Linkage of Indiana State Cancer Registry and Indiana Network for Patient Care

A collaboration between Regenstrief Institute, Indiana University, and the Indiana State Cancer Registry
Objectives

- Understand the value of linking cancer registry data with an electronic health record (EHR) data repository.
- Learn how we determined the linkage algorithm.
- Can the linkage identify missing cases for the cancer registry?
Background

- Cancer registries are widely used to study the epidemiology of various cancers.

- Cancers captured in registries are well characterized around the time of diagnosis, including type, stage, and initial treatment.

- However, follow-up information is limited regarding subsequent clinical course and recurrence.

- Is it possible to link the cancer registry population to a state-wide health information exchange containing individual electronic medical records to enhance clinical research in cancer?
The purpose of this project was to develop and validate linkage algorithms to match the cancer cases in Indiana State Cancer Registry (ISCR) to medical records in the Indiana Network for Patient Care (INPC), for patients with a primary diagnosis date occurring between 2005-2012.
Aims

1. Test and select the most optimal matching algorithm between the ISCR and INPC.

2. Determine match rate for the linkage of patients diagnosed with specific cancers in the ISCR with INPC.

3. Test the completeness of ISCR in comparison with INPC electronic records based on additional validated cancer cases in the INPC.
Established in 1985 "for the purpose of recording all cases of malignant disease and other tumors and precancerous diseases required to be reported by federal law or federal regulation or the National Program of Cancer Registries that are diagnosed or treated in Indiana, and compiling necessary and appropriate information concerning those cases, as determined by the state department, in order to conduct epidemiologic surveys of cancer and to apply appropriate preventive and control measures."

Reporting for both providers and hospitals began on January 1, 1987.

INPC

- A unique federated electronic health record (EHR) data repository containing data collected from a large population across various healthcare settings throughout the state of Indiana.
- The INPC was developed by the Regenstrief Institute (RI) which further developed an accompanying clinical data environment to allow quick access and extraction of information from medical charts.
- Each participating institution provides common data elements, which can include inpatient admission/discharge information; outpatient visit information; laboratory values; microbiology, pathology, radiology, and cardiology reports; and clinical notes that can be analyzed via natural language processing.
INPC Data and Services

Data Management

Data Access & Use

- Results delivery
- Secure document transfer
- Shared EMR
- Credentialing
- Eligibility checking
- Results delivery
- Secure document transfer
- Shared EMR
- CPOE
- Credentialing
- Eligibility checking
- Results delivery

- Surveillance
- Reportable conditions
- Results delivery
- De-identified, longitudinal clinical data
- Secure document transfer
- Quality Reporting
- De-identified, longitudinal clinical data
- Pharmacoeconomics
Cohort Selection

Three cohorts were selected from the ISCR
1. Complete cohort
   o Encompassing all cancer patients
2. Melanoma
3. Lung

*Specific cancer cases were selected from the ISCR cohort by histology code. The SEER ICD-O-3 list categorizes melanoma as 8700-8799 (http://seer.cancer.gov/ icd-o-3/). Lung cancer cases were selected from the complete cohort by including all SEER ICD-O-3 codes C340-C349.*
Aim 1: Which linkage algorithm?

An attempt to match all eligible cancer cases from the complete ISCR cohort to the INPC was made using two different linkage approaches.

Compare the performance of 2 algorithms based on their Positive Predictive Value (PPV).

Two existing algorithms:
- Global Match – a deterministic linkage algorithm used by INPC to link new patient visits to the Global ID in the INPC Global Registry
- RecMatch – a probabilistic linkage algorithm implemented at RI
RI Global-Deterministic Linkage

- Runs daily on the INPC production database to link newly generated clinical data to existing patient records in the INPC master file.
- Considered a conservative deterministic algorithm.
- For this study’s purposes, Global Linkage made use of name, date of birth, gender, zip code, telephone number, and social security number whenever these data elements were available.
RecMatch-Probabilistic Linkage

A majority of patients in the ISCR had a value representing the medical record number (MRN) of the submitting institution, which should have very high specificity if matched to the MRN in the INPC.

Separate probabilistic linkage processes were run, based upon whether the institution and MRN matched between ISCR and INPC among all possible pairs from the two data sources.
Aim 2: Match Rate
Validation of Optimal Linkage Method between ISCR and INPC

- Pairs of identifiers from ISCR and INPC that were declared as matches by both the Global and MRN/probabilistic algorithms were considered true matches.
- Pairs declared as matches by one algorithm, but not another, were manually reviewed by two reviewers to determine the “true” match status.
- Medical record review was used as the “gold standard” for evaluating the performance of the linkage algorithms.
Evaluation of the Performance of Linkage Algorithms-Match Rates

Cohorts

Complete

- 2005-2013, a total of 202,153 cases were in ISCR from institutions reporting to INPC.
- Overall, a total of 172,895 ISCR cases could be matched to the INPC using either of the two algorithms, resulting in an overall match rate of 85.5% for the complete cohort

Melanoma

- Match rate of 94.4%

Lung

- Match rate of 84.4%
Aim 3 Results: Completeness of ISCR

**Melanoma** – 9043 in INPC, of which 3083 (34.1%) linked to ISCR chart review (sample n=199):
- 44 (22%) were confirmed as correct,
- 46 (23%) could not be confirmed as having melanoma within the date range of interest
- 109 (55%) had melanoma prior to date range

**Lung Ca** – 21,259 in INPC, of which 13593 (63.1%) linked to ISCR chart review (sample n=200):
- 15 (7.5%) were confirmed as correct,
- 107 (53.5%) could not be confirmed as having lung Ca within the date range of interest
- 78 (39%) had lung Ca prior to date range
Discussion

- The linkage of a state cancer registry with EHR data leverages longitudinal, electronic data which documents care delivered to the general population served by several community-based health care institutions.

- EHR data linkages hold the promise of generating knowledge about cancers more common in younger populations, e.g. testicular cancer, thyroid cancer, lymphoma, and leukemia.

- Compared to administrative claims, EHR data also has the potential to provide more clinically detailed information, such as the results of lab or imaging tests, than the event-based billing information available in insurance claims.
Overall match rate of 88.5% is encouraging, suggesting information about longitudinal, follow-up care may be ascertained among a significant proportion of cancer patients shared between the ISCR and INPC.

Cancer control covers the continuum of care from prevention to end-of-life care. Merging both data repositories has the potential to create a unique resource for many types of epidemiologic studies and clinical research topics.

Trade-offs existed between deterministic and probabilistic algorithms. The probabilistic algorithm identified more matches, but the deterministic algorithm had a higher PPV.

Ultimately, both were implemented as the PPV associated with both was quite high. For the purpose of longitudinal, epidemiologic cancer control studies, this threshold is still determined to be reasonable.
Conclusion

- By linking the ISCR with the INPC, the ISCR is able to identify missing cancer cases.
- Identification of any missing cases in ISCR adds value to the overall accuracy of the ISCR, ensuring proper incidence and mortality can be assessed and targeted approaches for cancer control can be implemented across the state.
- We can ascertain that for epidemiological studies based on large databases such as a HIEs and EHRs, case identification using cancer registries that can be linked to EHRs will provide definitively diagnosed cancer cases with the added advantage of rich data on treatment, disease progression, and outcomes.
Conclusion, continued

- Most, but not all, patients with specific cancers identified by ICD-9 codes in the INPC could be linked to the ISCR.
- Among those who could not be linked, about half were found to be false negatives from the registry perspective, i.e., a cancer was present based on manual review of their EHRs in INPC.
- The potential of a HIE to capture cancer cases in real time, especially cases that are not otherwise identified by the state cancer registry, suggests future models for disease surveillance using EHR data.
Thank you to all of the contributors!

David Haggstrom, MD, Principal Investigator, VA Health Services, Indiana University School of Medicine, Regenstrief Institute, Inc.

Siu Hui, PhD, Co-Principal Investigator, Regenstrief Institute, Inc.

Paul Dexter, MD, Co-investigator, Indiana University School of Medicine, Regenstrief Institute, Inc., Eskenazi Health

George Eckert, Biostatistician, Indiana University School of Medicine

Fangqian Ouyang, Biostatistician, Indiana University School of Medicine

Joel Martin, Data Analyst, Regenstrief Institute, Inc.

Abby Church, MPH, Project Coordinator, Regenstrief Institute, Inc.

Laura Ruppert, MHA, Cancer Surveillance Section Director, Indiana State Department of Health

Jinghua He, PhD; Merck Project Lead
Acknowledgements

- This research was funded by a grant under the Merck-Regenstrief Program in Personalized Health Care Research and Innovation, a collaboration between Merck, Sharp & Dohme and the Regenstrief Institute.

- The journal article, that this presentation is based on, was supported by DP003884 funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

- Population data was provided by the Indiana State Cancer Registry which is funded by the Centers for Disease Control and Prevention under cooperative agreement 5 U58DP003884-04-00.