MINIMAL DATA SET TO OBTAIN OPTIMAL REGISTRATION DATA FROM PATHOLOGISTS

Mia Slabbaert, Harlinde De Schutter, Liesbet Van Eycken, Belgian Cancer Registry, Brussels.

Introduction

Whenever a hospital or a laboratory for pathologic anatomy diagnoses or treats a patient with cancer, they are both required by the law of 13/12/2006 (1) to register the case and to transfer this registration to the Belgian Cancer Registry (BCR). Since 2010, the pathologists are obliged to additionally register the diagnoses of the samples taken in the context of early cancer detection (2).

In that regard, all pathology laboratories receive data requests from the BCR at regular times. These requests encompass 4 projects: data related to Cancer and data related to Prevention (Breast, Colon and Cervix). For each project a structured file and a protocol file (text) are required.

Data collection and data transfer is followed by an extensive Quality Control of the received information in a dedicated application. To enable this crucial data cleaning, data delivery should be as adequate as possible. The use of a fixed format and specific classification, as made obligatory by the law of 2006, is therefore of utmost importance.

Accepted classifications for pathology laboratories

For Dutch speaking laboratories the CODAP-2007 classification has been developed, while French speaking laboratories can use the SNOMED 3.5 VF, by courtesy of the French government and this free of charge. As different classifications entail slightly different datasets (for Dutch/French speaking laboratories), the number of used coding languages should be limited to an absolute minimum.

Datasets for the different projects (Cancer/Prevention)

The variables included in the dataset of the structured file are limited to those considered indispensable and suitable to be assessed. The dataset for the **Cancer file** is based on datasets used by cancer registries over the world according to *international guidelines*. The dataset for the **Prevention files** is based on the *needs of the different centers responsible for organized cancer detection in Belgium.*

While the data of the Cancer files are completely managed by the BCR (from the reception till the publication of the results), the information of the Prevention files is used to support/sustain the centers responsible for organized cancer detection in Belgium in obtaining their goals.

As a manual extraction of the demanded dataset information from the pathology protocols at the BCR would be labor extensive, pathologists are asked to actively collaborate by providing encoded information in separate fields. Accurate encoding is necessary to enable dedicated IT-applications to extract and gather these fields into a coded data file.

Short description of the different variables present in one or more datasets

Within the dataset of the structured file, some variables are compulsory while others are optional. The content and format of the different variables is well specified in a document "BCR-protocol" that accompanies every data demand.

The identification of the patient is of upmost importance. Preferably, the National Social Security Number (INSZ/NISS) is provided, which the BCR is authorized by law to use as the unique patient identifier. In case the INSZ/NISS is lacking, one can deliver the last name, first name, sex, date of birth and postal code.

Date of death may be registered (optional variable) but is mostly unknown by the pathologist.

The country code is an important and obligatory variable because only patients with an official residence in Belgium are eligible to be included in the incidence numbers of Belgium.

The specimen number or protocol number is indispensable to link the structured data to the corresponding accompanying protocol present in the protocol file.

The date the specimen is taken is mandatory, as this date is used to establish the incidence date (date of the first microscopic evidence of the malignant tumour). Without a sample date, it is unclear which incidence year must be taken into account and as a consequence makes the registration untreatable.

The requesting hospital (being the hospital that asked for the pathology examination) is an optional variable. Some important information as laterality, the segment of the colon or the real incidence data is not always available to the pathologist. With the help of this variable, the BCR data managers can easily identify the center from which additional information can be obtained.

The diagnostic procedure is an optional variable that is only present in the dataset for SNOMED users, and refers to the procedure that has led to the microscopic cancer diagnosis.
For Cancer files, this variable enables to distinguish between cytology and histology, and to know if the diagnosis has been made on the examination of the primary tumour or its metastasis. For the Prevention files (cervix), this variable adds in establishing the amount of diagnoses based on pap smears and in verifying if an aberrant pap-result is followed by a biopsy (fail safe mechanism). The variable also helps to examine which cytological results triggered a biopsy.

The organ code is the localization code of site from which the diagnostic sample was taken. This can be the localization of the primary tumour but also the place of nodal or distant metastasis. It is recommended to provide detailed information about the specific place on the skin, in the colon, …. because, in conformity to specific cancer registration rules, tumours occurring at different, well-defined segments on the skin or colon are considered and registered as multiple tumours.

The lesion code (CODAP) or the **morphology code** (SNOMED), provides information on the cytology/histology of the lesion.

When a primary tumour has been resected, it is sometimes possible to determine a **pTNM**. Although since several years, the pTNM classification is mostly present in the pathological reports, a manual

extraction of the information by the BCR data managers is very time-consuming. Additional tools such as textual recognition are sometimes useful but inevitably associated with poorer data quality. Therefore, the collaboration of the pathologist is asked to provide this information in a coded data field.

A pathologist can inform us about the **certainty of the diagnosis** in an optional field.

Some variables are specific to the cervix Prevention file, such as the quality of the specimen and the variables describing the HPV test results.

The **quality of the specimen** (of pap smears) is important when one wants to set up a fail-safe system in which patients with an abnormal result or with a smear that not allows an evaluation, get the adapted care.

As cervical cancer is **HPV**-related and prevention campaigns include vaccination of young girls, data on HPV test results are very informative. In case an HPV test is performed, the presence of HPV, ideally including details concerning the detected subtypes, should be communicated if possible.

The Cancer Registry uses the **nomenclature numbers of the cervixfile** to make the difference between cytology and histology (extra control), screening and follow up, first exam or second reading,… The nomenclature numbers are also used to create correct exclusion lists for the cervical screening program to prevent unnecessary invitation of patients from which in fact an cervical smear was taken, but not always reimbursed by the health insurances (using only the data of the health insurances would invite too many women unnecessarily).
The nomenclature numbers provided by the pathologists will only be used in the BCR and never be communicated to external partners with a link to a certain lab.
This variable has been recently modified from obligatory to optional because of interpretation problems.

<u>Creation of files for the Belgian Cancer Registry</u>

The creation of the different **structured files**, asked for by the BCR, is depending on the use of specific inclusion criteria.
For the Cancer file, the inclusion criteria are based on the lesion/morphology codes regardless of the sample in which a certain pathology was found.
For the Prevention files, the inclusion criteria are based on the organ codes, regardless of the microscopic findings.

This has important consequences :

- Prevention files also contain negative test results or samples not allowing to make a diagnosis (eg because of insufficient quality of the sample)
- detected cancers of breast, cervix and colon will be present in two files : the Cancer file and the specific Prevention file.
- Absence of an organ code will make extraction for a Prevention file impossible
- Absence of a lesion/morphology code will make extraction for a Cancer file impossible.

Each structured file has to be accompanied by a **separate file** in which the **according protocols** are gathered. These protocols have to be anonymized and foreseen by a unique protocol number to enable linkage to the coded information of the structured file.

What will be done with this information ?

Cancer file : once data of the laboratories for pathologic anatomy or the oncological care programs are received at the BCR, they are made ready for import in the BCR-specific application. The data belonging to the same patient (INSZ/NISS) is coupled and compared to end up with **one registration for each primary tumour. This registration needs to be as complete as possible.** 75% of all received records are treated in an automated way, but yearly about 120.000 of the 480.000 received cancer related records are investigated more in depth. With the help of these detailed revisions, a Cancer Registry database of high quality is established.
Data from the Cancer Registry database are used to obtain incidence and prevalence numbers per year, region, tumour type, tumour stage, to perform tumour-specific survival analyses, to calculate Quality-Indicators for different tumour types with special interest in the diagnostic and therapeutic procedures, to create maps on (trends in) cancer incidence, to support research,…

Data of the colon file will be used to evaluate the recently started screening program e.g. to get an idea of the consequences of a positive iFOBT or a negative one, to learn more about interval cancers (cancers detected after a negative colonoscopy and before the next planned screening), …

Data of the breast file are compared with the results of the mammography results to study the consequences of positive/negative radiologic features, to calculate quality indicators and to get insight in interval cancers (cancers detected after a negative mammography and before the next screening examination).

Data of the cervix file are used to obtain an efficient *call-recall invitation model* to support the screening program. This includes the creation of exclusion lists, in order to limit the invitations to those women who really need a pap smear (eg all women who got a pap smear in recent history – regardless of reimbursement by the government - are excluded). Data are also used to set up a *fail-safe mechanism* : is an aberrant pap-result followed by adequate diagnostic or therapeutic actions ? Which actions are undertaken in case of an insufficient sample that not allowed any evaluation ?

As forewarned is forearmed, we hope that this additional information about the dataset and the underlying reasoning behind, will enhance the motivation to deliver the most complete and accurate data, including the expected format. As clarified, all of the delivered information is considered valuable, and is used for a huge variety purposes. In order to establish a firm collaboration with a limited help of IT-staff, the minimal data sets will be kept unchanged as long as possible. The accompanying document (the so called BCR-protocol) may be adapted e.g. to include extra explanatory information.

(1) *Wet houdende diverse bepalingen betreffende gezondheid van 13 december 2006, artikel 39. Belgisch Staatsblad, 22 december 2006 ; Loi portant dispositions diverses en matière de santé du 13 décembre 2006, article 39. Moniteur Belge, 22 décembre 2006.*
(2) *Wijziging van de Wet houdende diverse bepalingen inzake gezondheid van 13 december 2006, Belgisch Staatsblad, 2 juni 2010 ; Modifications de la Loi portant dispositions diverses en matière de santé du 23 décembre, Moniteur Belge, 2 juin 2010.*

	VARIABLES FOR CODAP users	DATASET FOR CANCER DIAGNOSES	DATASET FOR BREAST AND COLON PREVENTION FILE	DATASET FOR CERVIX PREVENTION FILE
		following international guidelines for cancer registries	***According to the need of the Centers for Cancer Detection in Belgium***	
1	National Social Security Number (INSZ/NISS)	C	C	C
2	Last name	O/C	O/C	O/C
3	First name	O/C	O/C	O/C
4	Sex	C	C	C
5	Date of birth	C	C	C
6	Date of death	O	O	O
7	Zip code = postal code	C	C	C
8	Country code	C	C	C
9	Specimen number	C	C	C
10	Date specimen was taken	C	C	C
11	Requesting hospital	O	O	O
12	RIZIV/INAMI number of the demander of the test		C	C
13	Quality of the specimen			C
14	Organ	C	C	C
15	Lesion	C	C	C
16	pT	O/C*		
17	pN	O/C*		
18	pM	O/C*		
19	Degree of certainty	O	O	O
20	HPV **high risk** test results			C if HPV test performed
21	HPV **high risk** types detected			O
22	Nomenclature number(s)		O	O

O = optional ; C = compulsory
O/C : compulsory if INSZ not delivered ; optional when INSZ is delivered
O/C* : only compulsory if TNM-classification applicable for the coded specimen
/// : not applicable

For details on the specific variables, see BCR-protocol (www.kankerregister.org or www.registreducancer.org).

Table 1 : dataset for CODAP-users for the different projects

	VARIABLES FOR SNOMED USERS	DATASET FOR CANCER DIAGNOSES	DATASET FOR BREAST AND COLON PREVENTION FILE	DATASET FOR CERVIX PREVENTION FILE
		following international guidelines for cancer registries	***According to the need of the Centers for Cancer Detection in Belgium***	
1	National Social Security Number (INSZ/NISS)	C	C	C
2	Last name	O/C	O/C	O/C
3	First name	O/C	O/C	O/C
4	Seks	C	C	C
5	Date of birth	C	C	C
6	Date of death	O	O	O
7	Zip code = postal code	C	C	C
8	Country code	C	C	C
9	Specimen number	C	C	C
10	Date specimen was taken	C	C	C
11	Requesting hospital	O	O	O
12	RIZIV/INAMI number of the demander of the test		C	C
13	Quality of the specimen			C
14	Diagnostic procedure	O or HR	O or HR	HR
15	Organ	C	C	C
16	Lateralization	O	O	
17	Morphology	C	C	C
18	Differentiation level	O		
19	pT	O/C*		
20	pN	O/C*		
21	pM	O/C*		
22	Degree of certainty	O	O	O
23	HPV **high risk** test results			C if HPV test performed
24	HPV **high risk** types detected			O
25	Nomenclature number(s)		O	O

O = optional ; C = compulsory ; HR : highly recommended if not feasible to deduce from other variables
O/C : compulsory if INSZ not delivered ; optional when INSZ is delivered
O/C* : only compulsory if TNM-classification applicable for the coded specimen
/// : not applicable

For details on the specific variables, see BCR-protocol (www.kankerregister.org or www.registreducancer.org).

Table 2 : dataset for SNOMED-users for the different projects