Record-Linkage Studies

Marwan Khawaja

UN-ESCWA

Beirut, Lebanon

Background -

- Evaluating completeness and coverage of mortality statistics traditionally rely on indirect methods

- Indirect methods (Generalised Growth Balance Method developed by W Brass in the early 1970s) use census and registration data jointly but are loaded with assumptions – accurate census enumeration, no net migration and constant incompleteness rate by age.

- Results are usually used for life table construction but not routine measures of mortality

- Remain controversial - assumptions & results

- Ken Hill (2004-5) & Tony Barnes sensitivity tests (2008)
  - Different methods give different results
  - Results are sensitive to errors in the estimates of the size and the age structure of the population

- Preliminary results on data from several Arab countries give different estimates of incompleteness

Alternative: Record linkage and direct methods

- The process of linking data from two or more data sets
  - Mostly used for linking records from census and registration data
  - Has a long history – with many uses
  - Case re identification in capture re capture studies
  - Data cleaning –duplication of records etc.
  - Causal inference – matching studies
  - More recent interest in the ‘big data’ world

- Method pioneered by **Fellegi-Sunter** (JASA ‘69)
- Method is challenging: Two types of data linkage errors
  - Mismatches and non-matches
- More recent advances & Algorithms available in widely used software

Two main methods

- Deterministic record linkages
  - AKA, exact matching
  - Linking on first and last name: marwan khawaja, marwan khawajah, marwan khawajeh
    marwan khawaja, marwan khawaja, marwan khawaja
  - Only two outcomes: matched/unmatched

- Probabilistic record linkage
  - More recent
  - % of string characters matched perfectly
  - Determining threshold (%) for matched status is up to the researcher
  - Algorithms & assigning agreement weights are widely available

- Causal inference studies – known as propensity score matching

Consistency index

- Records are linked in two data sets (census and registration) to assess consistency in refugee identification

	Register		
Cansus	Refugee	Non refugee	Total
Refugee	A	B	C
Non refugee	D		
Total	G		

Overall consistency index = A/(B+D+A) (refugee in both as % of total)
Consistent index for Census = A/C (refugee in both data sets as % of refugee in census)
Consistent index for Register = A/G (refugee in both data sets as % of refugee in Register)

1 = totally consistent

0.5 = number of consistent linked records the same as the number of inconsistent linked records

0 = totally inconsistent

Some challenges -

- Coverage problems in census & registers
  - ..as different from completeness
- Other data quality problems – for data used in matching:
  - Missing data on key variables (names, residence, date of birth, etc.)
  - Assumes fully updated registers – but often no the case (e.g., residence)
- Which source to use the master file (superior data set)?
- Inconsistencies in linked records can be substantial – and lead to biased estimates
- Privacy/ethical issues – e.g., census of Palestinian refugees in Lebanon

Next -

- EGM planned for the Fall, 2017
  - Needed Data from at least 2 countries to undertake a study

Thank you!