Linking HDSS data to health services data: Challenges and opportunities

Dickman GARETA, Siyabonga NXUMALO and Kobus HERBST

Africa Centre for Population Health

INDEPTH Scientific Conference, 11-13 Nov 2015
Overview

- Background
- Research setting
- Data sources
- Methods
- Data linkage process
- Results
- Challenges
- Opportunities
- Conclusion
Background

- HDSS collect huge volume of data - adequately enough?
- Increasing interest in using data linked from multiple sources - 6-fold increase in the last two decades (Bohensky et al, 2010)
- Data linkage centres have been established globally
  - >60 in Australia, 23 in Canada, 5 in UK, 1 Germany,
  1 New Zealand, 2 Austria and 7 in Brazil
Background

- Data linkage (DL) provides:
  - Wide coverage and scope of data
  - More complete info of individuals
  - Most cost-effective ways of supporting research (WT Report, 2015)
Setting

- High poverty
  - Unemployment (~67%, 2010)

- High disease burden
  - 28% adult HIV prevalence

- 200-bed district hospital outside the DSA

- HIV care is provided in the subdistrict through 17 primary health care clinics
Data sources

- HDSS research data
- Hospital information system
- Electronic ART record system (ARTemis/Tier.Net)
- Electronic routine lab results (NHLS)
- Electronic pharmaceutical data (iDART)
- Electronic TB Register (ETR) (in progress)

There is a need to link all these disparate sources - that is complex and time consuming and you need an effective and efficient work flow to do it routinely
Data linkage levels

- ART data (Tier.net) with ART data (Tier.net)
- ACDIS and ART data
- ACDIS and hospital data

Data from the same individual can appear in multiple systems and the ability to integrate this data depends on accurate record linkage.
Deterministic data linkage

Largely use national id to link data sets

- 77% (Resident & age > 15) have valid SA Ids in DSA

- 81% (n=30,129, ART initiators) ARTemis

- 57% (n=41675) TIER.Net
Deterministic data linkage

- Combined variables are used as linkage key
  - Individual names + DOB + sex

**Validation:** Mother/father names, location and closest clinic
Probabilistic data linkage

- Common/unique ids rarely available
- Pentaho Data Integration’s JaroWinkler Similitude is used
- Weighted score for every possible pair of records
  - Returns jaro similarity coefficient between two strings
  - \([\text{SurnameMatch}]+[\text{FirstNameMatch}]+[\text{SexMatch}]+[\text{DoBScore}]\)=Weighted Score
  - A weighted score greater than or equal to a cut off point means match

**Validation:** Mother/father names, location, closest clinic
Data linkage process

Data set A
Data cleaning and standardisation

Data set B
Data cleaning and standardisation

Record pair comparison
Classification
Non matches
Possible matches
Blocking
Matches
Manual check
HDSS data linkage with ART research data

- **HDSS database**
  - 92,000 (62,160 resident) currently
    - ~1.3 million person years
  - 10,000 resident adults (age > 15) consent for HIV annually
  - 43,307 ever tested, 11,763 positive

- **ART database**
  - 30,129 initiated ART (Include individuals outside DSA)
  - Of the HIV+ in HDSS, 4,049 (34%) linked to HIV database
HDSS data linkage with hospital info system data

Hospital information system (HIS) database

Serves a population=247,350 (2011 SA Census)- (37% from DSA)

-21,881 individuals hospitalized (Include individuals outside DSA)

-2,766 linked to ACDIS

HIS database with ART database

-4,942 are linked to ARTemis
Challenges

- Errors, variations and missing data on the info to link the datasets
- Differences in data captured and maintained by different databases
- Data dynamics as data regularly change over time
  - Names and addresses
- Data problems
  - Typos/misspellings (algorithm)
  - Incomplete words
- Methodological issue such as validation of linkages
Opportunities

- Data linkage helps:
  - Improve data quality and integrity
  - Reduce costs and efforts in data acquisition
  - Allow re-use of existing data sources for new studies
Conclusion

To fill the gaps in the existing data infrastructure, additional steps that integrate HDSS data with other data sources are required to improve data quality.

- Effective and efficient work flow to do it routinely.

Without such an effort, HDSS are likely to continue to build on fragmented and often costly system with limited access.

In the end, the potential for HDSS data will not be fully realized.