Using an Online Database Resource to Characterize Healthcare Data Linkage Capabilities

Anokhi J Kapasi, Sharmila A Kamani, **Judith K Jones**
DGI, LLC, Arlington, VA, USA

BACKGROUND

- **Linkage of data elements from multiple data sources** is becoming essential to epidemiology and health outcomes research, and allows query in a broader, more **diverse** data set, ideally with **granular** information.
- Separate from the complexities of extracting and merging data, it is important to note that there are many types and methods of data linkage and levels of data that can be obtained, many of which currently lack proper descriptive and operational definitions.
- OMOP's efforts to standardize data terms and to develop common data models (CDM) lend well to data linkage processes.
- **B.R.I.D.G.E. TO DATA®** (www.bridgetodata.org) is a centralized compendium of population healthcare database (DB) profiles worldwide that utilizes **standardized data fields** (Table 1) to describe the types of information captured within a DB, including data linkage capabilities.
- The structure of B.R.I.D.G.E. profiles can complement OMOP's CDM. The profiles contain 75 standardized data fields (Table 1), which may be mapped to CDM fields and concept tables. E.g., Drug Information maps to CDM Drug Exposure table, and drug generic name, dosage, days supply, coding system, can map to CDM fields such as drug_concept_id, refills, quantity, and days_supply.
- One major application of B.R.I.D.G.E. is to allow **comparison of data across multiple data sources**; therefore, it can be a useful tool in identifying DBs where CDM fields can be applied.

Table 1. Examples of Data Fields Used in Profiles (by Category)

Category	Data Fields
Summary	Database description, Database source, Years covered, Population type, Date of last update
Population Dynamics	Population size, Sample weights – Extrapolation factors
Demographic Data	Age, Gender, Date of birth, Death recorded, Other demographic data
Physician & Practitioner Info	Physician ID & Specialty, Pharmacy ID
Diagnoses/Signs & Symptoms	Diagnosis data, Diagnoses coded (coding systems), Max. number of codes, Physical exam findings, Environmental exposures, Behavioral data elements
Procedures	Procedure data, Procedures coded (coding systems), Laboratory information
Drug Information	Drug data, Drug dosage, Drug coding system(s), Additional drug information
Economic Data	Type of cost data (if applicable)
Validation & Linkage	Data validation, Access to medical records, Linkage to other databases
Administrative Data	Database contact data, Database usage restrictions, References of studies using/describing the database

LIMITATIONS: This analysis was done using DBs currently profiled within B.R.I.D.G.E. TO DATA®. More profiles of data sources are continually being added to this resource.

OBJECTIVE

To identify and define the types of DB linkages possible within or across various healthcare DBs and to describe the potential for CDM mapping across linked data sets.

METHODS

B.R.I.D.G.E. was used to identify DBs with data linkage capabilities by:

(1) A keyword search with **'link'** to identify various types of data linkages.

(2) A search with (criterion) **'Cross-sectional Population Databases'** AND (keyword) **'longitudinal'** to identify DBs with records linked across survey periods.

Out of 225 profiles as of 10/30/13, the searches resulted in 163 unique DBs. After manual screening of the search results, 31 DBs were excluded due to no data linkage capabilities. The remaining 132 DBs were reviewed for data linkage characteristics, which included type of data sources being linked, type of data being accessed via linkage, and variables used in establishing the linkage.

RESULTS – Part 1

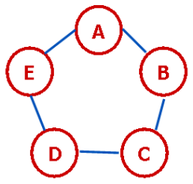
The set of 132 DBs had the following non-exclusive characteristics: 105 (80%) DBs directly linked to another DB (Figure 1A), 30 (23%) had indirect linkage capabilities (Figure 1B), and 41 (31%) were formed through DB linkages (Figure 1C). The primary linkage methods were using a unique ID or probabilistic matching at the patient level; however, other linkages also exist, e.g., encounter-level linkage.

Figure 1. Examples of Database Linkage Capabilities

A. Direct Linkage (n=105)

A.1 Direct Linkage (DB 'A' links to DB 'B')

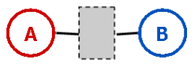
- Korean Health Insurance Review Agency (HIRA) DB links to Korea Central Cancer Registry via unique personal ID
- Multiple Risk Factor Intervention Trial (MRFIT) links to National Death Index (NDI) via unique personal ID



A.2 Multiple Direct Linkage (Network of linkages across DBs 'A' through 'E')

- All inhabitants in Denmark are registered in the Danish Civil Registration system with a unique 10-digit national PIN. The PIN is used to link Danish registries, such as the Danish Cancer, National Hospital Patient, National Pathology, and Death Certificate registers, allowing follow-up from birth throughout life
- Manitoba Population Health Research Data Repository via patient ID

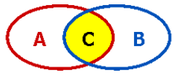
B. Indirect Linkage (n=30)



B. (Linkage of DB 'A' to DB 'B' requires an extra step)

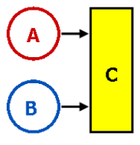
- Icelandic Cancer Registry needs approval prior to linkage via national PIN to Cause of Death Registry

C. Formed by Linkage (n=41)



C.1 Combination of Database Subsets (DB 'A' subset links to DB 'B' subset to form new DB 'C')

- SEER-Medicare Database (USA) is a linkage of SEER cancer registry data, and the Medicare enrollment and claims files via patient insurance ID and physician ID



C.2 Merged Databases (DB 'A' merges with DB 'B' to form new DB 'C')

- North American Research Committee on Multiple Sclerosis (NARCOMS) Registry formed by multiple regional MS registries
- AIHW National Diabetes Register (Australia) formed by the National Diabetes Services Scheme database (NDSS) and the Australasian Paediatric Endocrine Group's (APEG) state and territory databases via registration ID

RESULTS – Part 2

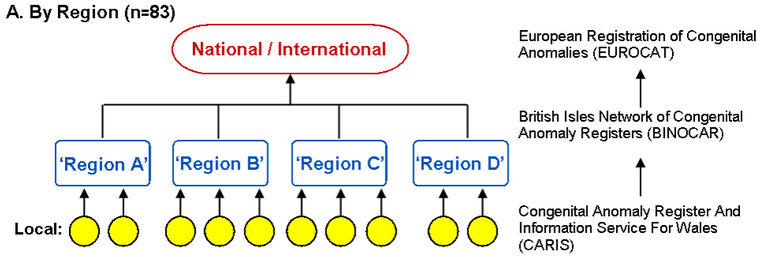
The most common patterns included linkages by:

- Type of health services, e.g., prescription, diagnoses, and hospitalization data (80; 61%);
- Region, e.g., national registers (83; 63%) (Figure 2A);
- Health status (64; 48%) (Figure 2B);
- Vital statistics (48; 36%) (Figure 2B); and
- Civil information, e.g., government administrative DBs (43; 33%).

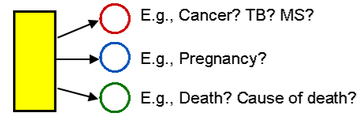
Some of the less common linkages were those by institution practice type, across survey years (Figure 2C), or study cohorts.

Figure 2. Examples of Types of Data Linkage Themes Across Healthcare Databases

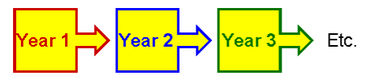
A. By Region (n=83)



B. By Health Status (n=64) or Vital Status (n=48)



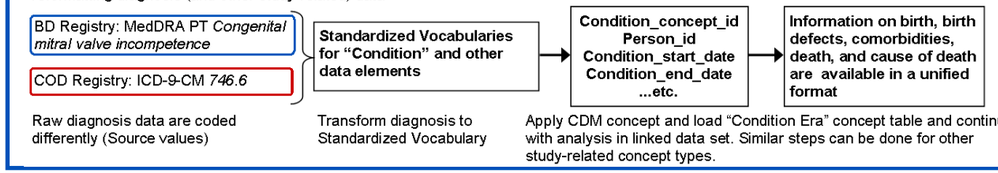
C. Across survey years/Longitudinal Cross-sectional



Data elements obtainable via linkage varied, but frequently included data on birth & death, cancer, hospitalizations, and prescriptions. Use of common terminology may be helpful. Figure 3 is a schematic showing how OMOP standardization and CDM formatting can be applied to source data prior to evaluation of linked data.

Figure 3.

Example: Linkage of a birth defects registry and cause of death registry to assess patients with congenital mitral valve insufficiency requires reformatting diagnosis (and other study-related) data



CONCLUSION

(1) Concurrent assessment in multiple data sources is important as a single data set is typically not sufficient to meet all outcome analysis requirements. This study highlights a growing number of databases with data linkage capabilities and defines linkage patterns. Specifically, 59% of the profiles in B.R.I.D.G.E. describe data linkages. The most frequent are to regional or health services DBs; common data elements obtained are on vital status and cancer data.

(2) One of OMOP's aims is to enhance estimates of association between treatment and outcome across multiple disparate observational data sets. In doing so, a CDM is being generated. The detailed profiles describing coding in B.R.I.D.G.E. facilitates mapping the data to OMOP CDMs. The next step in this study of data linkages would be to catalog further data elements to coordinate with the developing granular features found in the CDM.

This study was self-funded by DGI, LLC, a non-profit organization. Contact: info@bridgetodata.org

OMOP-IMEDS 2013 Symposium. Bethesda, MD. November 5-6, 2013.