WP4: Data management

WASABY - Water And Soil contamination and Awareness on Breast cancer risk in Young women

Roberto Lillini, Martina Bertoldi

Milan, 19.02.2018

Index

- 1. WP4 Objectives and actions
- 2. Participating Cancer Registries
- 3. Preliminary survey
- 4. Data collection protocol: proposal

1. WP4 Objectives and actions

Objectives

- To manage data from Cancer Registries (CRs)
- To coordinate the data flow (data exchange) between WPs
- To define the model for spatial analysis on cancer incidence data

1. WP4 Objectives and actions

Actions

- A **survey to European CRs** aimed to: a) identify CRs interested to participate in the action and b) collect information on relevant data availability in the CRs, with the aid of the European Network of Cancer Registries (ENCR). A preliminary survey was sent in the writing phase of the project.

- Writing **data collection protocol** linked with the production of breast incidence maps (see "Methods and means. Identification of risk areas across European Countries"). The protocol serves to describe data collection and to prepare the material for the Ethical Committee approval.

- The protocol will be sent to the **INT Independent Ethical Committee** and where necessary also to the **local ethical committees** according to CR's requirements. After the EC approval, data collection will be opened.

1. WP4 Objectives and actions

Actions

- Preparation of a **report on the model designed** and validated for incidence rates maps. This report uses information coming from the survey to the CR, as well as from the activities conducted in WP-4, WP-5 and WP-6

- **Collection** (and linkage where necessary) of data coming from CRs, on deprivation indexes coming from WP-5 and environmental data defined in WP-7. All data collected are systematically included in the open source software chosen by WP-6

- Regarding **other confounders** different from deprivation indexes (see WP-5), availability on adherence to screening is to be investigated (through scientific literature search, online website search, direct contacts with national statistic offices, screening registries, CRs)

1. WP4 Objectives and actions

Deliverables linked to this work package

- D4.1 – CR survey (M4)
- D4.2 – Protocol for CR data collection and Ethical Committees (M5)
- D4.3 – Model on data management for spatial analysis of cancer incidence data (M24)

Milestones to be reached by this WP

- M4.1 – INT Ethical Committee (M5-M6) and other Ethical Committees approval (M7-M10)
- M4.2 – Complete list of participating CRs (M7-M8)
- M4.3 – Available data from CRs (M7-M15)
- M4.4 – Complete database for spatial analysis on breast cancer incidence (M10-M28)
- M4.5 – Complete database for pilot environmental study (M28-M32)

2. Participating Cancer Registries

- 100 ENCR CRs were contacted during the application phase with a preliminary questionnaire (in your folder) to assess interest and feasibility.
- 30 replied positively.
- In January 2018, as the project started, we asked the 30 CRs to double-check the replies they had entered, and possibly confirm interest.
- To date, 6 CRs chose to not participate: Estonia, Hamburg and Munich, Firenze-Prato and Friuli Venezia-Giulia, and Lower Silesia.
- Feedback from 4 CRs are still pending: Parma, Lithuania, Kracow, Murcia.
- **20 CRs** have informally declared their participation, representing **8 European countries**.

2. Participating Cancer Registries

Cancer Registry	CR Director	Director's e-mail
Belgium	Liesbet Van Eycken	elizabeth.vaneycken@kankerregister.org
Bremen	Sabine Luttman	luttmann@bips.uni-bremen.de
Schleswig-Holstein	Alexander Katalinic	Alexander.Katalinic@uksh.de
Napoli 3 South	Mario Fusco	mariofusco2@virgilio.it
Palermo	Francesco Vitale	francesco.vitale@unipa.it
Ragusa	Rosario Tumino	rtumino@tin.it
Siracusa	Anselmo Madeddu	rtp@ausl8.siracusa.it
Trento	Silvano Piffer	Silvano.Piffer@apss.tn.it
Umbria	Fabrizio Stracci	fabrizio.stracci@unipg.it
Varese	Giovanna Tagliabue	Giovanna.Tagliabue@istitutotumori.mi.it
Greater poland	Maciej Trojanowski	maciej.trojanowski@wco.pl
Kielce	Stanislaw Gozdz	stanislaw.gozdz@onkol.kielce.pl
Silesia	Marcin Motnyk	Marcin.Motnyk@io.gliwice.pl
Central Portugal	Manuel António Silva	
Northern Portugal	Maria José Bento	mjbento@ipoporto.min-saude.pt
Slovenia	Maja Primic Žakelj	MZakelj@onko-i.si
Basque Country	Arantza Lopez De Munain Marques	arantza-lopez@euskadi.eus
Castellon-Valencia	Ana Torrella	torrella_ana@gva.es
Girona	Rafael Marcos-Gragera	rmarcos@iconcologia.net
Granada	Maria José Sanchez Perez	mariajose.sanchez.easp@juntadeandalucia.es
Northern Ireland	Anna T. Gavin	a.gavin@qub.ac.uk

3. Preliminary survey

All the 20 participating CRs currently answered to the preliminary questionnaire.

Also the non-participating and pending CRs answered to it.

The most relevant information regarded the available geo-coding level and if/how they could improve it.

3. Preliminary survey

Level of availability

- Municipality: all the CRs.
- Zip Code (6): Ragusa, Trento, Umbria, Kielce, Central Portugal, Northern Ireland.
- Census Sections (8): Naples 3, Trento, Umbria, Slovenia, Basque Countries, Castellon-Valencia, Girona, Murcia (only for some municipalities).
- Other levels of territorial classification higher than the municipalities and/or between the municipalities and the Census sections are available for some CRs.

3. Preliminary survey

Resource available for geo-coding cases at the smallest administrative/geographic area

- Internally to the CR (13): Belgium, Bremen, Naples 3, Palermo, Ragusa, Trento, Umbria, Silesia, Central Portugal, Northern Portugal, Slovenia, Castellon-Valencia, Northern Ireland.
- Paying for a service outside the CR (4): Kielce, Girona, Granada, Murcia.
- They must obtain information on (6): Parma, Syracuse, Lithuania, Kracow, Greater Poland, Basque Countries.

3. Preliminary survey

Availability of maps

- Already available at the CR (13): Belgium, Bremen, Naples 3, Palermo, Trento, Umbria, Silesia, Central Portugal, Slovenia, Basque Countries, Castellon-Valencia, Granada, Murcia.
- Acquisition free from external sources (3): Kielce, Northern Portugal.
- Purchasable by paying (2): Kracow, Girona.
- They must gather information (6): Parma, Ragusa, Syracuse, Lithuania, Greater Poland, Northern Ireland.

4. Data collection protocol: proposal

The protocol will be only focused on performing spatial analysis for breast cancer risk for WASABY in the participating European cancer registries.

Formally a CR will be considered a participating CR when it demonstrates to be able to geo-code own cases and an ethical committee will allow it to participate in the project.

The protocol will be prepared as *vademecum* for each CR interested to join the project. WASABY allows to CRs to define incidence years and type of geographic data: thus also methods to be applied can vary according to original data received.

4. Data collection protocol: proposal

The protocol steps will be:

1. Each participating CR will need to provide the list of first invasive breast cancer cases (coded as C50 according to the ICD-10) diagnosed during a specific period (to be defined separately for each participating CR, e.g. 2000-2010), together with age at diagnosis (or 5-year age groups), morphology and data on the place of residence at the time of diagnosis (exact x and y coordinates or smallest possible sub-area of residence).
2. Socio Economic Status (SES) data will be collected as main confounder in the spatial analysis: National or European Deprivation indexes will be utilized.
3. Maps of incidence will be estimated in order to identify CR sub-areas characterized by higher-than-CR average rates.

4. Data collection protocol: proposal

FILE WITH BREAST CANCER CASES

All primary invasive female breast cancer (ICD9 174*, ICD10 C50*), selected from cancer registries data during a specific period (ex: 2000 to 2009). Cancer registration criteria follow IARC rules.

Variable name	Description	Data type
CR	Cancer Registry name	Alphanumeric variable
PATIENT_ID	Patient identification code assigned by Cancer Registry. It is necessary not only to identify a single subject, but also to retrieve all necessary health and administrative data Note: synchronous breast cancer cases must be counted once	Numeric/Alphanumeric variable
DATE OF BIRTH	Date of birth of the patient	DD/MM/YYYY
DATE OF DIAGNOSIS	Incidence date based on histological or cytological confirmation of the malignancy, if available	DD/MM/YYYY
AGE	Age at diagnosis	Numeric variable
ICD_9	Complete ICD-9 code of incident case	Alphanumeric variable
ICD_10	Complete ICD-10 code of incident case	Alphanumeric variable
ICDO3_M	ICDO3 morphology code of incident case	Alphanumeric variable
STAGE	Stage at diagnosis according to TNM stage grouping I II III IV unknown	Alphanumeric variable

4. Data collection protocol: proposal

FILE WITH GEOGRAPHIC DATA

Residence addresses at diagnosis retrieved from the National or local Security system or from the anagraphical reference of each registry will be collected.

Variable name	Description	Data type
CR	Cancer Registry name	Alphanumeric variable
PATIENT_ID	Patient identification code assigned by Cancer Registry. It is necessary not only to identify a single subject, but also to retrieve all necessary health and administrative data.	Numeric / Alphanumeric variable
OPTION 1		
X	Longitude coordinate referred to the address where the patient was residing at the moment of the breast cancer diagnosis	Numeric variable
Y	Latitude coordinate referred to the address where the patient was residing at the moment of the breast cancer diagnosis	Numeric variable
Reference	The coordinate system used for X and Y: UTM WGS84 32N vs. UTM ED 1950 32N	Alphanumeric variable
OPTION 2		
MUNICIPALITY_CODE	Code of the Municipality where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable
MUNICIPALITY	Name of the Municipality where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable
OPTION 3		
CENSUS_BLOCK	Census block where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable

4. Data collection protocol: proposal

POPULATION FILES

For every CR, WASABY needs the reference population at the same geographic level of the incident cases. More specifically, the population files must contain the female population data by 5-year age groups, calendar year within time period and sub-area on which the incidence data would be estimated. Sub-areas refer to the smallest geographical area for which required data are available and may be different across countries.

Variable name	Description	Data type
CR	Cancer Registry name	Alphanumeric variable
AGE_CLASS	5-year age class	Numeric/Alphanumeric variable
YEAR	Calendar year	Numeric/Alphanumeric variable
REF_DATE	Reference date of population data (1st Jan, 31st Dec, ecc)	Date/Alphanumeric variable
SUB_AREA	MUNICIPALITY_CODE or CENSUS_BLOCK indicated in the file with geographic data (see page 5)	Alphanumeric variable
POP	Female population by 5-year age groups, calendar year within time period and sub-area on which the incidence data would be estimated	Numeric

4. Data collection protocol: proposal

SHAPEFILES

For every CR, WASABY needs a complete shapefile of the geographic area covered by its activity. The shapefile format is a digital vector storage format for storing geometric location and associated attribute information. It consists of a collection of files with a common filename prefix (e.g., Varese.shp, Varese.dbf, Varese.shx), stored in the same directory, with mandatory and optional files.

Mandatory files:

File name	Description	Data type
(CR area).shp	Shape format; the feature geometry itself	Alphanumeric
(CR area).shx	Shape index format; a positional index of the feature geometry to allow seeking forwards and backwards quickly	Alphanumeric
(CR area).dbf	Attribute format; columnar attributes for each shape, in dBase IV format	Alphanumeric

Other optional files, regarding spatial features not reported in the .dbf file, can be added but are not needed for a correct representation.

In the .dbf file an information about the minimum geo-coding level must be reported (i.e., census block, municipality, etc.)

4. Data collection protocol: proposal

Confounders

- Socio economic status (SES) and other confounders: the European Deprivation index (EDI) (see WP5 presentation).

- Individual factors, e.g. ethnicity, family history, age, reproductive factors, alcohol intake, weight, physical activity, hormone therapy and oral contraceptives. Adherence to organized screening programmes in areas covered by cancer registries: where possible, information on adherence to organized cancer screening is to be collected at census-block level. This could extend the considered age of incidence of the patients

4. Data collection protocol: proposal

The file with confounders is structured this way:

Variable name	Description	Data type
COUNTRY	Country name	Alphanumeric variable
SUB_AREA	MUNICIPALITY_CODE or CENSUS_BLOCK indicated in the file with geographic data (see page 5)	Alphanumeric variable
SES_SCALE	European Deprivation Index or specific national deprivation indices (according to the availability in the specific CR) by sub-area of incidence data. This is a scale variable	Numeric - Scale
SES_ORDINAL	European Deprivation Index or specific national deprivation indices (according to the availability in the specific CR) by sub-area of incidence data, classified by deprivation groups. This is an ordinal variable	Numeric - Ordinal
SCR_ADH	% of screening adherence by sub-area of incidence data (if available)	Numeric - Scale