Australian
National
University

Record linkage studies to assess completeness of death registration

Chalapati Rao

Department of Global Health

Research School of Population Health

Outline

- Background
- Statistical methods
- Assumptions
- Typology of study designs / historical overview
- Examples
  - Viet Nam, Indonesia
  - Other countries
- Limitations/advantages of record linkage studies

Background

- Record linkage studies are being considered as an alternative to indirect demographic techniques to measure completeness of death registration
- Involve linkage of records across different data sources, and are also referred to as dual record system studies; or matching studies
- Record linkage can be used for reconciling data across different sources, and as a basis for dual record system (DRS) analysis to estimate completeness
- DRS method can be defined as a method for estimating total population size (total deaths) when a full count of the total population is unavailable or unfeasible, but when there are two or more independent sources of information on individual members of the population

Conceptual basis

- Individuals are ‘captured’ from their record in one data source and ‘recaptured’ when the record for the same individual is matched in the second source

- Matching across key variables:
  - Personal details (UID/Name/age/sex)
  - geographical variables
  - Event details - Date of birth/death/registration

- Linkage produces 3 sets i.e Matched records; plus sets of unique records in either source

- Linkage allows data reconciliation to derive a larger set of empirical records than from either source

Conceptual basis contd

- Completeness of either source could be computed as a proportion of the total reconciled deaths

ALSO

- record linkage permits the application of another statistical procedure (based on certain conditions) to estimate deaths not captured by either source

- This estimate of missed events added to the reconciled deaths to derive an estimate of total deaths

- Subsequently, completeness of either source derived as a proportion of deaths recorded in it out of the estimate of the total deaths
- Other ‘hybrid’ models for estimating completeness, involving multiple data sources/partial data sources etc

Computation

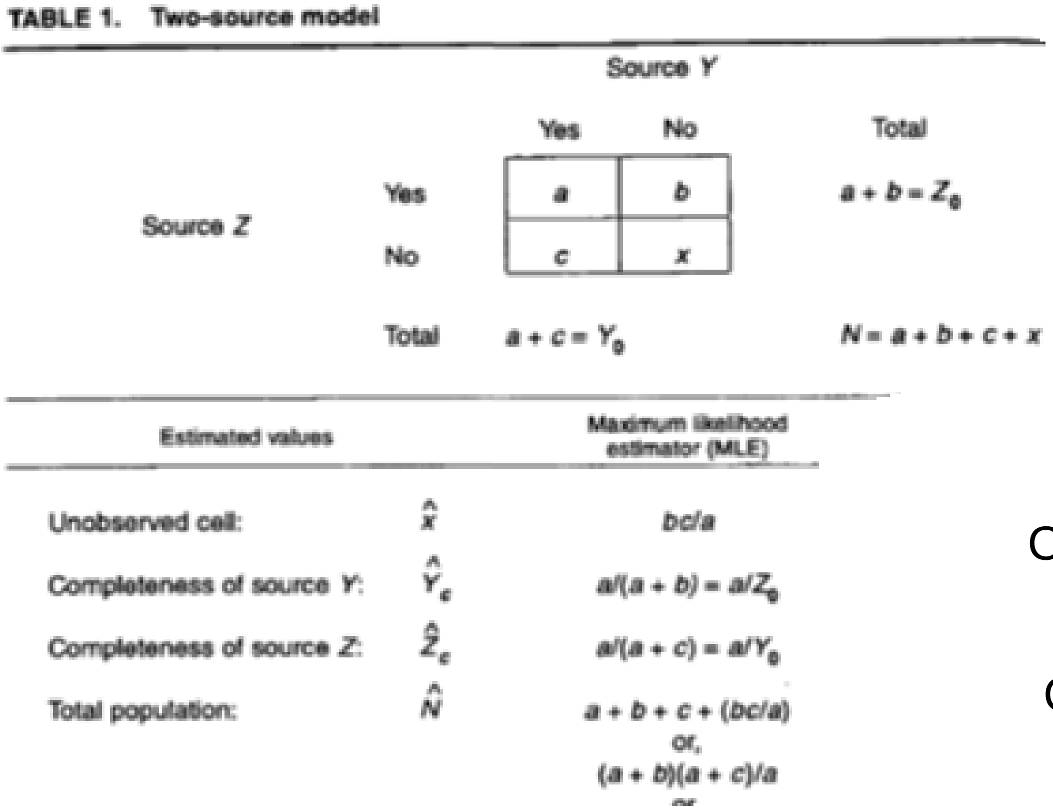


TABLE 1. Two-source model

		Source *Y*		
		Yes	No	Total
Source *Z*	Yes	*a*	*b*	$a + b = Z_0$
	No	*c*	*x*	
	Total	$a + c = Y_0$		$N = a + b + c + x$

Estimated values		Maximum likelihood estimator (MLE)
Unobserved cell:	$\hat{x}$	bc/a
Completeness of source *Y*:	$\hat{Y}_c$	$a/(a + b) = a/Z_0$
Completeness of source *Z*:	$\hat{Z}_c$	$a/(a + c) = a/Y_0$
Total population:	$\hat{N}$	$a + b + c + (bc/a)$ or, $(a + b)(a + c)/a$ or

Completeness of Y = $\frac{a+c}{a+b+c+x}$

Completeness of Z = $\frac{a+b}{a+b+c+x}$

- Hook, E.B. and R.R. Regal, *Capture-recapture methods in Epidemiology: Methods and limitations.* Epidemiologic Reviews, 1995. **17**(2): p. 243-64.

Conditions for DRS methods

- No ‘out-of-scope’ events in either source
  - All cases in each source are correctly diagnosed (true deaths)
  - All cases from each source are in the correct and same time-space frame
    - year of death/ address
    - Correct application of definitions of residence status
    - Study population is closed (no in/out migration)
- Homogeneity of capture probability in each source (in each data source each individual has equal probability of being captured)
  - No selective exclusion of specific sub groups - gender/age/ethnicity/geography/SES
- Independence of data sources (capture in one source does not influence capture in the second source)
- Accuracy of matching procedures and matching outcomes (no erroneous matches or erroneous non-matches)

Data sources for deaths

- Continuous recording systems
  - Vital registration systems
  - Sample registration systems (India/China/Bangladesh/Indonesia)
  - Health system records / parish registers/ ‘population committee’ registers
  - Specific disease/program registers (TB, MCH), police records (injuries)
  - social sector or insurance databases
  - Special registration sites (HDSS/INDEPTH network) – **limited generalizability**
- ↑ **likelihood of ‘dependence’ between multiple continuous systems in same popln**

- Periodic/one-off cross-sectional data collection systems
  - Censuses / intercensal surveys
  - Periodic household surveys – DHS, STEPS, MICS, SES surveys
  - ‘completeness’ surveys – China/India/Bangladesh

Typology of data sources for record linkage studies

Type of data collection	Primary source[1]	Secondary source[2]	Remarks
Continuous recording systems			
Civil registration	Yes		• Optimal source • annual data on routine basis
Alternate registration	Yes	Yes	• Health system vital records e.g Vietnam, Fiji • Church records in Christian societies
Sample registration	Yes	Can serve as a secondary source for evaluating CRVS	• Best alternative to CRVS • Indian SRS (ref) • Chinese DSP (ref) • Bangladesh SVRS (ref)
Special registration	Yes	Can serve as a secondary source for evaluating CRVS or SRS	• E.g. Health and Demographic Surveillance Sites in several countries (INDEPTH Network) (ref)
Age based registers		Yes	• Maternal/child health • senior citizens /pensioners databases
Disease surveillance systems		Yes	• tuberculosis • cancers • injuries • stroke
Periodic data collections			
Census (total population)	Yes	Yes	• Optimal 2nd data source (national coverage)
National sample surveys		Yes	• Inter censal surveys • DHS program • WHO NCD surveillance (STEPS) surveys • UNICEF MICS surveys etc
Special surveys designed to assess completeness		Yes	• Evaluation surveys for sample/special registration • sporadic research based examples

1 = data source for which completeness needs to be evaluated

2 = data source which will be used to evaluate completeness of the primary source

Parameters for study design

- Scope of analysis e.g national / sub national measures; by age; pop sub groups
- Availability/choice of primary & secondary data sources
- Reference time period of analysis
- Matching process
  - Manual/electronic
  - Deterministic/probabilistic/implicit rules
- Statistical procedures
  - Data reconciliation
  - Use of multiple parallel sources or partial data sources
  - DRS method (2source/multiple source models)
  - Hybrid models

Key concepts in DRS analysis

- There should be compatibility of data sources to minimize out of scope events
- Availability of multiple variables for matching
  - Enhances matching potential / validation of matching
- Assurance of data quality
  - Completeness and accuracy of all variables for each death record in each data source
- Matching procedures should be clearly defined
  - Manual / electronic / combination
  - Rules for matched cases – explicit rules vs implicit rules
  - Tolerable limits for specific criteria / deterministic matching / probabilistic matching
  - Mechanisms for field verification of matched/partially matched/ unmatched cases
- Analytical approach – reconciliation/DRS/hybrid approach
- Assessment of DRS conditions (potential for bias)
  - Description of design and data collection process / statistical evaluation
- Measure error of completeness estimate from sampling and bias
- Ethics and data confidentiality

Evaluating bias in completeness estimates

- Completeness of Y = $\frac{a+c}{a+b+c+x}$

- RMSE of completeness estimate: RMSE = $\sqrt{variance + bias^2}$
- Three sources of bias
  - ‘out-of-scope-bias’: results in under estimate of true matches; leading to an ↓ underestimate of completeness; and ↑ overestimate of the vital rate
  - Response correlation bias (from communication/data sharing between sources i.e lack of statistical independence): results in overestimate of true matches; leading to an over estimate of completeness; and underestimate of the vital rate
  - Matching bias: expressed as the *net matching error* which is the difference between the erroneous matches and erroneous non matches.
    - Net matching error is positive = same effect as response correlation bias;
    - if net matching error is negative = effect as ‘out of-scope’ bias
  - ***Due to varying directions; net bias is usually less than any individual source of bias***

Sampling error

- Periodic data collections (except censuses) are based on samples, and usually with cluster design
- Some study designs (e.g. DSP China) involves sampling in both data sources
- Sources of variance
  - Sample size
    - Measuring completeness for specific sub groups (sex, age, geography etc reduces the sample and therefore precision of the estimate
  - Cluster size and characteristics – need to account for design effect

Chandrasekar-Deming estimate of SE of completeness

- In 1949, CD proposed that SE of completeness = $\sqrt{Nq1q2/p1p2}$
- Where N = total number of events estimated by the method (Table 1)

  *p*1 = the probability that an event is recorded in data source 1

  p2 = the probability that an event is recorded in data source 2

  *q*1 = the probability that an event is missed in data source 1

  *q*1 = the probability that an event is missed in data source 2

- Assuming that
  - There is true statistical independence between the two data sources, and zero matching bias or out-of-scope events; and no variance from sampling etc
- Subsequently, various scientists ((Seltzer &Adlakha 1973; Greenfield 1976; Chandrasekar & Deming 1981; Nour 1982, Ayhan 2000, El Khorazaty 2000) proposed methods to estimate bias arising from lack of statistical independence
- Nour (1982) illustrates computation with a practical example with data from Malawi; and El Khorazaty illustrates a practical example with Egyptian data for 1974/75

Variations of record linkage studies

- <u>Variations in design</u>

- Matching all records from two sources of the study population – e.g sample registration system in India; Viet Nam study, Oman, Tonga

- Matching of records in only a sample of the study population – China, Thailand (2006), Indonesia, Malaysia (1995)

- <u>Variations in method for computation of completeness</u>

  - Data reconciliation after matching; no adjustment for cases potentially missed by both sources (Indian SRS; Tonga)

  - Data reconciliation after matching, with adjustment for potentially missed cases – Vietnam, Indonesia (Java)

  - Matching followed by adjustment, no data reconciliation – China, Thailand, Indonesia (other locations), Oman, Malaysia (1995)

Historical review of record linkage completeness studies

Study type	Countries	Remarks
Special registration with periodic surveys	1960-1975 Pakistan, Egypt, Liberia, Malawi, Philippines, Columbia, Morocco, Turkey, Kenya 2006/07 Indonesia	• Time bound projects (-3 years) in listed countries during 1960-1975; USAID PGE program • Tested range of data collection e.g direct household contact; use key informants; combinations • Tested range of recall periods (1,3,6, 12 months) • Completeness; estimated by CD method (ranging from 53 to 90% settings); no 95% CI • Crude birth/death rates adjusted for completeness; no age-specific rates reported; • Indonesian studies in 2006-2007 as sentinel sites, later transformed into national SRS; completeness for 2006 by data reconciliation (no 95% CI); in 2007 by CD method (with 95% CI)
National sample registration with periodic surveys	India – SRS since 1970 Bangladesh-SVRS - 1980 China DSP since 1990 Indonesia since 2014	• India & Bangladesh – continuous recording in sample clusters with total coverage in routine 6 monthly surveys; data reconciliation used to measure mortality, completeness not routinely reported • China – continuous recording in sample clusters with triennial sample completeness surveys; completeness estimated by CD method, results reported with uncertainty intervals for • Indonesia – completeness survey of 2014 discarded due to data quality issues; new survey 2017
Civil registration with periodic data sources	Thailand (2006) Oman (2010) Philippines (2012-14)* Palestine (2017)*	• Thai study involved civil registration and intercensal survey; completeness by CD method, no 95%CI • Oman study involved civil registration and national census; completeness by CD method with 95% CI • Philippines and Palestine – civil registration and census (studies yet to be implemented)
Multiple sources with overlapping recall periods	Philippines 2006/7 Viet Nam 2008/9 Kiribati (2001-2009) Tonga (2000-2009)	• Philippines study – Civil registration; health system; parish records; CD method; with 95%CI by Max Lik Est • Viet Nam study – civil registration; health system; peoples committee plus additional partial sources; completeness by variant of CD method with 95% Ci (by bootstrapping method) • Kiribati – civil registration; health information system; reproductive surveillance, data reconciliation; no CI • Tonga –civil registration; health information system; completeness by CD method; No 95% CI
Civil registration with HDSS	South Africa 2006-09	• Civil registration and HDSS; electronic linkage with deterministic & probabilistic matching; completeness not measured due to 'out-of-scope' coverage

Example: Viet Nam 2009

- Study population – 192 communes; 2.6 million pop
- Data sources – Commune health station/Population department- (source 1); Justice system (source 2); others – Farmer's union, Womens group, aged care
- manual matching at commune level, leading to reconciled list of unique events
- relaxation of matching criteria (age, date of death) owing to inaccurate recording in either source (exercise of local judgement critical to the matching process)
- Unobserved cell computed from two source analysis
- Reconciliation before ascertaining causes of death, hence reconciled data used as numerator for deriving completeness
- Completeness factor used to adjust life tables and later develop cause-specific mortality estimates for burden of disease analysis

Matching results

	Regions	Total in reconciled list	CHC ⟺	Population Dep	Justice system	Other
1	Ha Noi	2304	1723 (75%)	1580 (69%)	1669 (72%)	720 (31%)
2	Thai Nguyen	1185	999 (85%)	210 (18%)	183 (15%)	85 (7%)
3	Hue	2221	1768 (78%)	1043 (47%)	1311 (59%)	777 (35%)
4	Ho Chi Minh	2453	435 (18%)	571 (23%)	1871 (76%)	202 (8%)
5	Can Tho	1758	872 (49%)	758 (43%)	1081 (62%)	535 (30%)

- *A death could be recorded in more than one system*
- ⟺ *= interdependence*

Viet Nam 2009

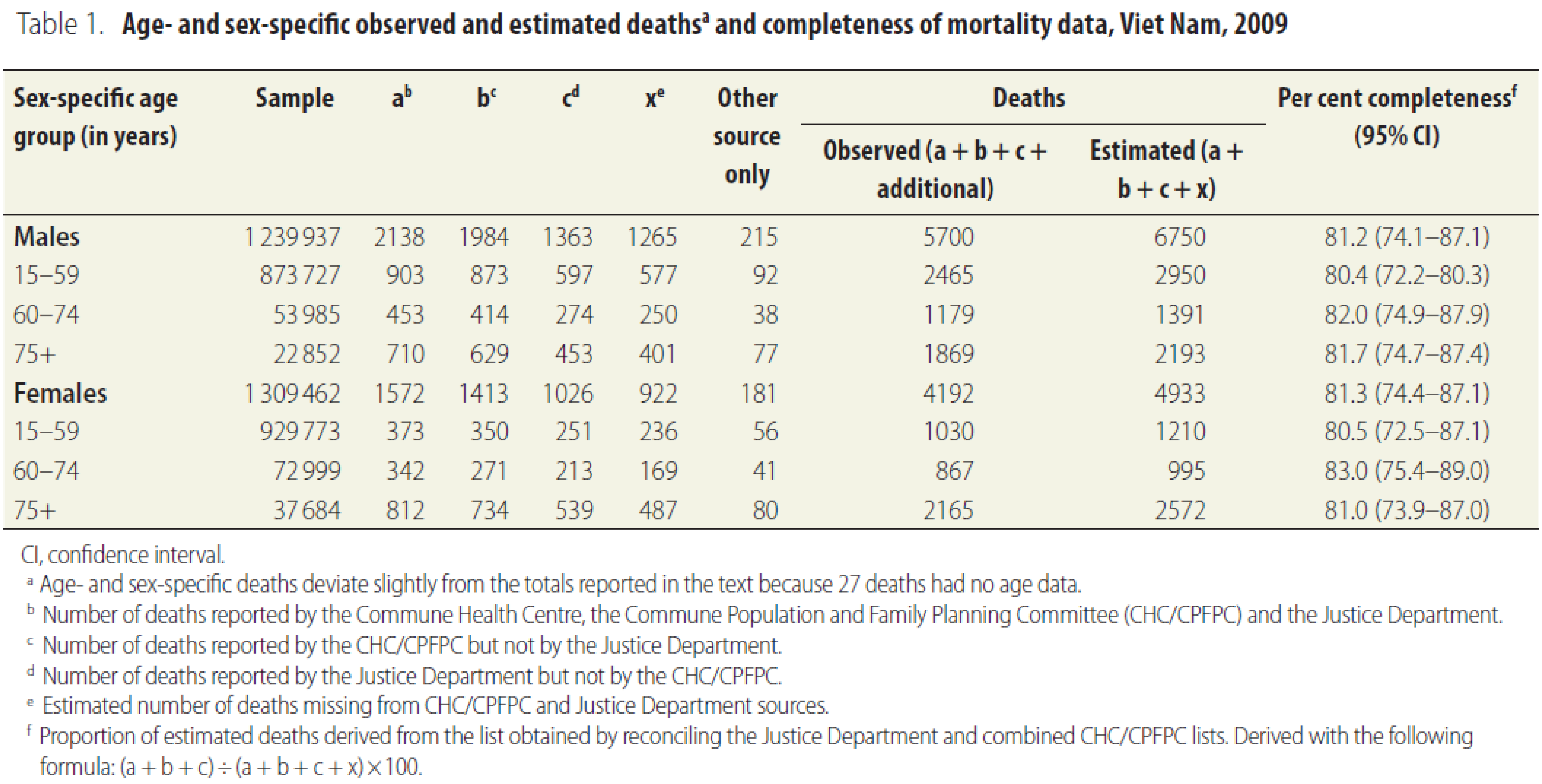


Table 1. **Age- and sex-specific observed and estimated deaths[a] and completeness of mortality data, Viet Nam, 2009**

Sex-specific age group (in years)	Sample	a[b]	b[c]	c[d]	x[e]	Other source only	Deaths		Per cent completeness[f] (95% CI)
							Observed (a + b + c + additional)	Estimated (a + b + c + x)	
Males	1 239 937	2138	1984	1363	1265	215	5700	6750	81.2 (74.1–87.1)
15–59	873 727	903	873	597	577	92	2465	2950	80.4 (72.2–80.3)
60–74	53 985	453	414	274	250	38	1179	1391	82.0 (74.9–87.9)
75+	22 852	710	629	453	401	77	1869	2193	81.7 (74.7–87.4)
Females	1 309 462	1572	1413	1026	922	181	4192	4933	81.3 (74.4–87.1)
15–59	929 773	373	350	251	236	56	1030	1210	80.5 (72.5–87.1)
60–74	72 999	342	271	213	169	41	867	995	83.0 (75.4–89.0)
75+	37 684	812	734	539	487	80	2165	2572	81.0 (73.9–87.0)

CI, confidence interval.

[a] Age- and sex-specific deaths deviate slightly from the totals reported in the text because 27 deaths had no age data.

[b] Number of deaths reported by the Commune Health Centre, the Commune Population and Family Planning Committee (CHC/CPFPC) and the Justice Department.

[c] Number of deaths reported by the CHC/CPFPC but not by the Justice Department.

[d] Number of deaths reported by the Justice Department but not by the CHC/CPFPC.

[e] Estimated number of deaths missing from CHC/CPFPC and Justice Department sources.

[f] Proportion of estimated deaths derived from the list obtained by reconciling the Justice Department and combined CHC/CPFPC lists. Derived with the following formula: $(a + b + c) \div (a + b + c + x) \times 100$.

- Hoa, N.P., Rao C et al., *Mortality measures from sample-based surveillance: evidence of the epidemiological transition in Viet Nam.* Bulletin of the World Health Organization, 2012. **90**(10): p. 764-772.

Adjusted mortality indicators

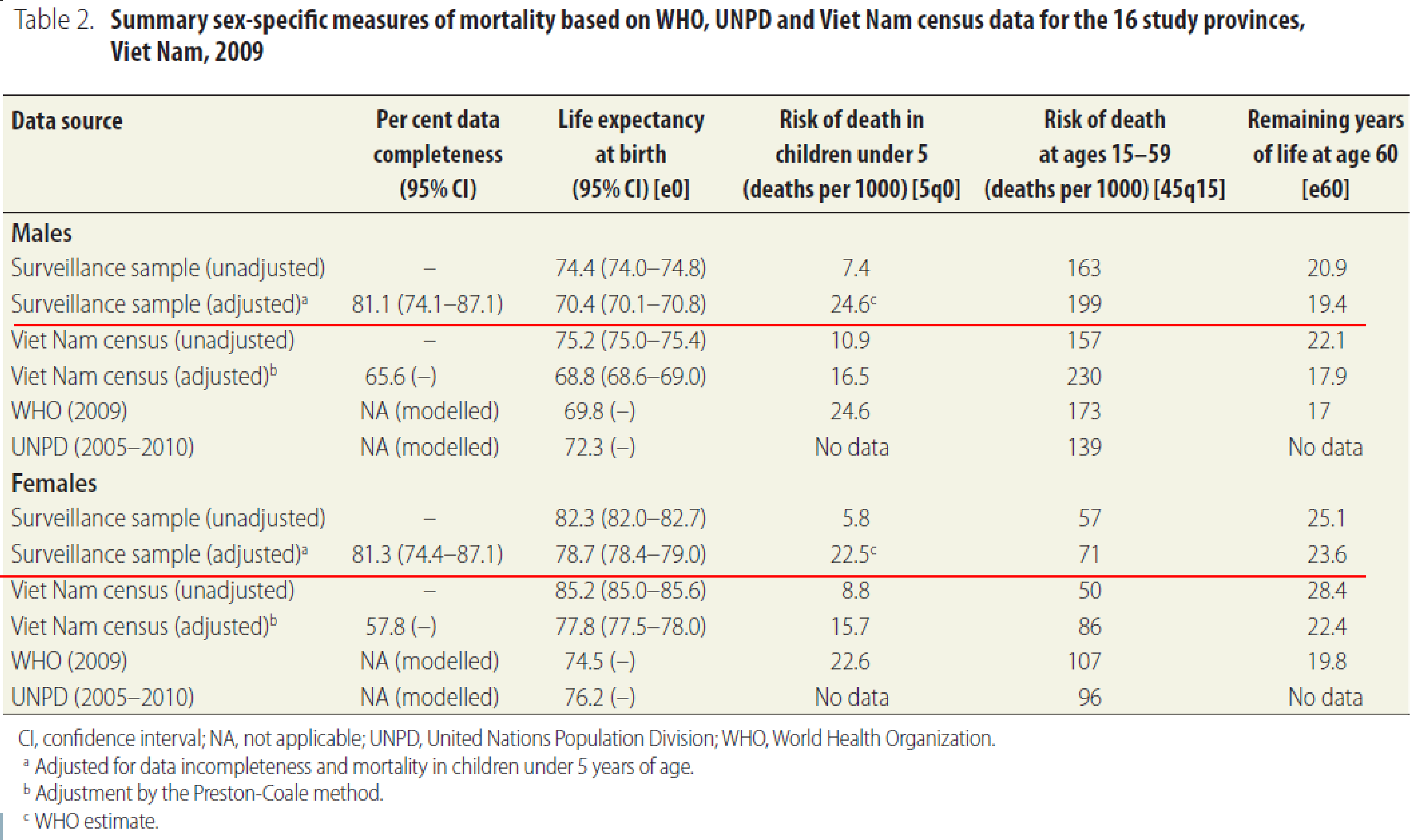


Table 2. **Summary sex-specific measures of mortality based on WHO, UNPD and Viet Nam census data for the 16 study provinces, Viet Nam, 2009**

Data source	Per cent data completeness (95% CI)	Life expectancy at birth (95% CI) [e0]	Risk of death in children under 5 (deaths per 1000) [5q0]	Risk of death at ages 15–59 (deaths per 1000) [45q15]	Remaining years of life at age 60 [e60]
Males					
Surveillance sample (unadjusted)	–	74.4 (74.0–74.8)	7.4	163	20.9
Surveillance sample (adjusted)[a]	81.1 (74.1–87.1)	70.4 (70.1–70.8)	24.6[c]	199	19.4
Viet Nam census (unadjusted)	–	75.2 (75.0–75.4)	10.9	157	22.1
Viet Nam census (adjusted)[b]	65.6 (–)	68.8 (68.6–69.0)	16.5	230	17.9
WHO (2009)	NA (modelled)	69.8 (–)	24.6	173	17
UNPD (2005–2010)	NA (modelled)	72.3 (–)	No data	139	No data
Females					
Surveillance sample (unadjusted)	–	82.3 (82.0–82.7)	5.8	57	25.1
Surveillance sample (adjusted)[a]	81.3 (74.4–87.1)	78.7 (78.4–79.0)	22.5[c]	71	23.6
Viet Nam census (unadjusted)	–	85.2 (85.0–85.6)	8.8	50	28.4
Viet Nam census (adjusted)[b]	57.8 (–)	77.8 (77.5–78.0)	15.7	86	22.4
WHO (2009)	NA (modelled)	74.5 (–)	22.6	107	19.8
UNPD (2005–2010)	NA (modelled)	76.2 (–)	No data	96	No data

CI, confidence interval; NA, not applicable; UNPD, United Nations Population Division; WHO, World Health Organization.

[a] Adjusted for data incompleteness and mortality in children under 5 years of age.

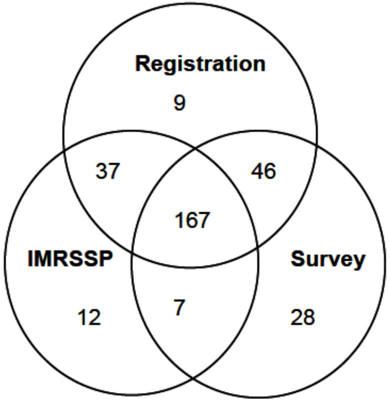
[b] Adjustment by the Preston-Coale method.

[c] WHO estimate.

Example 3: Indonesia (a)

- Central Java – record linkage/matching across three sources (health system, vital registration, independent survey)
- Independent survey and record linkage/matching conducted only in a sample of villages from the overall study population
- Completeness of health system data calculated as a proportion of total deaths obtained from the reconciled list of unique deaths

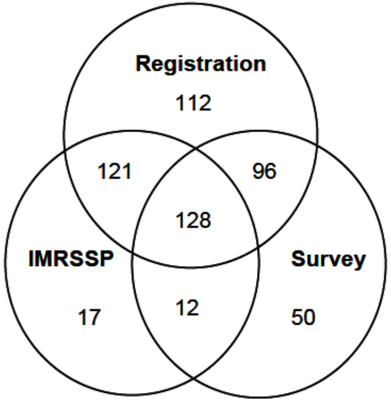
PEKALONGAN



Total deaths = 306

Completeness = 73%

SURAKARTA

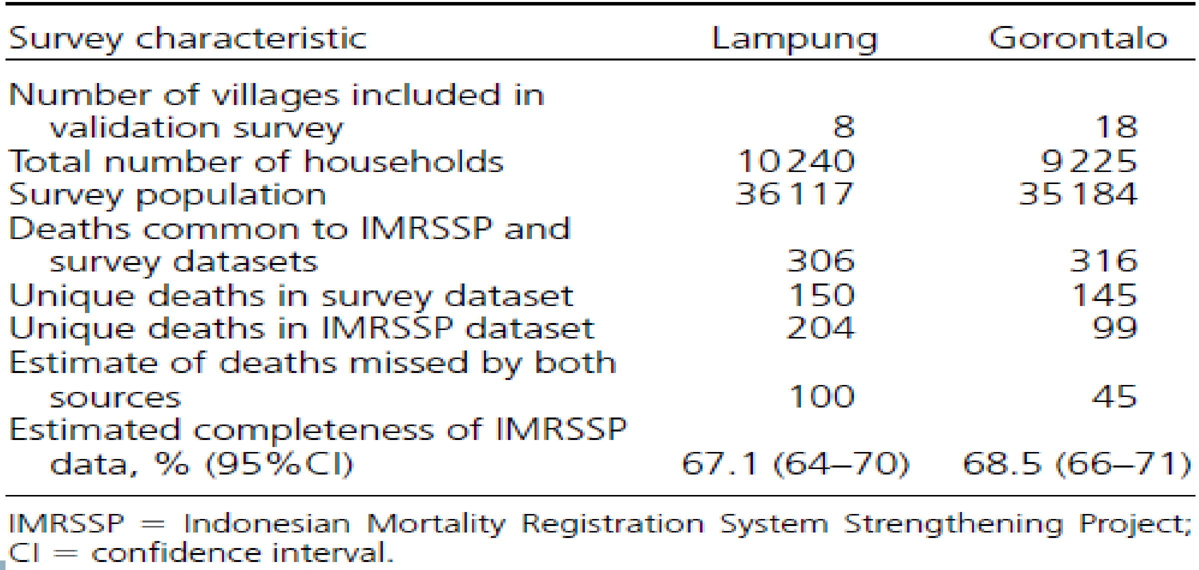


Total deaths = 536

Completeness = 55%

Example 3: Indonesia (b)

- Lampung/Gorontalo (2007-2008) – two data sources - health system records of facility and community deaths, and an independent survey
- Independent survey in a sample of villages from the overall study population, recall of deaths over two years, record linkage/matching across the two sources
- Analysis using capture recapture methods completeness computed as a proportion the total deaths including the estimated unobserved deaths



Survey characteristic	Lampung	Gorontalo
Number of villages included in validation survey	8	18
Total number of households	10 240	9 225
Survey population	36 117	35 184
Deaths common to IMRSSP and survey datasets	306	316
Unique deaths in survey dataset	150	145
Unique deaths in IMRSSP dataset	204	99
Estimate of deaths missed by both sources	100	45
Estimated completeness of IMRSSP data, % (95%CI)	67.1 (64–70)	68.5 (66–71)

IMRSSP = Indonesian Mortality Registration System Strengthening Project; CI = confidence interval.

Rao C, Kosen S, et al. Tuberculosis mortality differentials in Indonesia during 2007-2008: evidence for health policy and monitoring. Int J Tuberc Lung Dis. 2011;15(12):1608-14.

Limitations of DRS methods

- In PGE studies, several conditions for record linkage difficult to fulfil (e.g. absence of out-of-scope events, homogenous capture probability; statistical independence of data sources,; accuracy of matching)
- These occur as a result of the
  - nature of the events (e.g deaths in low SES strata less likely to be registered);
  - nature of data collection processes (passive or active)
  - Quality of data collected in each source
- All the above lead to potential bias in the completeness estimate
- Further, there is also sampling error / stochastic variation; which contribute to uncertainty in the completeness estimate
- In addition, there were considerable logistical challenges in implementing record linkages studies in terms of costs and manpower, as well as technical challenges in matching, evaluation of bias etc

Strengths of DRS methods

- Essentially the major conditions / assumptions of record linkage and DRS methods are statistical as compared to the demographic assumptions for indirect techniques (related to underlying fertility/mortality/population growth patterns in the study population)
- The data collection procedures allow assessment of bias and error, hence enabling a more informed assessment of uncertainty of the completeness estimate
- Findings enable completeness assessment and also help identify systemic weaknesses in registration system, including specific population sub groups
- Involvement of local staff in matching helps build awareness and capacity for strengthening registration
- Age specific measures of completeness
- Data reconciliation especially from additional fragmentary sources helps fill data gaps in cause of death information

Reasons for renewed interest

- Availability of computerised registration datasets as well as computerisation of periodic data collections (censuses, surveys); which will increase going forward
- Potential to improve data quality of recorded variables used in linkage (name spellings; address variables, age, date of death etc)
- Wider use and recording of Unique Identifiers which are invaluable for linkage
- Electronic linkage vastly reduces logistical challenges of manual matching
- Explicit rules and probabilistic approach using computerised datasets can be applied to test a range of scenarios and judge cut points for specific criteria
- Routine application of DRS method in India and China serve as robust examples of their general acceptability

Principles of DRS analysis

- Develop an efficient study design based on a careful choice of alternatives
  - E.g existing routine data sources vs special data collection
  - Scope of desired outcome measures (e.g by age, geography etc)
- Establish a clear understanding of data collection procedures to evaluate potential for and degree of bias
- Conduct a thorough analysis and evaluation of completeness estimates alongwith margins of error

Conclusions and recommendations

- Hierarchy of study designs (based on sample size; potential for meeting condition of independence; cost considerations; potential for sub group analysis)
  - CRVS with census based recall of deaths
  - CRVS with intercensal survey / nationally representative sample survey/special survey
  - SRS with periodic special surveys
  - Special registration in targeted surveillance sites with special surveys

Recommendations contd

- Focus on computerisation of all data sources
- Inclusion of relevant variables in all future potential data sources
- Emphasis on data quality (name spellings; address variables; accuracy of age, date of death; and where available Unique ID numbers)
- Promote follow up of electronic linkage with field verification of sample of matched/partially matched pairs and unmatched cases (to assess net matching error)
- Use all available evidence and methods to assess for bias and error in completeness; and where possible, conduct sensitivity analysis applying different methods

- Completeness estimates should be presented with margins of error, to assess impact on mortality indicators