SafetyCube

Guidelines for the determination of the number of serious road injuries

The Hague, 24 May 2016

Jean Louis Martin

Co-funded by the Horizon 2020
Framework Programme of the European Union

Using linked/matched police and hospital data

Interest in using data linkage

- Maximises use of available data sources
- Provides insight into the completeness of police and hospital data
- Reduces / identifies selection biases

General context for using linkage

		Hospital data	
		yes	no
Police data	Yes	Common data	Unlinked police data
	No	Unlinked hospital data	Unobserved / missed

Implementation of the linking process

- Deterministic: the best if unique identifier
- Otherwise probabilistic or distance based linkage, using variables common to the data sources:
  - *date of birth of the casualty*
  - *gender*
  - *date and time of the crash (and/or date and time of hospital admission)*
  - *location of the crash*
  - *severity of the crash*
  - *mode of transport.*

Objective: estimation of the number of serious road injuries defined as MAIS3+

1. MAIS3+ are almost all hospitalized
   (about 95% according to the Rhône Trauma Registry in France which includes in and out-patients)
   ➔We consider in the following that considering only hospitalized casualties leads to a very small underestimation of MAIS3+
2. The number of casualties hospitalized and not recorded in Hospital Discharge Register (HDR) is considered negligible.
3. MAIS is then derived from ICD. This allows selection of MAIS3+ (and exclusion of MAIS2-)
4. Hence injuries can be characterized as road casualties except if external cause is missing or incorrectly documented. This uncertainty will be removed by linking with police data

Using of Hospital Discharge Register (HDR)

	Hospitalized and in HDR			Hospitalized and not in HDR	Not hospitalized
	MAIS3+		MAIS2-		
	External cause: Traffic casualty	Other or missing external cause			
In police					
Not in police					

- A main consequence of removing the MAIS2- column is that the four remaining cells are restricted to MAIS3+ without knowing MAIS level in police data.

Using of Hospital Discharge Register (HDR)

	Hospitalized and in HDR MAIS3+	
	External cause: Traffic casualty	Other or missing external cause
In police	n_{PR}	$n_{P\overline{R}}$
Not in police	$n_{\overline{P}R}$	$n_{\overline{PR}}$

Hence injuries can be characterized as road casualties except if external cause is missing or incorrectly documented. This nto account by the

Hospitali common

Hospitalize common to recorded ➔ Conside road casualties

Hospitalized MAIS3+ recorded by police but not identified as road

Remaining question: How to estimate the part of MAIS3+ road casualties in this cell?

Estimating the unobserved subset by capture-recapture approach

		List B		
		B	$\bar{B}$	
List A	A	n_{AB}	$n_{A\bar{B}}$	n_{A}
	$\bar{A}$	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$	
		n_{B}		n

If one assumes that the probability of being registered in list A is independent of the probability of being registered in list B, this translates into:

$$n_{AB}/n_{B}=n_{A}/n$$

➔ from this we obtain the intuitive Petersen estimate:

$$\hat{n} = \frac{n_A \times n_B}{n_{AB}}$$

Numerical example from the Dutch data

		List B		
		External cause: Traffic casualty	Other or missing external cause	
List A	In police data	1752	90	1842
	Not in police data	5417		
		7169		

Numerical example from the Dutch data

		List B		
		External cause: Traffic casualty	Other or missing external cause	
List A	In police data	1752	90	1842
	Not in police data	5417	278	
		7169		7537

$$\hat{n} = \frac{1842 \times 7169}{1752} = 7537$$

- Using linkage method on this example adds 5% of MAIS3+ to the hospitalized MAIS3+ observed in HDR.
- This small increase is due to the low proportion of missing external causes for Dutch data.
- If a country has 20% of missing values for external causes, it leads to a 20% increase in the total estimated number.

Conditions for using capture-recapture method

(1) No entry or loss between the registrations (close population)

(2) perfect identification of subjects common to both registrations

(3) independence of recording between the registrations

(4) homogeneity of capture by a given registration

(5) same geographical area and same time period

(6) perfect identification of the subjects of interest

Perfect identification of the subjects of interest (C6)

- The criteria for defining a subject of interest must be very precise, and should be the same for the two (or more) registrations.
- When the definition of the subject of interest in one source is included in the definition of the other source, the most restricted definition hence applies. For the most common case described above , the HDR definition is "MAIS3+ and hospitalized" which is included in the police definition (Injured whatever the severity). The outcome is then restricted to MAIS3+ hospitalized

Independence of recording between the registrations(C3)

- The subjects' probability of being registered in one source should be independent of the probability of being registered by the other source. This is the basic underlying assumption for establishing the Petersen estimator
- Coming back to the practical case shown above, the condition of independence means that the probability of having a correct or missing external cause is independent of being registered or not by the police
- If there is a positive dependence, the obtained estimate of the number of road traffic hospitalized MAIS3+ is likely a lower bound of the estimated number

Homogeneity of capture by a given source/registration (C4)

- All subjects of interest should have the same probability of being registered by a given source, but this is usually not the case. Many characteristics usually influence the reporting probability: the number of vehicles involved in the crash, the road user type
- In such cases, the homogeneity of capture is only valid **within sub-groups** (ex: within cyclists, within car occupants, etc..
- Two ways to account for lack of homogeneity:
  - *To stratify on these sub-groups, i.e. to stratify on the variable which is associated with the probability of registration, and which defines the sub-groups. More precisely, one should estimate the number of subjects of interest in each stratum, and then one should sum up the estimates obtained over the strata to get the total number of subjects of interest (workable with 2 or 3 variables)*
  - *To use an explicit modelling and to include as covariates the variables that influence registration probability. The number of covariates one can take into account is hence higher.*

Key points and some recommendations when using linked police and hospital data

- The key idea is to use all available information (Police+Hospital+Other sources)
- Linking process can only be based on variables that are included in both records. The most ideal variable is a unique personal identification number (deterministic linkage), but this information is most likely not available for privacy reasons
- In the absence of unique identifier, probabilistic or distance based linkage is recommended. Linking variables commonly used are date and time of the crash (and/or date and time of hospital admission), location of the crash, gender and date of birth of the casualty, mode of transport
- MAIS3+ casualties are mostly hospitalized and recorded in hospital data, but external causes derived from ICD are often missing or misspecified.

Key points and some recommendations when using linked police and hospital data

- The number of traffic casualties recorded in hospital data but not identified as such can be estimated by linking these data with police data and using capture-recapture method
- The capture-recapture approach is based on six conditions, especially the three following ones:
  - *perfect identification of the subjects of interest :*
    the definition of the road casualty in the two data sources should be the same or included into one another
  - *Independence between the registrations:*
    when this hypothesis is weak, estimation is biased downwards in case of positive dependence, upwards otherwise
  - *Homogeneity of capture by a given registration:*
    stratification or modelling methods can be used when homogeneity assumption is only valid within subgroups

Topics/questions to discuss

- According to available data in one country, is it worth using linking method?