A Pilot Interactive Data Viewer for Cancer Screening

Ladislav Dušek, Jan Mužík, Matěj Karolyi,
Michal Šalko, Denisa Malúšková, Martin Komenda

Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University,
Brno, Czech Republic
komenda@iba.muni.cz

Abstract. The paper introduces processing, modelling, analysis and visualisation of data on cancer epidemiology and cancer care in compliance with a proven and validated methodology. We aim to provide online access to unique data on cancer care and cancer epidemiology, including an interactive visualisation of various analytical reports in order to provide relevant information to the general public as well as to experts, such as health care managers, environmental experts and risk assessors. The data viewer has been developed and implemented as a web-based application, making a very time-consuming process of data analysis fully automatic. The presented data contain dozens of validated epidemiological trends in the form of tables, graphs and maps.

Keywords: cancer care, epidemiology, data analysis, data visualisation, CRISP-DM, Czech Republic.

1 Introduction

Health care reporting and overviews nowadays involve not only the distribution and availability of health care, but also the standardisation of diagnostic and treatment approaches [1]. The European health systems underwent a great deal of reorganisation in the last decade. There has been a tendency to facilitate the expanding involvement of the private and public health care sector, a process which has occurred mainly in the countries of Central and Eastern Europe [2]. Cancer care is a prime example of multidisciplinary medical service which requires integration and a certain degree of centralisation in order to ensure an optimal use of resources available and to achieve optimal treatment outcomes. Cancer surveillance holds a privileged position, compared to other diseases, in terms of sources for collecting data, rich experience and availability of data [1]. Various cancer monitoring systems aim to collect data on cancer occurrence and provide much more detailed information on cancer, including diagnostic criteria and therapeutic procedures at the level of individual patients. The Comprehensive Cancer Control (CanCon) initiative aims to improve the quality of cancer care in the European Union. Involved cancer experts from across Europe have joined forces to advance cancer care and reduce cancer incidence by: (i) identifying key elements and quality standards for comprehensive cancer control in Europe and preparing an evidence-based European guide on quality improvement in comprehensive cancer control; (ii) facilitating

Dušek et al, p. 1, 2016.

cooperation and exchange of best practice between EU countries, to identify and define key elements to ensure optimal, comprehensive cancer care [3]. CanCon has been divided into nine work packages, of which three are horizontal (dealing with coordination, dissemination and evaluation), and six are core packages (focusing on developing the content of cancer control). The concept of Comprehensive Cancer Care Network (CCCN) has been introduced in the work package 6, providing synergy with all institutions that have complementary expertise. CCCN aims to: (i) promote the optimal use of advanced technologies; (ii) make innovative clinical trials accessible to the entire population in a certain area; (iii) identify the most suitable unit within the CCCN for the management of rare and complex cancers; (iv) promote common infrastructures within the CCCN; and (v) provide a forum for regular consultation among professionals. A pilot model of such CCCN has been set up in the Czech Republic, namely in the Vysočina Region and South Moravian Region. This model covers all components of cancer care: from cancer prevention and organised screening programmes through standard diagnostic and treatment procedures to follow-up plans; specialised care focused on rare cancers as well as palliative care are also included. This CCCN involves one highly specialised national comprehensive cancer centre, three regional cancer centres and four general hospitals. This consortium of core centres was at the very start of the pilot CCCN and took responsibility for the development of binding cancer care protocols, rules of multidisciplinary teams and quality assessment standards [4]. Beside the required organisation of cancer services in two specific regions (in order to provide the best possible care for their population), collecting data on cancer epidemiology, their processing, analysis, and visualisation is very important in order to summarise long-term trends in cancer burden and to provide up-to-date incidence and mortality overviews.

1.1 Problem definition

The paper introduces the domain of epidemiological, clinical and demographic data aggregation, analysis and interactive visualisation. We started by research question formulation, which helped us clarify what exactly we wanted to achieve. We have defined the following issue focusing on a comprehensive and fully representative overview of analytical reports. Specifically, we aimed to investigate how to perform efficient modelling, storage and visualisation of data on cancer epidemiology from an area of Czech hospital network (the CCCN pilot model).

2 Methodological background

The healthcare industry has continuously generated large amounts of data stored in various locations (e.g. national information systems and specialised registries) [5]. In general, processing, modelling and analysis of these data need to be in compliance with proven and validated methodologies. It allows the discovery of new knowledge and potential useful information based on data describing the particular domain of human interest [6]. In fact, analytical reports, overviews and visualisations should help plan,

understand, work through and reduce cost by detailing procedures to be performed in each of the steps. For our purposes, we decided to use the CRoss-Industry Standard Process for Data Mining (CRISP-DM) reference model, which provides a life cycle overview of a given research question [7]. The CRISP-DM model serves mainly as the methodological standardised guideline in practice (Fig. 1).

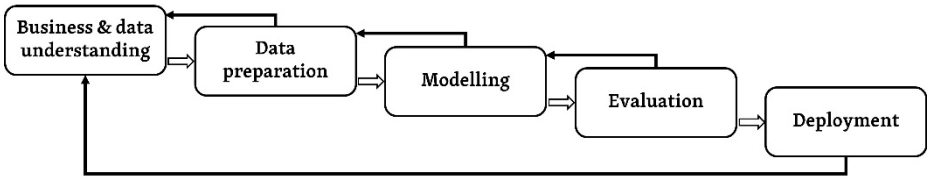


Fig. 1. CRISP-DM schema showing the relations between the different phases of the process.

(i) Business and data understanding introduces the defined objectives and requirements mapping from the research question perspective. It starts with initial data collection, identification of data quality problems, detection of interesting subsets regarding hidden and useful information. (ii) Data preparation covers construction of the final dataset from the initial raw data from various sources including table, record, and attribute selection, as well as data transformation and cleaning as pre-processed output files for modelling tools. (iii) Modelling represents a calibration of parametric values and the use of selected statistical and analytical techniques. (iv) Evaluation focuses on generating analytical reports (interactive data tables and graphs) assessing in terms of the usefulness, transparency and reliability. (v) Deployment phase consists of final reports implementation as well as testing and maintenance planning [8].

2.1 Business and data understanding

The Czech Republic is ranked among those countries with the highest cancer burden in Europe and worldwide [9]. Based on a thorough knowledge of the domain of cancer epidemiology, we needed to define the problem in the form of a research question corresponding to local regional estimates of incidence and mortality rates, which helps to choose the proper statistical and analytical methods and software tools selection. The main objective of our study entitled "CCCN pilot model: Interactive data views" was to design, to develop and to implement a web-based tool supporting scientific analytical reports on cancer data aggregated with demographic data from the Vysočina Region and the South Moravian Region. This stage also involved a more detailed fact-finding about all data sources. The Czech National Cancer Registry[1] provides fully representative long-term trends and consists of cases according to main risk factors and diagnostic descriptors including TNM classification of tumours in the following diagnostic groups [10]. These groups are presented in accordance with the 10th edition of the International Classification of Diseases (ICD-10) terminology [11], which is the standard

[1] http://www.uzis.cz/en/registers/national-health-registers/czech-national-cancer-registry

diagnostic tool for epidemiology, health management and clinical purposes. We used only Chapter 2 (C00-D48 diagnoses), which classifies all neoplasms:

- I. Head and neck cancers;
- II. Digestive cancers;
- III. Cancers of the respiratory tract and intrathoracic organs;
- IV. Bone and soft tissue cancers;
- V. Skin cancers;
- VI. Breast cancers;
- VII. Gynaecological cancers;
- VIII. Genitourinary cancers;
- IX. Cancers of the central nervous system and eye;
- X. Malignant neoplasms of lymphoid, haematopoietic and related tissue;
- XI. Endocrine cancers;
- XII. Other malignant neoplasms;

Data describing the main demographic characteristics of the Czech population (such as the total population, age structure or life expectancy) were provided by the Czech Statistical Office[2].

2.2 Data preparation and modelling

We decided to use the data warehousing concept, which makes it possible to integrate information from heterogeneous databases and to query very large databases efficiently [12]. We aimed to synthesise available data on cancer care and to store them together in a single repository. First of all, a four-step process – extraction, transformation and load (ETL) – was performed.

Step 1: Stage. The first step includes data retrieval from the Czech National Cancer Registry and the Czech Statistical Office. Raw data are mined from both databases and further processing (data cleaning and transformation) is needed due to data format unification. Moreover, we discovered key dataset features and characteristics, which also include the tables, records (rows), and attributes (columns) selection. In case of a wrong format or syntax of the uploaded CSV file (syntactical and semantical check is performed), the transaction is aborted and the entire step has to be repeated. The output of the step is one table called <<import_nador>>.

Step 2: Operational data store. The second step is represented by the database table <<fact_primar>>. Data are syntactically and semantically checked by function f_restart_fact_primar() right before being stored to this table. If the check is not successful, the table <<import_nador>> has to be fixed and an attempt to transform it into a table <<fact_primar>> has to be made again.

[2] https://www.czso.cz/csu/czso/home

Step 3: Primary Data Warehouse (PWD). The third step consists of fact tables and dimension tables. The fact tables are not directly connected to each other, but there are relational links between them, depend-ing on their dimensions contents. Dimension tables are simple tables which contain an abstract identifier as the primary key and columns with descriptive data; these are displayed on the screens of the data viewer. Based on the selection of one or more descriptive attributes, clustering of identifiers in the structure suitable for data selec-tion from the dimension tables is performed. It is implemented through the re-start_dimensions_function() function, which firstly erases data from allthe dimension tables and secondly inserts data to each dimension table by accumulation (imple-mented through GROUP BY clause) of <<fact_primar>>. In cases when the <<fact_primar>> table does not contain any descriptive attributes, data for the di-mension table are inserted to it through SQL commands or manually from external sources. In contrast to the table <<fact_primar>>, fact tables do not contain atomic data providing the base for all data transformations, but contain aggregated data created by the count(*) function over all the appropriate dimensions. The data are cumulated in them. This means that the index over all dimensions in each fact table is unique and the measuring columns contain the numbers of specific cases that oc-cur. Data selection is performed faster in this way. The only exception is the <<fact_demography>> table, which already contains aggregated data, and is import-ed directly from CSV file.

Step 4: Data Mart. This phase involves tables derived from the PDW data. These are so far represented only by the <<fact_agr_clinic_stage>> table. Data in the tables are derived from the fact tables <<fact_demography>> and <<fact_agr_patient_diagnose>>. Again, the data in the fact table are arranged similarly to PWD based on aggregation over existing dimensions. The whole ETL process of selected entities is shown below (see Figure 2). The external source (CSV file) is imported to the database entity <<import_nador>>, validated and transformed to the table <<fact_primar>>. At this point, several fact tables of PWD (e.g. table <<fact_agr_patient_diagnose>>) are created. The table <<fact_demography>> is created directly from the external source. The Data Mart includes one table entitled <<fact_agr_clinic_stage>>. The data for endpoint visualisations can be selected from all fact tables, but there are significant performance differences (e.g. between the selection of the same information from the <<fact_primar>> and the <<fact_agr_clinic_stage>> tables).

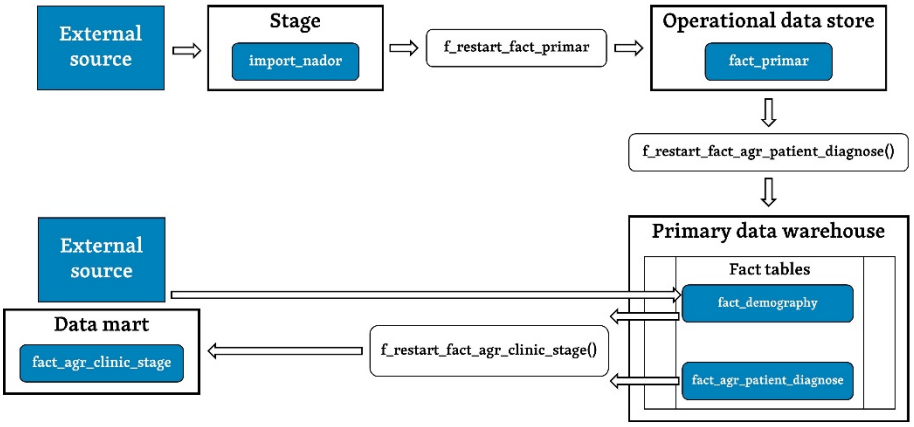


Fig. 2. Complete data flow in CCCN data warehouse subset.

The fact table <<fact_agr_clinic_stage>> (Fig. 3) is part of the CCCN Data Mart. This table is derived from fact tables included in the PDW and has the following dimensions, which correspond to the visualisation filters: dregion, dyear, dsex, dage_group, dage_group2 and ddiagnose_stage. The last dimension is a set of diagnoses and groups of diagnoses in accordance with ICD-10.

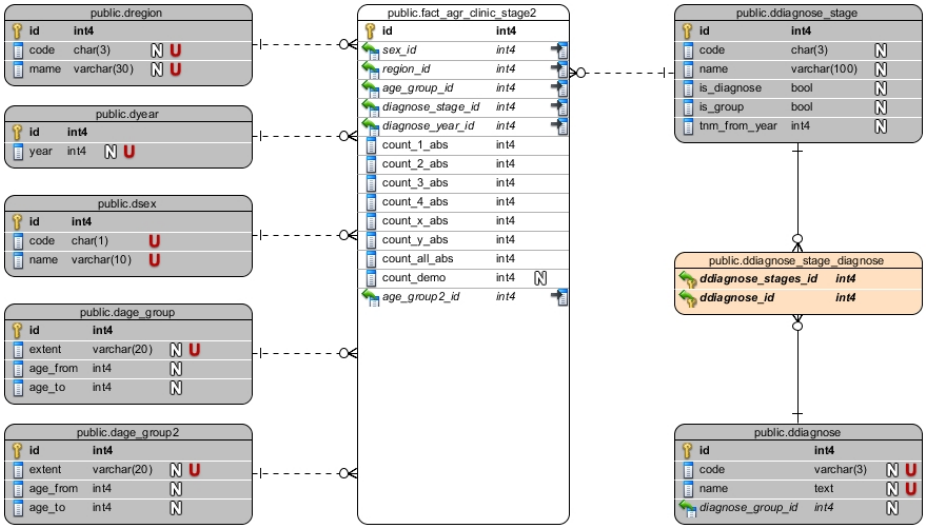


Fig. 3. Fact table of clinical stages where all relations between <<fact_agr_clinic_stage>> entity (white), parametric lists (grey) are shown.

2.3 Evaluation and deployment

The evaluation phase assesses the degree to which the analytical reports meet the given objectives and seeks to determine if there are any imperfections or inaccuracies in terms of graphs, table validity and general user understanding. We made a detailed computational validation of the outputs presented by tables and graphs using standard epidemiological statistical methods [13]. This statistical validation was computed using the SPSS 24.0.0.0 software. The CCCN pilot interactive data viewer is a web-based application written in the Symfony framework[3] version 3.1. The data warehouse is implemented as a PostgreSQL[4] 9.5 database, which is hosted on the database server. The application core is divided into two parts:

- AppBundle (Symfony project) – the main application bundle which provides data access layer, PHP entity mapping and all processes connected to the Model-View-Controller (MVC) architecture.
- VisualisationBundle – this bundle contains a functionality connected strictly with graphs, data tables and filters shown on all screens.

The Git[5] system was used for an efficient work with the code. All instances can be independently tested by a set of acceptance tests written in the integrated development environment for Selenium IDE[6] Mozilla Firefox plugin, which allows tests recording, editing, and debugging.

3 Results

The developed data viewer[7] allows the user to investigate general epidemiological trends for a particular diagnostic group. The viewer is divided into three individual modules, each of them containing a set of specific analytical reports visualised by interactive graphs and maps (Table 1).

[3] http://symfony.com/what-is-symfony
[4] https://www.postgresql.org/about
[5] https://git-scm.com
[6] http://www.seleniumhq.org/projects/ide
[7] http://cccn-viz.onconet.cz

Table 1.

Viewer module	Analytical report
Cancer epidemiology	Incidence and mortality trends over time Prevalence trend over time Age structure of incidence and mortality Age-specific incidence and mortality rate
Regional benchmarking	Incidence rates in regions Mortality rates in regions
Cancer diagnostics	Distribution of clinical stages Incidence trends by clinical stages Distribution of clinical stages by age Comparison of incidence of clinical stages by age

The user can access the analyses of each module through a module navigation homepage (Fig. 4).

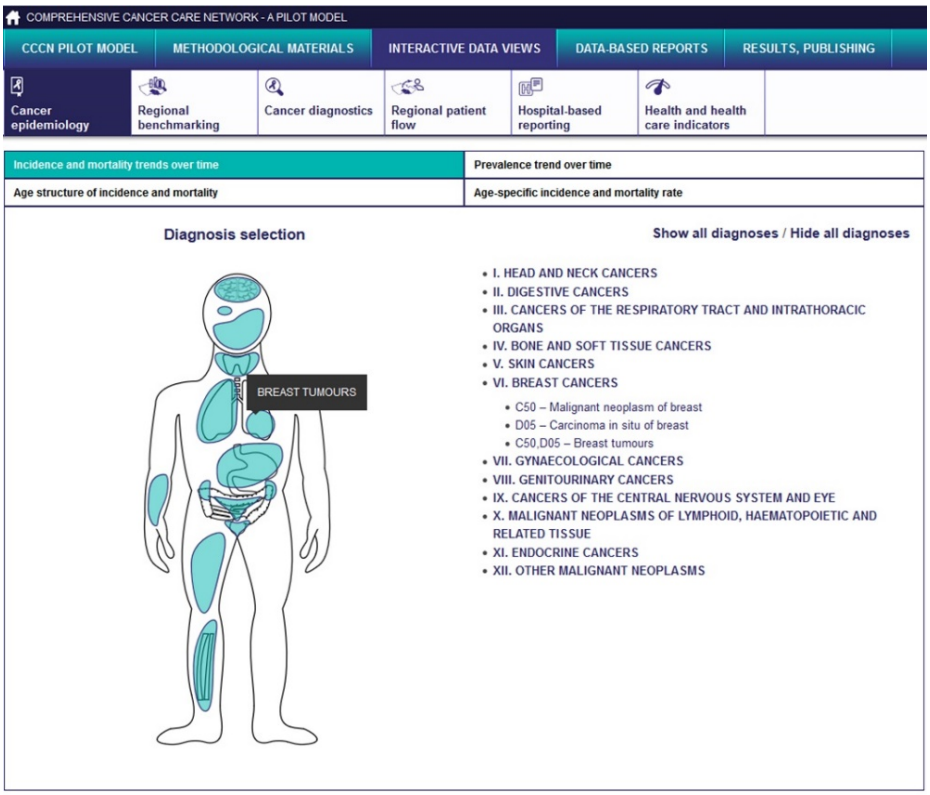


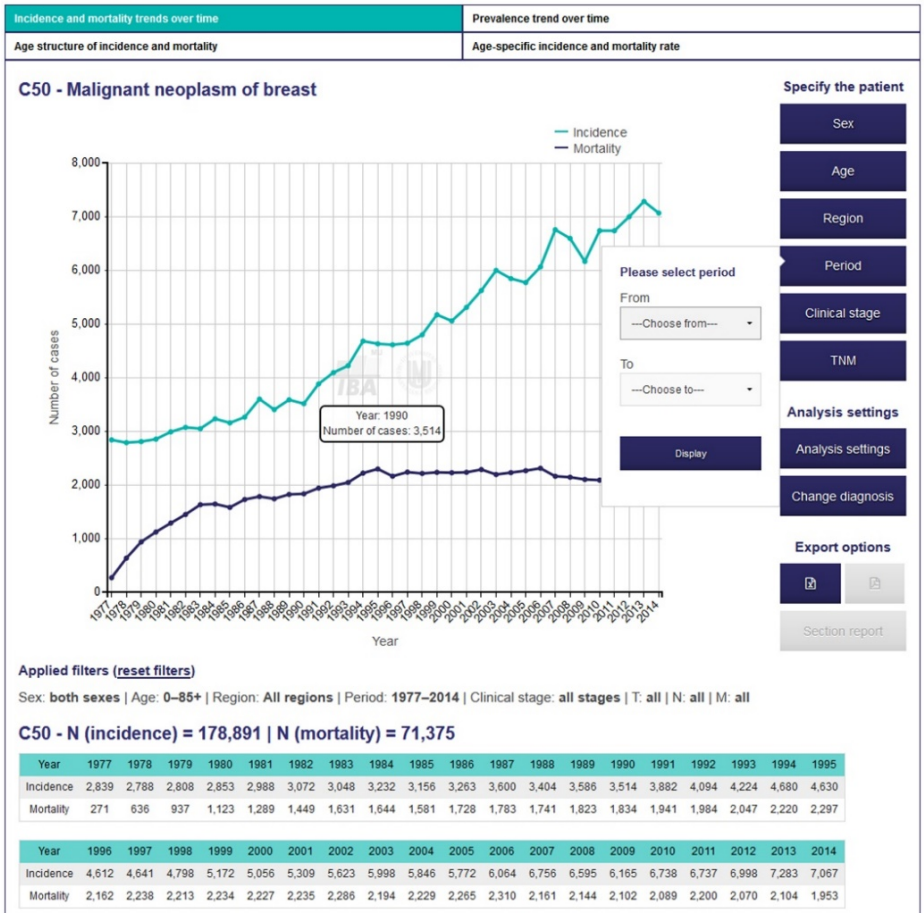
Fig. 4. Diagnosis selection page provides a human body interactive silhouette including tooltips (on the right) in combination with a complete list of all available diagnoses (on the left).

The selection of individual analyses is at the top of the screen. The user has an overview of all provided data views in the current module, and can choose the one that interests him/her. The remaining part of the navigation page contains a selection of a particular diagnosis. The diagnoses are divided into twelve major cancer groups, which were defined in the domain and data understanding section. The groups can be freely expanded and browsed directly in the tree list (on the right) or by using the schematic picture of a human silhouette with tooltip elements and information about the part of human body being hovered over (on the left). After selecting a specific diagnosis – or a subgroup of diagnoses – the screen with the required analytical report appears again. For illustrative purposes, four epidemiological views and analytical reports of the "Cancer epidemiology" viewer module are described as following.

- Incidence and mortality trends over time show annual incidence (newly diagnosed cases of a selected diagnosis) since 1977, and annual mortality (deaths caused by a selected diagnosis) since 1977 – absolute numbers or rate per 100,000 population.
- Prevalence trend over time shows annual prevalence (alive persons with disease or its history) since 1990 (point prevalence at 31 December of each year or interval prevalence during years) – absolute numbers or rate per 100,000 population.
- Age structure of incidence and mortality, and age-specific incidence and mortality rates show the numbers of new cases (incidence) or deaths caused by diagnosis (mortality) according to five-years age groups (absolute numbers, percentages or rate per 100,000 population).

When a user accesses the above-mentioned analyses in a standard way (there is also the option of proceeding through an URL with the already pre-filtered content), the graph and the table displays the complete information about all records in the registry. Additionally, we have provided a set of filters that correspond to fact table dimensions in the data warehouse. If the filter is disabled, the user is informed by the button's grey colour and the value is not included in the list of applied filters. The complete list of filters with analysis settings is on the right side of the screen (Fig. 5) and contains the following types of filters:

- Sex filter – men, women, both sexes.
- Age filter – five-year groups from 0 to 85+.
- Region filter – selection from 14 regions of Czech Republic plus the CCCN pilot model.
- Period filter – years from 1977 to 2014.
- TNM filter – extension of the primary tumour (T), regional lymph nodes (N), and distant metastases (M) [14].
- Clinical stage filter – combination of the TNM staging system, which determines the stage of cancer for each person (four stages: stages I to stage IV).



Year	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Incidence	2,839	2,788	2,808	2,853	2,988	3,072	3,048	3,232	3,156	3,263	3,600	3,404	3,586	3,514	3,882	4,094	4,224	4,680	4,630
Mortality	271	636	937	1,123	1,289	1,449	1,631	1,644	1,581	1,728	1,783	1,741	1,823	1,834	1,941	1,984	2,047	2,220	2,297

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Incidence	4,612	4,641	4,798	5,172	5,056	5,309	5,623	5,998	5,846	5,772	6,064	6,756	6,595	6,165	6,738	6,737	6,998	7,283	7,067
Mortality	2,162	2,238	2,213	2,234	2,227	2,235	2,286	2,194	2,229	2,265	2,310	2,161	2,144	2,102	2,089	2,200	2,070	2,104	1,953

Fig. 5. Visualisation of incidence and morbidity trends over time.

4 Conclusion

We have designed, developed and implemented a prototype of interactive viewer of data on cancer care and cancer epidemiology, which shows the contemporary analytical overview of cancer incidence, prevalence and mortality, including the evidence on cancer care and cancer epidemiology in the CCCN pilot region and on the national level. In future, we might be able to extend the viewer in order to ensure regional and also international benchmarking analyses. Generally, the visual representation of cancer screening data is very heterogeneous (free text, parametric text, numerical and graphical format). The trend aims to show available data sources in a form of interactive visualisations, specifically graphs, maps and diagrams. Modern users want to easily understand aggregated information in a comprehensive and validated shape for further decision-making activities. The presented data visualisation concept heads towards

data-driven approach, which is quite useful for health professionals in situations where a large amount of data must be presented in the most comprehensible way.

5 References

1. Azevedo, A.I.R.L., 2008. KDD, SEMMA and CRISP-DM: a parallel overview.
2. CanCon: oficial webpage [WWW Document]. URL http://www.cancercontrol.eu/ (accessed 2.8.17).
3. Comprehensive Cancer Care Network: A pilot model [WWW Document]. URL http://cccn.onconet.cz/ (accessed 2.8.17).
4. Dušek, L., 2009. Czech cancer care in numbers 2008-2009. Grada.
5. Dušek, L., Mužík, J., Gelnarová, E., Fínek, J., Vyzula, R., Abrahámová, J., 2010. Cancer incidence and mortality in the Czech Republic. Klin Onkol 23, 311–324.
6. Dušek, L., Mužík, J., Koptíková, J., Brabec, P., Žaloudík, J., Vyzula, R., Kubásek, M., 2005. The national web portal for cancer epidemiology in the Czech Republic, in: Proceedings of the 19th International Conference Informatics for Environmental Protection (Enviroinfo 2005). Masaryk University, Brno.
7. Edge, S.B., Compton, C.C., 2010. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann. Surg. Oncol. 17, 1471–1474.
8. Golfarelli, M., Maio, D., Rizzi, S., 1998. The dimensional fact model: A conceptual model for data warehouses. Int. J. Coop. Inf. Syst. 7, 215–247.
9. Micheli, A., Coebergh, J.W., Mugno, E., Massimiliani, E., Sant, M., Oberaigner, W., Holub, J., Storm, H.H., Forman, D., Quinn, M., Aareleid, T., Sankila, R., Hakulinen, T., Faivre, J., Ziegler, H., Tryggvadòttir, L., Zanetti, R., Dalmas, M., Visser, O., Langmark, F., Bielska-Lasota, M., Wronkowski, Z., Pinheiro, P.S., Brewster, D.H., Plesko, I., Pompe-Kirn, V., Martinez-Garcia, C., Barlow, L., Möller, T., Lutz, J.M., André, M., Steward, J.A., 2003. European health systems and cancer care. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO 14 Suppl 5, v41–v60.
10. Organization, W.H., 2004. International statistical classification of diseases and related health problems. World Health Organization.
11. Raghupathi, W., Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. 2, 3.
12. Romero, C., Ventura, S., De Bra, P., 2004. Knowledge discovery with genetic programming for providing feedback to courseware authors. User Model. User-Adapt. Interact. 14, 425–464.
13. IARC Publications Website - Statistical Methods in Cancer Research Volume IV: De-scriptive Epidemiology. [Online]. Available: http://publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications/Statistical-Methods-In-Cancer-Research-Volume-Iv-Descriptive-Epidemiology-1994. [Accessed: 04-Apr-2017].
14. Sastry, S.H., Babu, P., Prasada, M.S., 2013. Implementation of CRISP Methodology for ERP Systems. ArXiv Prepr. ArXiv13122065.