# The Belgian Virtual Tumourbank (BVT) Project: Data Quality Control

Belgian Cancer Registry

**Araceli Diez-Fraile[1], Eva Van der Stock[1], Karen Vos[1], Mia Slabbaert[1], Liesbet Van Eycken[1], Vincent Grégoire[2] on behalf of the Steering Committee of the Belgian Virtual Tumourbank**

**(1) Belgian Virtual Tumourbank - Belgian Cancer Registry (BVT-BCR), Koningsstraat 215 bus 7 - Rue Royale 215 boîte 7, 1210 Brussels (Belgium).**
**(2) Radiation Oncology Dept & Center for Molecular Imaging and Radiotherapy and Oncology (MIRO), Université Catholique de Louvain, St-Luc University Hospital, 10 Ave Hippocrate, 1200 Brussels (Belgium).**

## Introduction

The BVT is a tissue-banking web application for biospecimen annotation and localisation among eleven Belgian university hospitals. Data quality control (QC) of the BVT includes control measures at every stage of the data process (Fig.1) guaranteeing a high quality of the data on the biospecimens requested by scientists working in translational research in oncology. It has been previously emphasised that linkages of cancer and biobank registry data optimise their value in cancer research [1]. However, to the best of our knowledge, a validity assessment linking information from biological samples to population-based cancer registries has not been previously reported in the scientific literature.

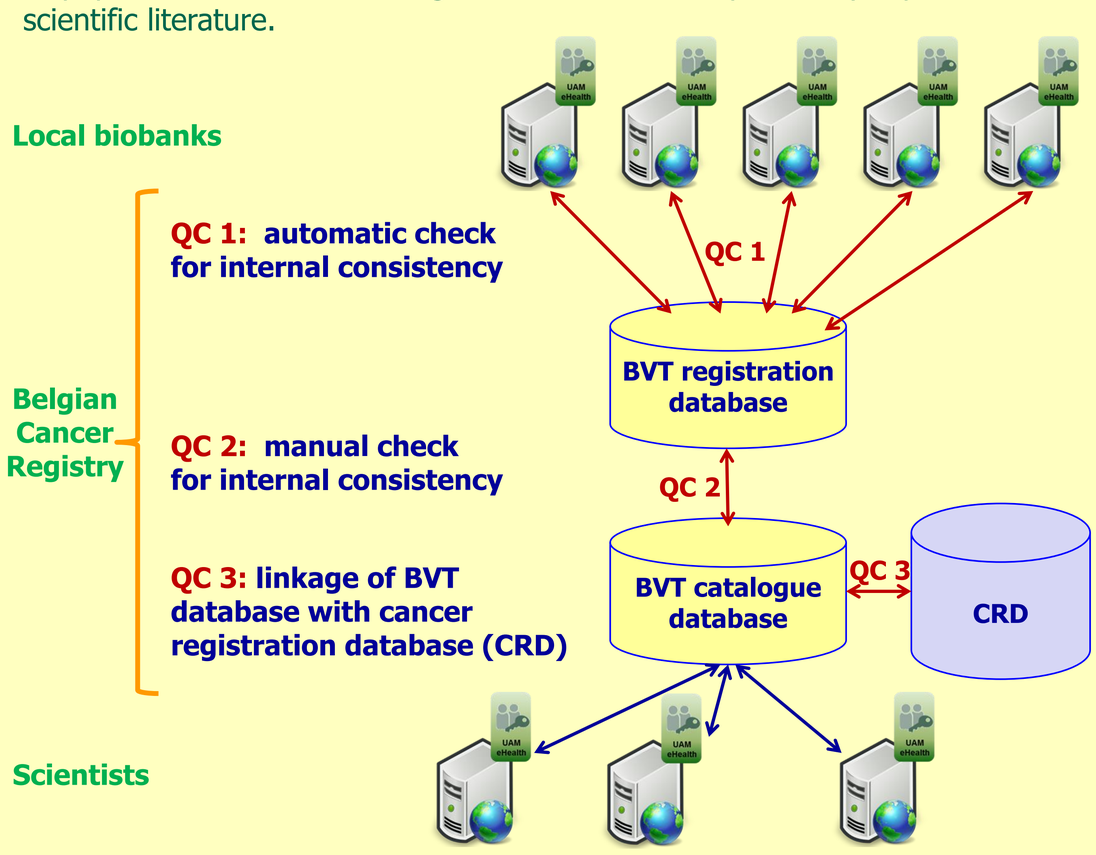


**Figure 1.** Organisational scheme and steps of data quality analysis of the BVT system with decentralised input by authorised local biobanks, the BVT central databank with two modules (registration and catalogue) and authentication service providing authorised users access to the data. The BVT catalogue database is the interface that allows queries based on patient, technical and oncological variables.

## Aim

To perform a first formal evaluation of validity of the BVT database.

## Methods

In this pilot study, the primary renal tumours collected during the year 2010 and registered in the BVT database were contrasted with the Belgian cancer registration database (CRD) by linking both databases using the social security identification number (SSIN). Registrations lacking the SSIN and patients that could not be found at the CRD on basis of the SSIN were linked by means of the laboratory, biopsy number, birth date and sample date. When the case could not be validated with the CRD, the pathological report and/or information from medical files was requested from the corresponding local biobank.

In analogy with a previous hallmark article for evaluating data quality in cancer registries [2], validity for the current study was defined as the proportion of cases in the BVT with a given characteristic which truly have this attribute. Validity measures within this study include the proportions of overlapping variables applicable for renal tumours (Table 1).

**Table 1.** Overlapping information between the BVT dataset and CRD that apply for the current study.

| Patient variables | Technical variables | Oncological variables |
|---|---|---|
| SSIN | Sample date | Sample type |
| Gender | Biopsy number | Sample localisation |
| Birth date | | Morphology (histology and behaviour) |
| | | pT * |

*Non-mandatory field for the BVT database and CRD.

## Results

A total of 247 registrations corresponding to 192 biopsy samples from 190 patients with primary renal tumours for the year 2010 and originating from 6 local biobanks were retrieved from the BVT registration module in June 2013. Eleven patients (5.7%) were reported to have non-malignant tumours (i.e. benign and borderline). The available material comprises majorly fresh frozen (n=192; 100%) and formalin-fixed paraffin-embedded (FFPE) tumour tissue samples (n=149; 76.6%). The registered number of corresponding normal tissue samples consists of 21.9% (n=42) for fresh frozen and 11.5% (n=22) for FFPE biospecimens.

A total of 189 biopsy samples out of 192 underwent a validity assessment for the overlapping information in both databases (i.e. BVT and CRD). The 3 additional biopsy samples, corresponding to non-malignant cases from three patients, were validated with the pathological report and/or information from medical files provided by the corresponding local biobank (Table 2). Two registrations were added to the CRD as consequence of this investigation.

**Table 2.** Validity of primary renal tumours from biopsy samples (n=192) collected in 2010 and registered at the BVT.

| Variables | Registered in BVT database n (%) | Mismatch in BVT database n (%) | Mismatch in CRD n (%) |
|---|---|---|---|
| SSIN | 185 (96)[1] | 0 | 1 (0,5) |
| Birth date | 192 (100) | 2 (1.0) | 0 |
| Gender | 192 (100) | 1 (0.5) | 0 |
| Sample date[2] | 192 (100) | 7 (3.6) | 18 (9.4) |
| Biopsy number | 74 (38.5)[1] | 4 out of 7 (57.1)[3] | 0 out of 7[3] |
| Sample type | 192 (100) | 1 (0.5) | 0 |
| Sample localisation | 192 (100) | 4 (2.1) | 1 (0.5) |
| Histology | 192 (100) | 8 (4.2) | 2 (1.0) |
| Behaviour | 192 (100) | 4 (2.1) | 2 (1.0) |
| pT | 74 (38.5) | 2 (2.7) | 0 |

[1] Although SSIN and biopsy number are mandatory fields for the local biobanks, missing information could be observed for both variables since registrations in the absence of data for mandatory fields may be uploaded by BVT data managers. Two cases lacked the SSIN at both databases. All other missing information concerning the SSIN and biopsy number could be retrieved from the CRD database.

[2] Mismatches for the variable sample date were defined as deviations larger than one month earlier or three months later for the sample date at the BVT as compared to the incidence date at the CRD. One out of 7 biopsy samples defined as mismatches was, indeed, incorrectly annotated at the BVT database. Four additional cases concerned recurrent tumours, and the sample date of the 2 additional cases was correctly registered. All mismatches detected at the CRD were due to an error during data handling for the registrations of August 2010 from one laboratory.

[3] The biopsy numbers from the patients without SSIN (n=7) were evaluated. Upon linkage of the databases we observed that for one biobank (n=4) the biopsy numbers registered at the BVT did not correspond to the ones registered at the CRD.

## Conclusions

- Both BVT database and CRD were found to have a good quality of data. Due to a more severe data cleaning, less errors can be observed for the CRD as compared to the BVT database.
- Currently, recurrent tumours in the kidney are improperly classified as primary renal tumours in the BVT application. Identification of recurrent tumours for the data published in the BVT catalogue can be done by coupling the BVT to the CRD database when biospecimens are requested for research. In addition, the possibility to specifically code recurrent tumours will be added to the BVT application.
- This report reinforces the relevance for a joint evaluation of biobank and cancer registry information in order to guarantee the highest quality of associated data from biospecimens used in translational cancer research.

## References

[1] Langseth et al. (2010) Acta Oncologica, 49:368-377.
[2] Bray and Parkin (2009) European Journal of Cancer, 45:747-755.

## Acknowledgements

The BVT project is financed by the Belgian Ministry of Public Health as part of the Belgian Cancer Plan.