OPENING HEALTH DATA

Martin Komenda, Boris Turek, Michal Vičar, Andrea Pokorná, Ladislav Dušek

METHODOLOGICAL BACKGROUND IN HEALTH DATA SPACE

Data processing and subsequent visualisation are integral parts of this publication. The field of health data is fundamentally complicated concerning the sensitivity of published information and personal data protection. It is essential to conceptually address and highlight what data will be published (the content), how (the appropriate format and interface), and to whom (target groups). A key role is played by the electronic health record (EHR), which has not only made patients' medical information easier to read and available from almost any location in the world, but also changed the format of health records, and thus changed health care [1]. Individual datasets must be adopted as an official and guaranteed source for outputs of third parties, including public authorities, non-governmental organisations, scientists, and online news portals [2]. Thus, most chapters of this book are working with the concept of open data. From a general perspective of working with data, it is elementary to realise that there are several modes or approaches to working with health data. In this specific domain of healthcare, there are logically very specific cases that need to be systematically dealt with in accordance with current legislation and data protection. It is also necessary to mention the importance and significance of the data's origin (sourcing or collection) in an ethical context. In academic settings, there are often various guidelines or internal regulations in faculties that describe the moral code of a scientist or researcher. The main objective is to define clear and transparent rules for handling, publishing and archiving data for the retrospective validation of the results obtained. This chapter introduces this issue with the aim of introducing the concept of open data and the fact that historically there has been, and often will continue to be, a completely individual consideration of each dataset. This is from all relevant perspectives, such as methodological, technical, analytical and legal.

Sharing datasets, preferably in open data format, provides a systematic way to make selected datasets available for further manual or machine processing in a uniform and technically well-defined form (= dataset ready for "re-use"). The development of the open data domain, together with the operation of the

National Catalogue of Open Data[1], was for many years coordinated by the Ministry of the Interior of the Czech Republic; today, this role is secured by the newly established Digital and Information Agency (DIA). In the health information domain, a new working group under the leadership of the Ministry of Health of the Czech Republic is being set up to systematically coordinate activities related to the opening of health data (approval, creation, validation, and publication of datasets from the National Health Information System and other departmental information systems). The target group involves all stakeholders who aim to work with health data on a one-off or continuous basis (business, scientific and research infrastructures, academia, media and news, working groups of public authorities, regional governments, professional and lay public and others). Securing all the necessary input in the form of management and executive-level staff, a robust information technology infrastructure and methodological leadership are essential aspects of the systematic development of health data opening.

All regimes of the data provision from National Health Information System (NHIS) strictly require a certain degree of legislative regulation and must fully meet the criteria set for NHIS by the Czech legislation, particularly Act No. 372/2011 Coll., on Health Services and Conditions of Their Provision. In other words, publishing open data cannot be misinterpreted as the publication of primary records without any regulation and standardisation; the term "open data" does not necessarily describe prior database records (the data may be aggregated, statistically processed, etc.).

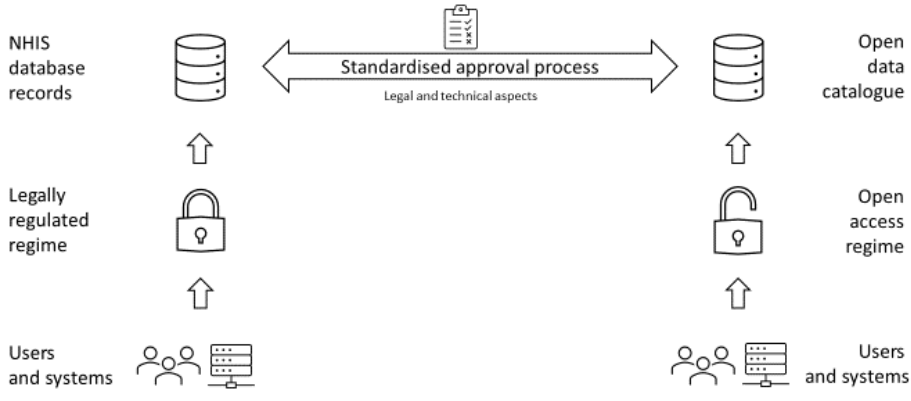

Figure 6: National Health Information System: Data provision schema

1 https://data.gov.cz/datasets

The dataset design, preparation, and publishing should respect the algorithm of dataset preparation shown below, which always respects several principal rules:

1. Both direct and indirect identification of individual person cannot be possible.

2. Explicit identification of providers and other healthcare-related subjects must not be identifiable unless expressly stated by the law.

3. Secondary processing must lead to the pseudonymisation of the dataset.

4. The purpose of the dataset publication must correspond to the NHIS purpose.

5. The standardised approval process and publishing must be adhered to (Figure 7).

6. Each dataset has clearly defined authors and associated licensing is strongly recommended (if applicable).

The essential requirement of a comprehensive process, which this approach meets, is to ensure the necessary completeness, validity, and overall quality of data [2].

Step 1 Concept design	Step 2 Concept evaluation	Step 3 Feasibility analysis	Step 4 Dataset production	Step 5 Review	Step 6 Publication
Proposed by state administration, external subjects (health insurance companies, expert societies, research institutions)	Purpose, data availability, feasibility, legal perspective	Data extraction, processing, analysis, validation	Structure, methods of production, metadata description	Personal data protection, factual content, IT solution	National Catalog of Open Data

Figure 7: A chart of dataset production and publication

A universal and comprehensive methodological description of designing, preparing, validating, and publishing datasets is essential information with which the wider open data community in healthcare must be familiar. As the case of NHIS open data sets, the final output is always a work of authorship dataset (a value of the main idea introducing dataset structure, meaning and purpose), including complete disclaimers and a clear definition of licensing rules (if applicable) for appropriate citation in professional publications and other results.

An integral part of this is the categorisation of the individual datasets of the National Health Information System (NHIS). This categorisation, together with

the explicit statutory purpose and level of aggregation, determines the possible mode of "openness" of particular subsystems of the NHIS as follows:

— systems not public by law,

— systems accessible only to legally defined readers/editors,

— sources of reference statistical data for the identified purpose and for the identified applicants,

— sources of published statistical data in the form of open datasets,

— sources open in primary "open data" mode.

When preparing datasets, it is recommended to follow the dataset creation scheme (Figure 7), where it is always crucial to respect the following rules: to comprehensively grasp the complex issue of data opening in Czech healthcare system, following three categories for data handling are proposed, defining different approaches according to the content of the information to be published.

1. Freely available primary data (very rarely used); examples may include service providers as defined in the respective acts, machines, swimming pools, chemical substances or drugs, and their primary characteristics.

2. Primary data publishable after necessary processing (used most frequently).

3. Data requiring reference interpretation – reference statistics, which means "a comment or summary to data output given by an expert in the field, usually given to data and values that have a more sophisticated background and require careful interpretation concerning objective uncertainties" [2].

The individual case studies mentioned in this book mainly refer to either open datasets or datasets that meet all technical aspects of open data except publication in a local or national catalogue (those datasets are not open but somehow FAIR). The book is intended to serve, among other things, as study material. However, this does not automatically mean that many samples or pilot data must be automatically published in an open data catalogue. However, the openness and availability of the data without access restrictions remain, and any user can freely work/re-use (with) the data referenced in the book.

TARGET GROUPS AND COMMUNICATION

An important part of the processing of open data in healthcare is connected with a proper definition of human resources and stakeholders dealing with healthcare open data. A universal and comprehensive methodological description of the process of design, preparation, validation and publication of data is essential information with which the broad community (composed of several general target groups) dealing with open data in healthcare must be familiar to be able to reuse the datasets. Ensuring clear and consistent communication about open data in healthcare is essential to a functional system. User profiling should consider not only communication goals but also information, health and data literacy. The outcome of such thinking is a division into three basic levels, which may overlap in preference of output format (Figure 8). Clarity, level of detail and, above all, a guarantee of validity and correct interpretation are essential attributes in terms of minimum requirements and demands. Such a concept requires very close collaboration between experts across the necessary expertise (theoretical physician or clinician, member of an expert medical society, healthcare expert, computer scientist, analytical guarantor, guarantor of data visualisation, open data guarantor). Depending on the format of the output, these people are actively involved in the actual process of communicating with the selected target group. In particular, data explorers and data experts have an excellent opportunity to disseminate the results correctly among stakeholders who have the mandate to reach the general lay public, i.e. data novices.

Figure 8: Communication of health data

Similarly important is the communication aspects towards the identified target groups. The lay public, together with the informed and professional public,

must be able to be informed about the current status, the publication plan and the overall vision for open data within the specific healthcare domain, if needed. Communication tools that will be continuously used not only to communicate the current status but also to gather suggestions, requests and feedback, include:

— National and regional conferences and webinars

— Discussion panels and educational seminars

— Individual consultations

— National Health Information Portal[2] as a source of guaranteed and verified information for the general public

— Online platform data.nzis.cz[3] for up-to-date information and news

— National Catalogue of Open Data[4]

— Social media (Facebook, Twitter, Instagram)

— Press conferences and official opinions of the Ministry of Health and the Institute of Health Information and Statistics of the Czech Republic or other public sector authorities

— Specialist publications and dedicated web portals

Systematic communication across the professional community at the international level is also an integral part of this effort; the aim of this communication is to share and transfer experiences, especially directly in the specific domain of open data in healthcare. Table 2 below describes the idea defining the three focus/target groups of (open) data professionals, their level of information processing, data knowledge and skills, and the main objectives underpinned by key motivating factors for each group.

2 https://nzip.cz
3 https://data.nzis.cz
4 https://data.gov.cz/datov%C3%A9-sady

Table 2: Description of target groups in terms of health data communication

Domain	Data novice	Data explorer	Data expert
Information and data literacy	The individual has basic knowledge of using information technologies. Can recognise, collect and share information in a digital environment. Understands the basic principles of assessing the credibility of sources. He/she occasionally searches for information in areas of his own interest.	The individual has more advanced knowledge in the use of information technologies. He/she is able to work with data, analyse it and draw conclusions from it. Understands the basic principles of data processing, such as collecting, organising, analysing and visualising data. Has a basic understanding of data analysis tools and understands the importance of critical thinking when interpreting results.	The individual has more advanced or expert knowledge in the use of information technologies. He/she is an expert in the field of information and data literacy. He/she is able to evaluate sources, process and analyse data, recognise sophisticated forms of misleading and disinformation. He/she has a deep understanding of data rights, ethical aspects of data processing and is able to comprehensively and critically evaluate information in the digital world.
Goals and motivation	The individual searches for information and data from the health sector on an ad hoc basis. He/she is more interested in the results than the journey. Articles and data overviews are accessed primarily via a search engine or via social networks. Charts and articles are shared via URL, in image form, video, or as PDF exports. To fulfil the need for information, he/she looks for mostly unstructured data. He/she approaches information more emotionally.	The individual often searches for information and data from the health sector for his/her work. Articles and data reports are accessed directly. Knows the basic sources of data reports and recognises their relevance. He/she approaches information critically and can draw conclusions from it and interpret it. He/she looks for mostly secondary sources of data. He/she wants to have all reports at hand as quickly as possible. Searches for structured and unstructured data to fulfill its information needs. He/she is intrinsically and explicitly motivated to education in data issues. He/she approaches information critically and ethically.	He/she searches for information and data from the health sector every day for his work. The data are accessed directly. Knows most sources of data reports and recognises their relevance. He /she approaches information critically and is able to evaluate, process, analyse, identify trends and subsequently interpret them. He/she looks for mostly primary data sources. He/she wants all reports to be complete and easily navigated (filtering). To fulfil the need for information, he/she mostly searches for structured data. There is a need for as much input data as possible to work with data. He/she is intrinsically and explicitly motivated to educate himself/herself in data issues to actively contribute to his/her know-how (professional publications, annual reports). He/she approaches information critically, ethically and legally.
Device usage	20 % laptop 80 % mobile	50 % laptop 50 % mobile	80 % laptop 20 % mobile
Characteristics	Fragile Low dissemination rate Risk of unintentional misinterpretation	Influenceable High dissemination rate Risk of intentional misinterpretation	Resistant and stable Medium dissemination rate Low risk of misinterpretation High level of distrust

The key attributes listed in Table 2 were essential characteristics in designing the open data communication matrix in healthcare. They helped define three main target groups based on information literacy, general goals and overall motivation for delivery and distributing the information, typical device usage habits, and ability to form conclusions and further dissemination.

REFERENCES

[1] Evans RS. Electronic Health Records: Then, Now, and in the Future. Yearb Med Inform. 2016;Suppl 1(Suppl 1):S48–S61.

[2] Komenda M, Jarkovský J, Klimeš D, et al. Sharing datasets of the COVID-19 epidemic in the Czech Republic. PLoS One. 2022;17(4):e0267397.