release but support access by other investigators (beyond those with IRB)
  - De-identified
  - Support limited access (such as through dbGAP)
    - Which allows review and control
  - Automatically notify registries of additional investigators who access the data
  - Using cloud resources
    - Control access
    - Do not permit download of data only results (similar to CDC applications and NCHS data controls)
- Work with registries to assure reduced risk of re-identifiability
  - Combined age to age groups when numbers are small
  - Do not release geographic area (i.e. registry)
- Evaluate potential for use in releasing new SEER data
  - High volume
  - Many new and complex data might risk re-identifiability
  - Controls access by permitting analysis but not downloading all data

Update on Hashed Tokenized Linkage Process for De-duplication and MPC Estimation

NIH NATIONAL CANCER INSTITUTE

Challenge for registries in accurate case counts.

- Mobility of the US population varies from 5-11% of residents moving to another state each year
  - Variation by age, demographics and economy
- This mobility may result in
  - Lack of complete information on each cancer case
  - Duplication of case reporting
  - Inability to accurately assess multiple primary cancer incidence

- Contiguous states routinely perform data exchange for cases who are residents of the exchanging states
- But…
  - This does not provide de-duplication either among contiguous nor among states that do not routinely perform exchange
  - Nor does this capture cases or information where the patient provides as their residence in two different states over time

Importance

- We do not have accurate data assessing the magnitude of duplicate cases (especially for non-contiguous states)
  - Impacts both incidence and survival trends
- Nor do we have accurate estimates of multiple primary (MPC) incidence.
  - This is especially important in the era of precision medicine and genomic classification of tumors

Goal of national VPR matching

- Provide a secure system using hashed encrypted tokens (representing unique patient identification data) to:
  - Securely identify potential duplicate cases and multiple primary cancers
  - While maintaining registry confidentiality requirements

Initial pilot study results

- Three pilots to evaluate ability to perform encrypted tokenized linkages across multiple state borders
- Patient data encrypted/tokenized behind registry firewall
- Encrypted tokens
  - submitted to central linkage facility or
  - Linked by IMS
- Potential duplicates flagged and provided back to the individual registries for review and evaluation
- Each study used to
  - Improve accuracy of the tokenized matching process
  - Provide information for automated processing and adjudication of the matches where feasible
- Third study included 6 registries (Results pending)

Next steps

- Iteratively improve the hashed linkage process
- Develop algorithmic rules for deduplication and identification of multiple primary cancers
- Further test the software and system on a larger scale
- Ultimate goal to perform across all registries in the US

Thank you

NIH NATIONAL CANCER INSTITUTE