WASABY

WASABY

WAter and Soil contamination and Awareness on Breast cancer risk in Young women

D4.2 Data protocol

WP4 – Fondazione IRCCS "Istituto Nazionale dei Tumori"

V1 - 23rd February 2018

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

PROTOCOL FOR
MAPPING OF BREAST CANCER RISK FOR THE WASABY PROJECT
V1 – 23rd February 2018

1. INTRODUCTION

The WASABY project (WAter & Soil contamination and Awareness on Breast cancer risk in Young women) focuses on the geographical analysis of population based cancer incidence data in connection with environmental factors, using breast cancer and water/soil contamination as an exemplification replicable to other cancer sites. The following table presents the participants to the project:

Participants	Acronym	Country	Work Packages (WP)
FONDAZIONE IRCCS ISTITUTO NAZIONALE DEI TUMORI	INT	Italy	WP1: Coordination of the project WP4: Data management WP7: Environmental risk factors & breast cancer
ASSOCIATION EUROPEENNE DES LIGUES CONTRE LE CANCER ASBL	ECL	Belgium	WP2: Dissemination of the project
UNIVERSITÄT ZU LÜBECK	GER	Germany	WP3: Evaluation of the project
UNIVERSITE DE CAEN NORMANDIE	FRA	France	WP5: Deprivation indexes
ONKOLOSKI INSTITUT LJUBLJANA	SLO	Slovenia	WP6: Methods and analysis

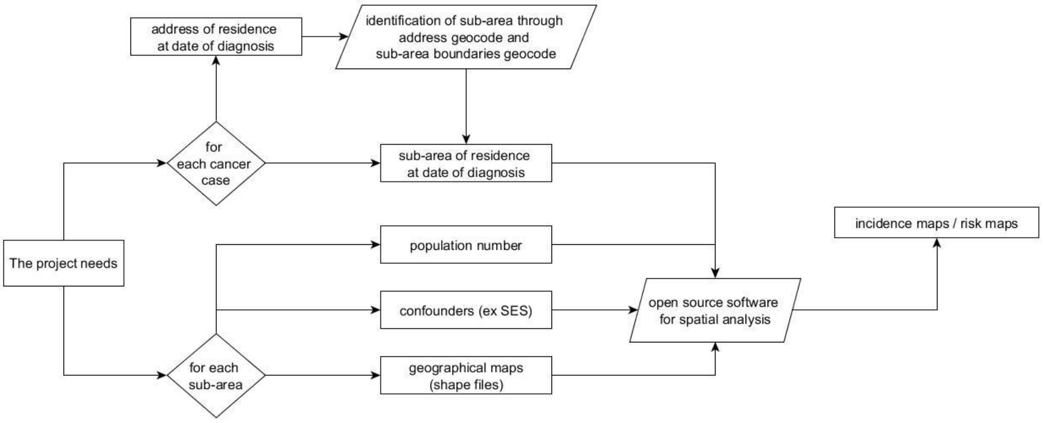
The present protocol focuses only on the WASABY project activity of spatial analysis for breast cancer risk in the participating cancer registries. The list of potential participating cancer registries (CRs) is in Annex 1. The project requires to have at least 15 CRs from 6 European countries. Formally a CR will be considered a participating CR when a) it demonstrates to be able to geo-code own cases and b) an ethical committee will allow it to participate in the project.

The present protocol is prepared as *vademecum* for each CR interested to participate in the project. WASABY allows CRs to define incidence years and type of geographic data, for this reason different methods may be applied according to original data received.

2. SINTHESIS OF THE PROJECT

Since individuals represent the basic unit of spatial analysis in cancer research, each CR shall assign geographic information (exact x and y coordinates or smallest possible sub-area of residence (SU)) to every breast cancer case corresponding to the location of their place of residence.

The following schema shows the process of data collection for the present protocol.


address of residence at date of diagnosis
identification of sub-area through address geocode and sub-area boundaries geocode
for each cancer case
sub-area of residence at date of diagnosis
The project needs
population number
confounders (ex SES)
for each sub-area
geographical maps (shape files)
open source software for spatial analysis
incidence maps / risk maps

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

According to the schema above, the protocol steps are:

1. Each participating CR is required to provide information on breast cancer cases (coded as C50 according to the ICD-10) diagnosed during a specific ten-year period (to be defined separately for each participating CR, e.g. 2001-2010), 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 SU). See Annex 2 for details on data to be collected
2. Socio Economic Status (SES) data will be collected as main confounder in the spatial analysis: National or European Deprivation indexes by SU will be utilized. See Annex 3 for details on SES and other potential confounders
3. Maps of incidence will be estimated in order to identify CR SU characterized by higher-than-CR average rates. See Annex 4 for methods to be applied according to data available in each CR.

Before the start of the data collection, every CR is required provide the following general information, which will be added to the ones collected with the preliminary survey:

- The calendar years by which the CR can provide incidence data at the most disaggregated geographic level (exact x and y coordinates or SU).
- The calendar year by which such geographic level has changed (e.g., census tract changes between two different Census data collection).
- Any confidentiality problems likely to arise if/when pursuing the approval to the Ethical Committees, locally.
- Detail limits in publishing maps (see confidentiality above).

3. Data storage

Two modalities of data storage can be envisaged:

- OPTION 1
  - Data will be centrally stored at FONDAZIONE IRCCS ISTITUTO NAZIONALE DEI TUMORI and, only for the selected number of cancer registries involved in WP6 analyses, data will be shared with the ONKOLOSKI INSTITUT LJUBLJANA
  - Data will be stored individually (but anonymously). If a CR is to send breast cancer cases by SU, data will be stored at aggregated level
  - Data will be stored in a dedicated server not connected to the web, and according to the standard requirements for data security
  - Data handling will be conform with the EC General Data Protection Regulation (2016/679)
- OPTION 2
  - Only results of the analysis (performed by the CR) will be shared in the WASABY project. This will be the case of CRs with the entire population geocoded.

4. INT contacts for data collection

Every information request and submission, in particular regarding the four points above, which update the questionnaire data, must be addressed to:

roberto.lillini@istitutotumori.mi.it (Roberto Lillini)

and CC:

lifetable@istitutotumori.mi.it (Paolo Baili)

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

5. Publication Policy of the entire project WASABY

All publications performed in the WASABY context must mention the WASABY Working Group. A suitable authorship formula being: Authors A, B, C, ... and the WASABY Working Group, with all members listed in a footnote or appendix to the article.

The WASABY Working Group will be realized with the following members:

- All members of the Steering Committee (SC)
- All members of the Management Support Team (MST)
- Up to two members of each Partner indicated in the introduction (in addition to those included in the SC and MST)
- Up to two members of each Cancer Registry participating in WASABY
- All experts actively participating in the work packages of the project
- Up to one member for each participating area working in geocoding activities (unless included in the previous points)

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 1 – TENTATIVE LIST OF PARTICIPATING CANCER REGISTRIES

Nation	Cancer Registry
Belgium	Belgium
Germany	Bremen
Germany	Schleswig-Holstein
Italy	Napoli 3 South
Italy	Palermo
Italy	Parma
Italy	Ragusa
Italy	Siracusa
Italy	Trento
Italy	Umbria
Italy	Varese
Lithuania	Lithuania
Poland	Greater Poland
Poland	Kracow
Poland	Kielce
Poland	Silesia
Portugal	Central Portugal
Portugal	Northern Portugal
Slovenia	Slovenia
Spain	Basque Country
Spain	Castellon-Valencia
Spain	Girona
Spain	Granada
Spain	Murcia
UK	Northern Ireland

ANNEX 2 – DATA TO BE COLLECTED FOR EACH CANCER REGISTRY

FILE WITH BREAST CANCER CASES AND GEOGRAPHIC DATA

Primary invasive female breast cancer (ICD9 174*, ICD10 C50*), selected from cancer registries data during a specific ten years period (ex: 2001 to 2010) are included in the project. It is mandatory to collect data with age at diagnosis less than 50 years of age, while it is not mandatory to collect data for all ages. Synchronous and metachronous breast cancer cases must be counted once. Cancer registration criteria must follow European Network of Cancer Registries (ENCR) rules.

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

Data can be collected in two different modalities.

OPTION 1 – individual level

		Variable name	Description	Data type	Mandatory
BREAST CANCER VARIABLES		CR	Cancer Registry name	Alphanumeric variable	Yes
		PATIENT_ID	Patient identification code assigned by Cancer Registry.	Numeric / Alphanumeric variable	Yes
		DATE OF DIAGNOSIS	Incidence date based on histological or cytological confirmation of the malignancy	DD/MM/YYYY	Yes
		DATE OF BIRTH	Date of birth of the patient	DD/MM/YYYY	Yes (one of the two variables)
		AGE	Age at diagnosis	Numeric variable	
		ICDO3_M	ICDO3 morphology code of incident case	Alphanumeric variable	Yes
		SUBTYPE_ER	Estrogen Receptor value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_PGR	Progesterone Receptor value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_HER2	HER-2 value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_KI67	KI-67 value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_FISH	FISH value at diagnosis	Numeric / Alphanumeric variable	No
GEOGRAPHIC VARIABLES	OPTION A	X	Longitude coordinate referred to the address where the patient was residing at the moment of the breast cancer diagnosis	Numeric variable	Yes, data from one option
		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 B	SU	Smallest administrative unit (SU) where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable	
	OPTION C	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 2 – aggregated level

		Variable name	Description	Data type	Mandatory
BREAST CANCER VARIABLES		CR	Cancer Registry name	Alphanumeric variable	Yes
		YEAR DIAGNOSIS	Incidence year based on histological or cytological confirmation of the malignancy	Numeric variable	Yes
		AGE	Age class at diagnosis	Alphanumeric variable	Yes
		ICDO3_M	ICDO3 morphology code of incident case	Alphanumeric variable	Yes
		SUBTYPE_ER	Estrogen Receptor value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_PGR	Progesterone Receptor value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_HER2	HER-2 value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_KI67	KI-67 value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_FISH	FISH value at diagnosis	Numeric / Alphanumeric variable	No
GEOGRAPHIC VARIABLES	OPTION B	SU	Smallest administrative unit (SU) where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable	Yes, data from one option
	OPTION C	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	
DATA		NR_CASES	Number of primary invasive female breast cancer by all the previous variables	Numeric variable	Yes

POPULATION FILES

For every CR, WASABY needs the reference population at the same geographic level on which that CR intends to study 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 SU (sub-areas refer to the smallest geographical area for which required data are available and may be different across countries).

All the variables are mandatory.

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
SU	MUNICIPALITY_CODE or SU indicated in the file with geographic data (see pages 5 or 6)	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

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

Files must be combined with information on calendar years of validity (in case of administrative changes of SU in the incidence years studied).

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.)

ANNEX 3 – CONFOUNDERS

Socio economic status (SES) and other confounders: Deprivation index

Since this study includes different European countries, it is important that measurement of socioeconomic deprivation be comparable or at least transferable between different European countries, despite their socio-cultural differences, to improve the comparability and reproducibility across countries. The European Deprivation Index (EDI) measures the social environment in a comparable manner across countries, despite the differences in the census variables available, and to incorporate the social and cultural specificities of each country concerned. The ecological deprivation indices are built according to shared methodological principles, by selecting fundamental needs associated with both objective and subjective poverty, and they use the same theoretical concept of relative deprivation using a European survey dedicated to relative deprivation (Eu-Silc) regularly conducted on national samples from the all European countries. The method for constructing national versions of this EDI is described in different papers [Pornet C, JECH 2012; Guillaume E, JECH, 2016] and national versions of EDI are already available for 5 European countries (Italy, Portugal, Spain, England and France). This index is based on two elements:

- The European survey on deprivation European Union Statistics on Income and Living Conditions (EU-SILC) is a cross-sectional and longitudinal sample survey providing data on income, poverty, social exclusion and living conditions in the European Union. From these data, the statistical office of the European Union (Eurostat—http://ec.europa.eu/eurostat/web/main) produces a European standardized questionnaire that is specifically designed to study deprivation. It consists of nine questions, common to European Union members, evaluating needs that directly or indirectly induce financial inability. For each European Union member, the sum of weights for the sample design and the response rate to a national questionnaire were tailored on the basis of the national population size. All analyses were weighted for non-response and adjusted for sample design, to ensure the representativeness of the results for each member.
- The ecological data of the national population censuses. Ecological data came from the last exhaustive national population censuses, which were conducted in 2001 for Italy (Italian National Institute of Statistics: ISTAT), Portugal (National Institute of Statistics: INE), Spain (National Institute of Statistics: INE) and England (Office for National Statistics: ONS), and, in 1999, for France (National Institute for Statistics and Economic Studies: INSEE). To minimize the unavoidable ecological bias as much as possible, the smallest area for which census data were available was identified.

Also in this case the efforts performed are for the number of participating countries and not for the number of participating CRs. As already mentioned, at the time of writing this methodology has been developed for 5 countries but it may replicable in other European Union member states. Therefore, we will evaluate the construction of the EDI for countries with areas covered by participating cancer registries. In countries without available data to construct the EDI the collection of national deprivation indexes will be envisaged.

Other confounders

Individual factors, e.g. ethnicity, family history, age, reproductive factors, alcohol intake, weight, physical activity, hormone therapy and oral contraceptives, have been found to influence the risk of breast cancer. Adherence to organized screening programmes in areas covered by cancer registries, lead to an increment of incidence in those areas [Pacelli, Eur J Public Health, 2014]; such information, however, is not available at individual level. Where possible, information on adherence to organized cancer screening is to be collected at SU level. If data are collected only for ages <50 screening adherence is not required.

The file with confounders is structured this way:

Variable name	Description	Data type	Mandatory
COUNTRY	Country name	Alphanumeric variable	Yes
SUB_AREA	MUNICIPALITY_CODE or SU indicated in the file with geographic data (see pages 5 or 6)	Alphanumeric variable	Yes
SES_SCALE	European Deprivation Index or specific national deprivation indices (according to the availability in the specific CR) by SU of incidence data. This is a scale variable	Numeric - Scale	Yes
SES_ORDINAL	European Deprivation Index or specific national deprivation indices (according to the availability in the specific CR) by SU of incidence data, classified by deprivation groups. This is an ordinal variable	Numeric - Ordinal	Yes
SCR_ADH	% of screening adherence by SU of incidence data.	Numeric - Scale	No

Co-funded by the Health Programme of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 4 – METHODS

Identification of risk areas across European Countries Background

We will conduct this project using open source GIS software such as QGis [http://www.qgis.org/en/site/]. Furthermore, a single ArcGis desktop 10.0 license is also available for INT group and it will be used during the project to improve, if necessary, geographical and spatial analysis. GIS system software allows users to create maps with many layers (raster or vector) using different map projections. The vector data is stored as either point, line, or polygon-feature. Different kinds of raster images are supported, and the software can georeference images. Maps can be assembled in different formats and for different uses.

Spatial analysis

When large spatial units are used, the heterogeneity of exposure and different population characteristics can be missed. On the other hand, in small spatial units, the number of cancer cases is usually low and analysing the observed spatial pattern proves to be inefficient, as the population base, from which these cases arise, is often very low too. This can lead to unstable and misleading estimates of the true value. Modern approaches to relative risk estimation often rely on smoothing methods. The basic idea of mapping smoothed estimates is to borrow information from neighbouring regions to produce more stable and less noisy estimate associated with each geographical area and thus separate out the spatial pattern from the noise [Waller LA, Applied Spatial Statistics for Public Health Data, Wiley, NJ, 2004]. Taking into account these considerations we perform different statistical methods according to the available data, as follows:

- Estimate of a census block level breast cancer incidence risk using Generalised Additive Models (GAMs), a form of non-parametric or semi-parametric regression offering the possibility to analyse contextual data while adjusting for covariates and taking into account spatial autocorrelation [Woods SN, Chapman and Hall, USA 2006]. This model takes into account the spatial dependence of the data and the incidence rate variability that is due to the small number of events per geographic unit, by using a locally weighted regression smoother to account for geographic location as a possible predictor of incidence rate [Webster T, Env Health Persp, 2008]. With this model it is possible to estimate the relative risk by adjusting for covariates.

- Estimate of a census block relative breast cancer risk, by using the Besag, York and Mollié (BYM) model [Besag J, Ann Inst Stat Math, 1991], since it assumes the existence of two sources of extra variation, one spatial and the other non-spatial. The BYM model can be specified as a generalised linear mixed model (GLMM) with Poisson response variables, and considering the expected cases as an offset. The non-spatial random effect, also called heterogeneity, is usually assumed to be distributed with zero mean and constant variance. For the random effect, which captures spatial variability, a conditional autoregressive (CAR) [Clayton DG, Int J Epidem, 1993] model is used. The BYM model enables us to obtain smoothed estimates in each sub-area and, on the other hand, to estimate the effects of possible explanatory variables, such as the deprivation index.

Open source software is used for data manipulation and statistical analyses such as R and WinBUGS.

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)