27-28 September 2022

7th IRTAD Conference

Linked police and health data: how to apply capture-recapture to correct for under-reporting and bias

E. Amoros, C. Aksoy, A. Ndiaye, B. Laumon, B. Gadegbeku, J-L. Martin

Université Gustave Eiffel

Université Claude Bernard Lyon 1

UNIVERSITÉ DE LYON

Linked police and health data

1) police data : in most countries

(killed= well-recorded in industrialized countries)
Injured= large under-reporting and bias

2) Health data :
National hospital discharge data = inpatients
Emergency departments (ED) ≈ outpatients
The Rhone road trauma registry ≈ outpatients and inpatients

Linked police and health data : under-reporting, Rhone example

annual mean	Police data	Health data	linked	total	Tot./pol	Tot/health
2006-2016	2800	7600	1700	8600	3,1	1,1

France, Rhone county (1,8 M)

Heath data: Rhone road trauma registry = outpatients + inpatients

		In health data (B) ?		
		yes	no	total
In police data (A) ?	yes	**1700**	**1100**	2800
	no	**5900**	**?**	
	total	7600		n=8600+ ?

Correcting for under-reporting: capture-recapture on linked police and health data

		In health data (B) ?		
		yes	no	total
In police data (A) ?	yes	n_{AB}	$n_{A\bar{B}}$	n_A
	no	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$ = ?	
	total	n_B		n= ?

Simple 2-list method (IF capture-recapture conditions are met):

Petersen estimate :

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

Capture-recapture conditions

2 implicit conditions:

- same geographical area and same time period
- perfect identification of subjects of interest (injured, in a road crash)
  (implies same definition in both sources)

4 key conditions:

- close population
- perfect record-linkage
- independence between sources (registrations)
- homogeneity of capture : for a given source/ registration (ex: police),
  the different casualties have the same probability of being recorded,
  whatever their characteristics

discussion of some capture-recapture conditions

1) positive dependence between hospital and police data
=> capture-recapture estimate will be a lower bound

2) capture by police-reporting is not homogenous ;
it usually varies with

- injury severity
- mode of transport (pedestrians, bicycle, M2W, car, etc)
- single-vehicle / multi-vehicle crash or crash opponent (yes/no)
- type of road network
- driver / passenger
- type of police

3) health-reporting slightly varies with

- injury severity

under-reporting and **biais** : example

annual mean	**police**	**health**	**linked**	**total**	**Tot./pol.**	Tot/health
M2W, with opponent	557	865	371	1052	1,9	1,2
M2W, without opp.	96	1063	65	1094	11,4	1,0
bicycle, with opp.	157	332	100	389	2,5	1,2
bicycle, without opp.	6	916	4	919	146,4	1,0
car, with opp.	1092	2415	647	2859	2,6	1,2
car, without opp	269	1050	178	1140	4,2	1,1
total	**2789**	**7563**	**1733**	**8619**	**3,1**	1,1

France, Rhone county, 2006-2016 (health data= outpatients+ inpatients)

capture-recapture on each strata : mode * crash opponent

		In health data (B) ?		
		yes	no	total
In police data (A) ?	yes	n_{AB}	$n_{A\bar{B}}$	n_A
	no	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$ = ?	
	total	n_B		n= ?

Petersen estimate :

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

=> Capture-recapture with stratification on mode* crash opponent:

annual mean	Police data	Health data	linked	total	Tot/ police	Tot/ health	**CRC**	CRC/ pol.	CRC/ health.
M2W, with opp.	557	865	371	1052	1,9	1,2	**1300**	2,3	1,5
M2W, without opp.	96	1063	65	1094	11,4	1,0	**1573**	16,4	1,5
bicycle, with opp.	157	332	100	389	2,5	1,2	**523**	3,3	1,6
bicycle, without opp.	6	916	4	919	146,4	1,0	**1437**	229,1	1,6
car, with opp.	1092	2415	647	2859	2,6	1,2	**4073**	3,7	1,7
car, without opp	269	1050	178	1140	4,2	1,1	**1583**	5,9	1,5
total	2789	7563	1733	8619	3,1	1,1	**12016**	4,3	1,6

But capture-recapture (CRC) with stratification on mode* crash opponent: pb with #MAIS3+

annual mean	Police data	Police MAIS3+	**Pol prop3+**	Health data	Heath MAIS3+	**Health prop3+**	total	**Tot. Prop3+**	CRC	CRC mais3+	**CRC prop3+**
M2W, with opp.	557	103	**18%**	865	95	**11,0%**	1052	11,7%	**1300**	237	**18%**
M2W, without opp.	96	29	**30%**	1063	61	**5,7%**	1094	6,3%	**1573**	470	**30%**
bicycle, with opp.	157	20	**13%**	332	21	**6,2%**	389	6,9%	**523**	68	**13%**
bicycle, without	6	2	**37%**	916	37	**4,0%**	919	4,1%	**1437**	533	**37%**
car, with opp.	1092	57	**5%**	2415	49	**2,0%**	2859	2,3%	**4073**	212	**5%**
car, without opp	269	31	**12%**	1050	38	**3,6%**	1140	4,1%	**1583**	183	**12%**
total	**2789**	**338**	**12%**	**7563**	**386**	**5,1%**	**8619**	**5,6%**	**12016**	**1952**	**16%**

Estimated number of MAIS3+ is too high and biased

discussion of some capture-recapture conditions

1) positive dependence between hospital and police data
=> capture-recapture estimate will be a lower bound

2) capture by police-reporting is not homogenous ;
it usually varies with

- **injury severity**
- mode of transport (pedestrians, bicycle, M2W, car, etc)
- single-vehicle / multi-vehicle crash or crash opponent (yes/no)
- type of road network
- driver / passenger
- type of police

3) health-reporting slightly varies with

- injury severity

Linked police and health data : predict MAIS3+

France, Rhone county (1,8 M)

Heath data: Rhone road trauma registry = outpatients + inpatients

		In health data (B) ?		
		yes	no	total
In police data (A) ?	yes	**1700**	**1100**	2800
	no	**5900**		
	total	7600		

Construct P(MAIS 3+ /1-2) on the linked dataset (MAIS from Health data)

P(MAIS 3+/1-2) as a function of crash and injured road user characteristics (from police data)

Apply the model to the subset "police data only"
=> predicted or observed MAIS3+ for all casualties observed in the Rhone county

Capture-recapture on mode * crash opponent * **MAIS** (1-2/3+):

annual mean	Police data	Police MAIS3+	**Pol prop3+**	Health data	Heath MAIS3+	**Health prop3+**	total	**Tot. Prop3+**	**CRC**	CRC MAI3+	**CRC prop3+**
M2W, with opp.	557	103	**18%**	865	95	**11,0%**	1052	11,7%	**1314**	130	**9,9%**
M2W, without opp.	96	29	**30%**	1063	61	**5,7%**	1094	6,3%	**1634**	82	**5,0%**
bicycle, with opp.	157	20	**13%**	332	21	**6,2%**	389	6,9%	**527**	30	**5,6%**
bicycle, without	6	2	**37%**	916	37	**4,0%**	919	4,1%	**1464**	56	**3,8%**
car, with opp.	1092	57	**5%**	2415	49	**2,0%**	2859	2,3%	**4096**	72	**1,8%**
car, without opp	269	31	**12%**	1050	38	**3,6%**	1140	4,1%	**1600**	52	**3,3%**
total	**2789**	**338**	**12%**	**7563**	**386**	**5,1%**	**8619**	**5,6%**	**12175**	**549**	**4,5%**

Capture-recapture :

However:

Some strata may contain **small frequencies**:
ex: injured cyclists without crash opponent in police data

More than 3 variables are associated with under-reporting:

- injury severity,
- mode of transport
- single-multi vehicle crash,
- type of road network,
- driver/passenger
- type of police

=> multivariate modelling

Multivariate multinomial model

Multinomial response variable Y :

- 1= Casualties recorded in police data only
- 2= Casualties recorded in health data only
- 3= Casualties recorded in both

(disjoint subgroups)

Explanatory variables: those associated with under-reporting

- injury severity,
- mode of transport
- single-multi vehicle crash,
- type of road network,
- driver/passenger
- type of police

model with interaction between the 3 variables = stratification on 3 variables
model with interaction between 2 var + var3 as main effect = no equivalent

Multivariate multinomial model

with SAS software:

```
PROC LOGISTIC data = collBUR out=modelCRC;
class source (ref = "1") mode_oppon (ref="carWithO") MAIScode (ref= "MAIS3p") / param = ref;
model source = mode_oppon MAIScode var3 var4 var5/ link = glogit;
weight decimal_freq; * freq=integer_freq;
format _all_;
ods output ParameterEstimates=est;
run;
```

Source : where the casualty is registered
1= in police data only / 2= in health data only, 3= in police AND health data

Multivariate multinomial model

with R software:

```
modelCRC <- multinom(source ~  mode_oppon + MAIScode + var3 + var4 +var5,
data=collBUR, weights=freq)

summary(modelCRC)
betas_modelCRC <- coef(modelCRC)
```

Multinomial model on French data :

- Type of police (3 categories) * type of road
- Daytime/nighttime

- Mode of transport * crash opponent
- MAIS (1-2 / 3+)
- Hospitalized (yes/no)
- (Age)
- (Gender)
- Driver /passenger

Thank you for your attention

UMRESTTE

UNITÉ MIXTE DE RECHERCHE
ÉPIDÉMIOLOGIQUE ET DE SURVEILLANCE
TRANSPORT TRAVAIL ENVIRONNEMENT

Sous la co-tutelle de :

UCBL • UNIVERSITÉ CLAUDE BERNARD LYON 1
UNIVERSITÉ GUSTAVE EIFFEL

Université Gustave Eiffel

ARVAC

emmanuelle.amoros@univ-eiffel.fr

Additional slides

Correcting for under-reporting and bias with capture-recapture on linked police and health data

ex: French Rhône county, 1996-2004, average annual frequencies

		In hospital data (B) ?		
		yes	no	
In police data (A) ?	yes	**1700**	**1100**	2800
	no	**5900**	**?**	
		7600		n=8600+ ?