# Analysis of data elements in cancer registries for defining single consistent clinical dataset

Sachiko Okada[*1] Paul Taylor[*2] Navin Ramachandran[*3] Ian McNicoll[*4]
Wai Keong Wong[*3]

[*1]Seta Clinic
[*2]Centre for Health Informatics & Multiprofessional Education, University College London
[*3]University College Hospital [*4]openEHR Foundation

# Analysis of data elements in cancer registries for defining single consistent clinical dataset

Sachiko Okada[*1] Paul Taylor[*2] Navin Ramachandran[*3] Ian McNicoll[*4]
Wai Keong Wong[*3]

[*1]Seta Clinic
[*2]Centre for Health Informatics & Multiprofessional Education, University College London
[*3]University College Hospital [*4]openEHR Foundation

Cancer registries provide information for cancer prevention, diagnosis and treatment. However, collecting detailed data required by each of the registries causes problems for the front line clinicians who have to record the information. Defining a single consistent dataset including all required data would reduce the effort involved in data collection. The aim of this study was to compare datasets in multiple cancer registries and to clarify the differences between them, to explore the possibility of defining a single consistent clinical dataset for cancer. Prostate cancer was selected as an exemplar target. All data elements from five cancer registries in UK were categorized in six groups: demographics, referral, imaging, diagnosis, treatment, and miscellaneous. The definitions of data elements were checked in detail by the first author. A domain expert judged the relationship between similar elements and explained the reason for the judgements when it was not clear.As a result, a set of unique data elements was created by eliminating overlaps in data elements. In addition, similar but distinct data elements were grouped together. The total number of data elements (451) was reduced by 22% from the simple sum (581) of the entries in each registry. There are large differences in treatment and miscellaneous, which reflects the interests of each registry. Although the ratios vary by groups, from 40% (25/62) for demographics to 83% (33/40) for referral, it is anticipated that the single consistent clinical dataset will improve the efficiency of data gathering.

Keywords: data set, Cancer registry

## 1. Backgrounds

As cancer is one of the major as well as severe diseases worldwide, collecting clinical data for cancer is critical both for research and for each patient's treatment.

There are many independent cancer registries, each of which has a detailed specification, reflecting the priorities of different agencies. Differences lie in various parts, for example, in target, e.g. type of cancer, viewpoint, i.e. whole process or focused only to some part such as chemotherapy and pathology, granularity of data, and so on. Therefore, hospitals currently collect data for each registry separately, which causes problems for the front line clinicians who have to record the information required by each of the registries.

Therefore, if difference between registries' datasets are made clarified and data elements are reduced by eliminating overlaps, it would reduce cost of data collection. Moreover, defining a single consistent dataset including all required data would be beneficial in viewpoint of data quality.

## 2. Aim

The aim of this study was to compare datasets in multiple cancer registries and to clarify the differences between them in order to explore the possibility of defining a consistent clinical dataset for cancer.

## 3. Methods

Prostate cancer was selected as an exemplar target for this research, because prostate cancer is the most common solid cancer in men in UK and US.

We chose five registries below as sources of data elements, because they are all the registries that University College London Hospital have to submit data to, for prostate cancer patients.

a) Cancer Outcomes and Services Dataset (COSD) [1)]

b) Systemic Anti-Cancer Therapy Dataset (SACT) [2)]

c) National Radiotherapy Dataset (RTDS) [3)]

d) 100,000 Genomes Project (Genomics) [4)]

e) National Prostate Cancer Audit (NPCA) [5)]

Firstly, all data elements belonging to the five registries were listed up. Genomics dataset includes data elements about genome sample element, but we used only part of datasets from Genomics, e.g., core clinical data. We excluded include essential sample metadata because it was information about the sample and participant, and more focused on sample.
Next, all the data elements are categorized in six groups, that is, diagnosis, imaging, demographics, treatment, referrals and miscellaneous (misc), firstly by the sections in each registry and secondly by the name of the data elements, for example, such as 'patient characteristic' in NPCA for demographics and 'Care episode – clinical diagnosis (ICD)' in RTDS for diagnosis.
Table 1 shows the numbers of data elements belonging to each group.

Table 1 numbers of data elements belonging to each group

| | diagnosis | imaging | demographics | treatment | referral | misc | total |
|---|---|---|---|---|---|---|---|
| COSD | 84 | 9 | 19 | 35 | 22 | 32 | 201 |
| Genomics | 58 | 10 | 8 | 19 | 3 | 78 | 176 |
| SACT | 5 | 0 | 10 | 27 | 0 | 1 | 43 |
| RTDS | 8 | 0 | 18 | 45 | 14 | 27 | 112 |
| NPCA | 18 | 0 | 7 | 18 | 1 | 5 | 49 |
| total | 173 | 19 | 62 | 144 | 40 | 143 | 581 |

Then, the definitions of data elements described in the specifications were checked in detail. Domain experts judged the relationship between similar elements across registries. Data elements which they judged equivalent were cross referenced to each other. During the process, staffs without domain knowledge supported experts' task and asked about the reason of the judgement.
As for some data elements, domain experts judged that they are not exactly the same, but quite similar. We put those data elements together as a group.

## 4. Results

### 4.1 Data elements and groups

Table 2 is a part of our results as an example from the diagnosis category.
Here the data element 'primary diagnosis' is included in COSD, Genomics, SACT, and RTDS but not in NPCA. SACT requires the primary diagnosis at start systemic anti-cancer therapy, but COSD document said 'The primary diagnosis is normally agreed at the MDT Meeting where the patient is discussed.' so we treated SACT 10 element as different from others but put all of them in one group.
The data element 'tumour laterality' is only in COSD.

Table 2 Sample data elements

| COSD | | Genomics | | SACT | | RTDS | | NPCA | |
|---|---|---|---|---|---|---|---|---|---|
| CR0370 | PRIMARY DIAGNOSIS (ICD) | 33183.1 | Diagnosis (ICD) | | | CD2 | PRIMARY DIAGNOSIS (ICD) | | |
| | | | | 10 | Primary_ diagnosis | | | | |
| CR0380 | TUMOUR LATERALITY | | | | | | | | |
| CR2030 | DATE OF DIAGNOSIS (CLINICALLY AGREED)* | | | | | | | 10 | DATE OF DIAGNOSIS (CLINICALLY AGREED) |

### 4.2 Number of data elements

Table 3 shows the numbers of source elements, data elements after eliminating overlaps and groups belonging to each group.

Table 3 number of data elements after eliminating overlaps

| | diagnosis | imaging | demographics | treatment | referral | misc | total |
|---|---|---|---|---|---|---|---|
| source | 173 | 19 | 62 | 144 | 40 | 143 | 581 |
| element | 129 | 12 | 25 | 116 | 33 | 136 | 451 |
| (ratio) | (75%) | (63%) | (40%) | (81%) | (83%) | (95%) | (78%) |
| group | 106 | 12 | 23 | 106 | 33 | 136 | 416 |

As shown in Table 3, the number of all the data elements extracted from five registries were 581, 438 when excluding misc. After eliminating overlapping elements, the total number of the data elements became 451 (78% of simple sum of source elements), 315 (72%) without misc. Here, we counted data elements with different timing separately, because each of them will require input as one data element.
Table 4 describes the number of data elements grouped by how many sources each of them has. For example, Primary diagnosis (ICD) in Table2 has three sources, and Tumour Laterality has only one source.

Table 4 unique data elements and number of overlaps

| no of sources | diagnosis | imaging | demographics | treatment | referral | misc |
|---|---|---|---|---|---|---|
| 1 | 89 | 5 | 9 | 97 | 26 | 127 |
| 2 | 34 | 7 | 5 | 11 | 7 | 8 |
| 3 | 6 | 0 | 4 | 6 | 0 | 1 |
| 4 | 0 | 0 | 4 | 2 | 0 | 0 |
| 5 | 0 | 0 | 3 | 0 | 0 | 0 |
| total | 129 | 12 | 25 | 116 | 33 | 136 |

As for groups, the total numbers of the groups by category are also shown in Table3.

## 4.3 Coverage in clinical studies

We compared data elements used in the recent prostate cancer clinical studies and the data elements in our dataset in order to evaluate the coverage of our dataset in practical clinical studies.

First, we collected 4 journal articles, for clinical studies as type, including 'prostate cancer' in title or in abstract from Pubmed. We excluded studies for patients before diagnosis, then chose top 4 articles when sorted by publication date [6-9,11].

Next, we listed up all the data elements in patient recruitment section, because data elements used in intervention and evaluation are very specific to each research protocol. On the other hand, data elements used for patient screening are considered more common and it will be beneficial if they are standardized when extracting data from EMRs or EHRs.

Then, we checked if they were included in our dataset. Most data elements in the articles has complicated structure such as 'metastatic prostate cancer'. In such cases, we divided it into basic elements such as 'metastasis', 'prostate cancer'.

As a result, there were 43 parameters in patient recruitment section in the four articles and they were divided into 73 basic parameters. Among them, 42 (58%) were included in our dataset and 31 (42%) were not. As for 43 original parameters in the articles, 16 (37%) are assumed to be covered using data elements in our dataset and 27 (63%) aren't.

# 5. Discussion

## 5.1 Related work

A lot of research has been done related to clinical dataset. The authors found more than one thousand articles with 'clinical', 'dataset' and 'cancer' in titles/abstract in PubMed, and 163 reviews for these 10 years among them. Most of them were about clinical studies using dataset, not focusing on dataset itself.

Several projects have been trying to standardize clinical dataset using clinical experts' knowledge[12-14]. However, the authors could not find any studies about comparison of data elements in each dataset so far.

## 5.2 Evaluation of results

### 5.2.1 Number of data elements

As shown in Table 3, the number of data elements was reduced about 20% by eliminating overlaps in all categories. But the degree of reduction varies according to the categories, that is, around 40% in demographics whereas less than 20% in misc, treatment and referral.

There are more treatment-specific data elements in treatment and referral category. For example, SACT requires detailed information about chemotherapy in treatment category. RTDS needs detailed appointment information in referral category. It can be assumed that conduct radiotherapy needs special facilities and referral to them, while chemotherapy can be done within clinic.

On the other hand, most registries require patient identification and basic characteristics such as gender and date of birth, which is considered the reason of the highest percentage in demographics category.

Zachary suggested to set a much smaller minimum core data set for public research and additional data elements for research [10]. In fact, COSD has such structures, core and other specific areas such as urology and lung. In addition, Genomics and NPCA provides linkage information between their data elements and elements in COSD in their specifications. However, we had to look into the definition of each element carefully because some of them look similar but are slightly different such as data acquisition timing. More description would be beneficial about the linkage and difference between other registries in each registry's specification.

### 5.2.2 Coverage in clinical studies

We found about one third of parameters used in clinical studies were covered in our dataset. Example of uncovered data elements are 'life expectancy', 'ongoing chemotherapy', 'cognitive impairment' and 'treatment end date'.

We use registries as sources of data element, which harnesses type of data, because cancer registries doesn't collect status, whole patient condition such as concurrent disease which are required for patient recruitment for clinical research. Registries collect common data, and they don't include specific data required for prospective clinical research.

During the evaluation, we faced several difficulties in translating clinical phrases into simpler data

elements, especially staffs without clinical knowledge. Clinical doctors can judge easily those information such as 'disease progression' from patients' overall condition or derive them from consult letter. But covering this information using simple dataset is not so easy.

In addition, we found that the some data element required in patient recruitment are not covered in our dataset, though very similar elements are included. For example, values such as performance status and PSA values at certain time points are required, but data element in our dataset is time-specific such as 'PSA at diagnosis'. In order to solve such problems, longitudinal data structure is preferable and it will enhance our dataset use in clinical practice and studies.

## 5.3 Limitation

We selected prostate cancer as target of analysis. It is true that each type of cancer has its special variables and options according to the characteristic, such as TNM staging, specific tumour marker, and specific investigation. In fact, 55 data elements (12%) in our dataset are prostate cancer specific. We need more different data elements for another type of cancer. However, the method can be used to for analysis of other cancer and the type of difference is not unique for prostate cancer. Therefore this method is considered generally available.

## 5.4 Future work

In this research, we tried to obtain a clinical dataset for prostate cancer from multiple dataset of cancer registries.

As the next step, we are planning to organize this dataset into clinical concepts, which will be used for an archetype-based information model to support a range of applications, including one for data entry. It is expected to enhance efficiency in data input by using these integrated dataset.

# 6. Conclusion

The aim of this study was to compare datasets in multiple cancer registries and to clarify the differences between them in order to explore the possibility of defining a consistent clinical dataset for cancer. Data elements collected from five registries about prostate cancer were checked in detail and same and similar elements are put together in groups. The total number of data elements were reduced by about 20% after eliminating overlaps. It is anticipated that the single consistent clinical dataset will improve the efficiency of data gathering

# References


[1] Cancer Outcomes and Services Dataset (COSD). http://www.ncin.org.uk/collecting_and_using_data/data_collection/cosd.

[2] SACT homepage. http://www.chemodataset.nhs.uk/home.

[3] National Radiotherapy Dataset - RTDS. http://www.natcansat.nhs.uk/rt/rtds.aspx.

[4] The 100,000 Genomes Project | Genomics England. https://www.genomicsengland.co.uk/the-100000-genomes-project/.

[5] National Prostate Cancer Audit: About. http://www.npca.org.uk/.

[6] Bahl A et al. Final quality of life and safety data for patients with metastatic castration-registantprostate cancer treated withcabazitaxel in the UK Early Access Programme(EAP(NCT01254279). BJU International 2015; 116: 880-887.

[7] de Bono JS et al. Peednisone plus cabazitaxelor mitoxantrone for metastatoc castration-resistant prostate cancer progressing after docetaxel treatment:a randomised open-label trial. Lancet 2010; 376: 1147-1154.

[8] Moore CM et al. Determination of optimal drug dose and light dose index to achieve minimally invasive focal ablation of localised prostate cancer using WST11-vascular-targeted photodynamic (VTP) therapy. BJU International 2014; 116: 888-896.

[9] Renard-Penna R et al. Multiparametric Magnetic Resonance Imaging Predicts Postoperative Pathology but Misses Aggressive Prostate Cancersas Assessed by Cell Cycle Progression Score. The Journal of urology, 2015; 194: 1617-1623.

[10] Zhang AY et al. Effects of Patient Centered Interventions on Persistent Urinary Incontinence after Prostate Cancer Treatment: A Randomized, Controlled Trial. The Journal of Urology 2015; 194:1675-1681.

[11] Kench JG et al. Dataset for reporting of prostate carcinoma in radical prostatectomy specimens: recommendations from the International Collaboration on Cancer Reporting. Histopathology2013; 62: 203-218.

[12] Grasner JT et al. EuReCa and international resuscitation registries. Current opinion in critical care 2015; 21(3): 215-219.

[13] Weintraub WS et al. ACCF/AHA 2011 key data elements and definitions of a base cardiovascular vocabulary for electronic health records: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Clinical Data Standards. Journal of the American College of Cardiology. 2011;58(2):202-22.

[14] Zachary I, Boren SA, Simoes E et al. Information Management in Cancer Registries: Evaluating the Needs for Cancer Data Collection and Cancer Research. Online Journal of Public Health Informatics.2015: 7(2):e213,.