CENTRE DE RECHERCHE EN DÉMOGRAPHIE ET SOCIÉTÉS, UCL

ANALYSIS-ORIENTED LONGITUDINAL DATA MANAGEMENT:
CHANGING THE ANALYTICAL PERSPECTIVE IN DHS AND HDSS

20 March 2012

Philippe Bocquier, Université Catholique de Louvain

Outline

Conviction: longitudinal data process model influences the quality of the analysis

- Breaking the myths on EHA
- Process for descriptive analysis
  - Conventional approach
  - EHA approach
- Process for regression analysis
  - Conventional approach
  - EHA approach

Some myths around Event History Analysis (EHA)

20 March 2012

- EHA is essentially about:
  - Lifetime data
  - Survey data
  - Right-censored data
  - Multivariate analysis
- EHA is inappropriate or a waste of time for:
  - Short-time data
  - Register or census data
  - Left-censored data
  - Basic, descriptive or aggregate data

(Terminology used in this presentation:
Event history analysis = biographical data analysis)

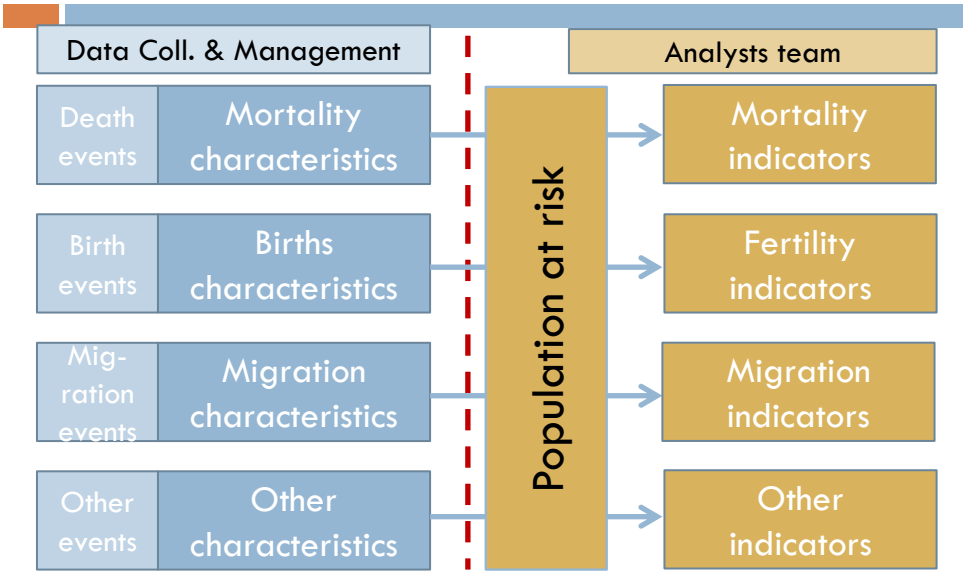
Event History Analysis (EHA)

20 March 2012

EHA should not be:	EHA should be:
… starting from after data cleaning	… starting from data collection and data management
… separated from cross-sectional analysis	… forming a continuum with cross-sectional analysis
… used only at senior (PhD) level	… part of basic training (Master's level)
… left to analysts only	… integrated into data processing

Conventional model of descriptive data process: "Module approach"

20 March 2012



Conventional model of descriptive data process: "Module approach" in words

20 March 2012

- Numerator is aggregated on fixed time intervals
- Denominator is aggregated on fixed time intervals:
  - Either mid-year/period population
  - Or mean population (multiplied by number of years in the interval to approximate person-years at risk)
- Rates are obtained by dividing numerator by denominator
- Assumption of equal distribution over time interval for both numerator and denominator
  - Necessary when cohort data is used to compute calendar time (period) indicators, or *vice versa*

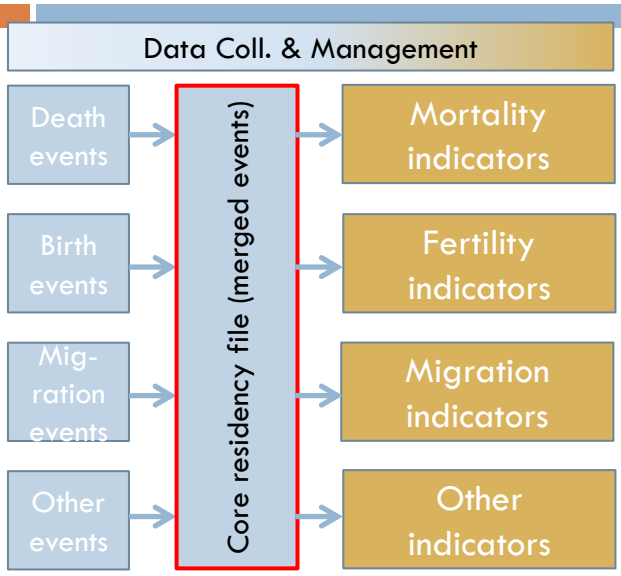
Conventional model of descriptive data process: "Module approach" pitfalls

20 March 2012

- Computing the number of events (numerator) usually not a problem
- Risk of computing denominator wrongly:
  - Assumption of equal distribution is wrong in many instances
  - Person-years at risk depends on inclusion and exclusion criteria defined using other files
- Probability-to-rate conversion (or vice versa) heavily depends on equal distribution assumption

EHA model of descriptive data process: "Core residency file approach"

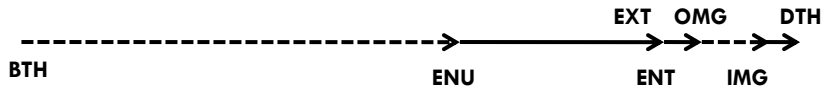
20 March 2012



What is a core residency file? ("Core individual level residency file")

20 March 2012

	Individual_ID	Household_ID	Event	DoB	DateEvent
A	G0010010010001	G00100100100	ENU	17 Nov 1947	21 Aug 2002
	G0010010010001	G00100100100	EOB	17 Nov 1947	31 Dec 2010
B	G0010010010002	G00100100100	ENU	1 Jul 1976	21 Aug 2002
	G0010010010002	G00100100100	OMG	1 Jul 1976	1 Jul 2007
C	G0010010010003	G00100100100	ENU	23 Aug 1985	21 Aug 2002
	G0010010010003	G00100100100	EXT	23 Aug 1985	1 Jul 2007
	G0010010010003	G00203000104	ENT	23 Aug 1985	2 Jul 2007
	G0010010010003	G00203000104	OMG	23 Aug 1985	10 Nov 2007
	G0010010010003	G00100100111	IMG	23 Aug 1985	30 Mar 2008
	G0010010010003	G00100100111	DTH	23 Aug 1985	15 Oct 2008
D	G0010010010004	G00100100100	ENU	1 Jul 1988	21 Aug 2002
	G0010010010004	G00100100100	EXT	1 Jul 1988	1 Jul 2007
	G0010010010004	G00203000104	ENT	1 Jul 1988	2 Jul 2007
	G0010010010004	G00203000104	OMG	1 Jul 1988	10 May 2008
E	G0010010010005	G00100100100	BTH	1 Jul 2005	1 Jul 2005
	G0010010010005	G00100100100	EOB	1 Jul 2005	31 Dec 2010
F	G0010010010006	G00100100100	IMG	1 Jul 1983	31 Aug 2007
	G0010010010006	G00100100100	OMG	1 Jul 1983	8 Apr 2008



EHA model of descriptive data process: Event and date consistency checks

20 March 2012

- Check non-logical sequence of events such as:
  - ⇒ BTH after DTH/ENU/IMG...
  - ⇒ Succession of same event
  - ⇒ First event not ENU, IMG or BTH
  - ⇒ Last event not OMG, DTH or end of observation (EOB)
  - ⇒ Internal moves: EXT not followed by ENT
- Check dates:
  - ⇒ Date of birth, date of last observation...
  - ⇒ Non-logical sequence often originate in errors of dates
  - ⇒ Need at least one day between BTH and DTH (born alive)

Kintampo HDSS: before corrections

20 March 2012

event	ENU	BTH	IMG	OMG	EXT	ENT	DTH	CUR	.	Total
ENU	320	0	346	48,569	22,941	34	4,503	49,530	14	126,257
BTH	0	6	28	3,530	3,036	0	1,091	16,335	0	24,026
IMG	5	1	367	26,058	11,452	74	1,025	45,959	11	84,952
OMG	0	0	10,856	306	210	435	6	329	74,853	86,995
EXT	0	0	875	65	538	44,837	2	280	72	46,669
ENT	0	0	362	8,455	8,455	475	563	27,404	161	45,875
DTH	0	0	0	0	0	0	19	1	7,192	7,212
CUR	3	0	0	0	0	1	0	346	139,978	140,328
.	0	1	2	12	16	10	3	42	149	235
Total	328	8	12,836	86,995	46,648	45,866	7,212	140,226	222,430	562,549
Percentage of records with inconsistencies: 2.26										

Kintampo HDSS: description of corrections

20 March 2012

Error type	Description	Number	Percentage	Solution
Wrong IDs	Arose due to correction of wrong IDs of individuals concerned but old IDs still in database	179	15.6%	Old IDs deleted from database
Inconsistent dates	Arose due to correction of wrong dates which led to inconsistencies with other dates previously entered concerning an individual's residency	317	40.3%	Corrections made on database
Duplicates	Arose mainly due to problems with merging data tables	188	27.7%	Removed by data correction program
Hanging cases	Migrations not reconciled	462	16.7%	Removed from current analysis by data correction program
Total		1146	100.0%	

EHA model of descriptive data process: "Core residency file approach" in words

20 March 2012

- Denominator is computed from exposure using all events defining inclusion (birth, enumeration, in-migration) and exclusion (death, last observation, out-migration) at the individual level
- Numerator is also computed at individual level
- Rates are obtained by dividing numerator by denominator
- Assumption of equal distribution is not needed
  - Rates can be computed for any time interval

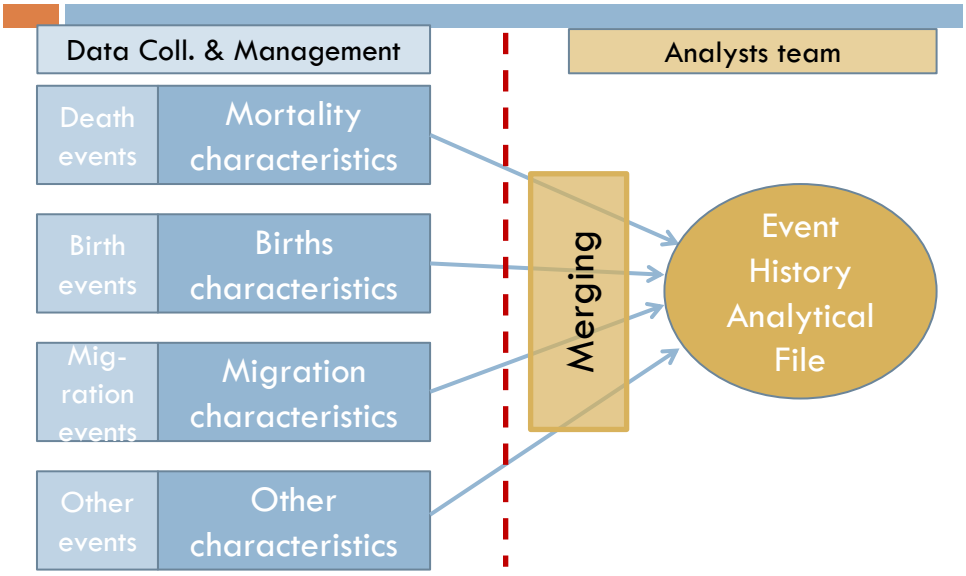
EHA model of descriptive data process: "Core residency file approach" advantages

20 March 2012

- Basic demographic rates can be computed directly from the core residency file
  - Computer capacities allow large datasets handling
  - Regular and reliable production of indicators made easier
- No risk of computing denominator wrongly:
  - No assumption of equal distribution
  - Built-in person-years at risk computation based on inclusion and exclusion criteria
- Probability-to-rate conversion (or vice versa) is not necessary

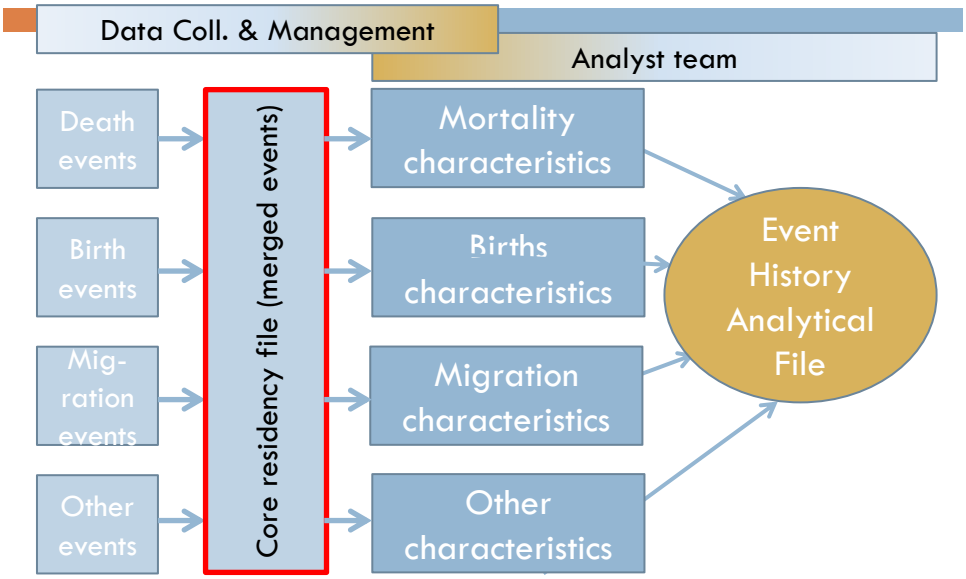
Conventional model of EHA data process: Modules

20 March 2012



Recommended integrated process: Core residency file

20 March 2012



"Core residency file approach" : advantages for determinants (regression) analysis

20 March 2012

- Core residency file is:
  - Controlled at data management level
  - Basic demographic rates and regression analysis are consistent

  ⇒ No need for (differently endowed) analysts to merge files, with risk of producing different results
- Core residency file can easily be expanded on demand:
  - Adding event attributes (e.g. cause of death, child rank, migration destination and origin…)
  - Adding other status events that do not modify exposure in the reference population (e.g. changes in employment, education, marriage, etc.)

  ⇒ Flexibility
- EHA-oriented data management training needed at Masters level (data manager, research assistant…)

20 March 2012

Many thanks for your kind attention