Addressing Quality Issues in Secondary Use of Health Data

Kalinka Kaloyanova [1, 2], Ina Naydenova [2] and Zlatinka Kovacheva [2, 3]

[1] *Faculty of Mathematics and Informatics – Sofia University St. Kliment Ohridski, 5 James Bourchier Blvd., Sofia 1164, Bulgaria Sofia University, St. Kliment Ohridski, 15 Tsar Osvoboditel Blvd., Sofia, 1504, Bulgaria*
[2] *Institute of Mathematics and Informatics – Bulgarian Academy of Science, 8 Acad. Georgi Bonchev Str., Sofia, 1113, Bulgaria*
[3] *University of Mining and Geology "St. Ivan Rilski", Sofia, 1700, Bulgaria*

Abstract

During the last two decades, medical data digitalization has grown constantly. This process raises a lot of challenges regarding data privacy, data interoperability, and data quality. Despite the variety of systems that manage and analyze medical data, in many cases, data is not properly collected and used. A significant part of these problems can be identified and overcome when the collected data is reused. Recent European initiatives to establish a common space for health data also create opportunities for more efficient secondary use of data. The paper discusses basic quality issues in the secondary use of data and how they could be addressed.

Keywords

Data quality, quality attributes, health data, secondary use of data, European Health Data Space (EHDS)

1. Introduction

Many innovations during the last decades influence the health sector. In addition to new drugs and methods of treatment, new devices and software applications were used and large amounts of medical data were generated. Unfortunately, there are many cases where this data is not properly collected and documented. Most frequently mentioned flaws concern health data interoperability, missing data, and low data quality. The secondary use of already obtained data can be applied not only as a mechanism for gaining more value from data but also as a mechanism that reveals and solves many problems with data quality.

Information Systems & Grid Technologies: Fifteenth International Conference ISGT'2022, May 27–28, 2022, Sofia, Bulgaria
EMAIL: kkaloyanova@fmi.uni-sofia.bg (K. Kaloyanova); naydenova@gmail.com (I. Naydenova); zkovacheva@hotmail.com (Z. Kovacheva)
ORCID: 0000-0003-0222-7607 (K. Kaloyanova); 0000-0002-9995-8299 (I. Naydenova); 0000-0001-7401-3072 (Z. Kovacheva)

The secondary use of medical and health data can be explored in different directions. Data can be used for improving health care for patients, as well as for optimizing health systems services at different levels – "personal care planning, medicines development, safety monitoring, research, and policymaking" [11]. These optimizations cannot be achieved if the information collected does not meet certain quality criteria.

The secondary use of data is different from the primary use of data in many aspects. In the case of health data, the primary data use is connected mainly with individual care for patients. For example, clinical data is accumulated from diagnoses, treatment recommendations, prescribed medicine, etc. Personal data of patients, as well as health insurance data, is also included. However, the health data could include much more details – for example, data coming from different medical devices or smartphone applications. In addition, the secondary use of health (medical) data is connected with the use of aggregated data, coming from different sources "…such as electronic health records, health insurance claims and health insurance data" [4]. This data can be reprocessed for new purposes – different types of research on the data, seeking cost-effectiveness for products and services, resolving problems, etc.

Most of the research discusses data quality in the case of the primary use of health data. The reuse of data, on the other side, may set new requirements for the data to change the criteria for their quality.

In this paper we outline basic quality issues, concerning health data secondary use and propose useful recommendations for data processing with a focus on data quality. We also briefly discuss some aspects related to the confidentiality of the medical data and the legal basis for their processing for purposes other than the original ones.

2. Secondary use of health data

Secondary use of health data is related to the use of medical data for purposes other than the reasons they were collected and stored initially. Medical data reuse has many advantages over primary data use:

- significant volumes of medical data are available as they are stored and processed in a variety of applications;
- data is structured, in many cases even summarized and generalized;
- data are collected in a certain period of time;
- there is no need for physical interventions or other ways of collecting data.

Apart from all considerations that traditionally are important when processing data, several other aspects, such as legal and ethical ones, are of big importance with regard to health-related information.

2.1. Privacy, legal and ethical considerations

All European countries and institutions are seriously considering data privacy issues. The EU "General Data Protection Regulation (GDPR)" presents the rules for the use of personal data that must be followed by all organizations. The GDPR aims to ensure secure methods for data processing. In addition, this regulation requires rules to be defined and implemented to achieve this goal. It introduces six main principles that need to be followed when personal data is processing: (1) lawfulness, fairness, and transparency; (2) purpose limitation; (3) data minimization; (4) accuracy; (5) storage limitation; and (6) integrity and confidentiality [6].

The General Data Protection Regulation is focused on the protection of individual data. However, medical data has the potential to be used for purposes that affect a large part of society, even in a form that does not contain personal information. It is therefore essential that the ethical aspects of the use of medical data be regulated, too.

As for data reuse, Recital 50 of GDPR indicates that the secondary use of personal data should be compatible with the reasons for the initial collection and use of data [6]. Furthermore, according to Article 9 health-related personal data is considered as "sensitive" and it is differentiated as a "special category" of data. The special categories require extra attention and need more protection because of their sensitivity. Ten conditions for processing special category data are presented in Article 9. For lawfully processing of special category data, both a lawful basis under Article 6 and a separate condition for processing under Article 9 should be identified [6]. The two justifications should not be linked. To avoid unacceptable distribution of sensitive information, two main techniques are used: anonymisation, where personal information is deleted (or permanently replaced by unrelated characters), and pseudonymisation, where sensitive data is encrypted in a way that allows it to be re-identified with the help of additional information.

Further, different countries could provide at the national level specific initiatives, procedures, and rules. In 2007, the American Medical Informatics Association provided a broad discussion on the issues, related to the secondary use of data [14]. The Finnish model for Secure use of data presents a detailed view of the ethical aspects of national health data policy [1].

2.2. EU regulations focused on health data space

European countries had made great efforts to create common principles for the processing of medical data [2], [5].

In May 2022, the European Commission (EC) published a proposal for a *REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space* [5]. This proposal aimed at the establishment

of a common framework for health data sharing. “The general objective of the intervention is to establish the rules governing the European Health Data Space to ensure natural persons’ access and control over their own health data, to improve the functioning of the single market for the development and use of innovative health products and services based on health data, and to ensure that researchers, innovators, policy-makers and regulators can make the most of the available health data for their work while preserving trust and security” [13].

The scope of health data is expanded to include health records, social data, administrative data, genetic and genomic data, public registries, clinical studies, research questionnaires, and biomedical data such as biobanks [7].

The document presents the new rights of patients regarding their personal electronic health data records such as the right to have free of charge access to a readable and accessible form of their personal electronic health data, “for example through the personal health data access service” [13]. It also explains when secondary use is allowed: “Data users are allowed to re-use health data only after receiving a data permit from a competent authority” [10].

It is expected the new regulation to encourage scientific research, as well as the development of advanced products and services in the health area. In addition, it will strengthen the cross-border exchange of health data between the different Member States.

3. Health data quality aspects

Data quality dimensions represent measurable data quality characteristics. The international standard ISO/IEC 25012 introduced a Data Quality Model with fifteen major data quality characteristics – accuracy, completeness, consistency, currentness, accessibility, creditability, compliance, efficiency, confidentiality, availability, recoverability, portability, as well as precision, traceability, and understandability [8]. The importance of quality characteristics of health data is broadly discussed in many publications and a lot of efforts are invested in achieving them, but the results are still not satisfactory [15], [9], [16].

The standard presents a common understanding of the importance of data characteristics, but the particular domain, where data is used, also has a strong influence on data characteristics and their prioritization [9]. In the table below, the quality dimensions, that are most relevant to health data, are listed and briefly described.

Table 1
Priority Health Data Quality Dimensions

Dimensions	Description
Accuracy	Degree of correct representation of the object
Completeness	Reflects the presence of values of all required attributes
Relevance	Presents how usable is data
Timeliness	The time expectation for accessibility and availability of information
Consistency	Data is presented in the same format
Security	Security access to data
Accessibility	Presents the degree of retrievability of the data

The prioritization of these quality characteristics can further differ depending on the type of records. A number of sources reported major challenges to the data quality of electronic health records [3], [9],[15]:

- Incompleteness – missing important details (attributes) of information;
- Inconsistency – incompatible, conflicting information between different data sources or even in the same EHR record;
- Inaccuracy – partially or completely incorrectly entered values.

For secondary use, data quality is no less important [12]. But in this case, quality dimensions can be viewed in a different way, compared to their use for primary purposes.

The incompleteness usually is reported as a leading data quality issue in the cases of the primary use of data but it could be overcome in some cases of reuse. When massive data sets are processed, missing or wrong values in some parts of them will not have a significant impact on the conclusions. Inconsistency also could be dismissed, if the detected cases are not too many and can be ignored. Nevertheless, the data of a particular individual is not of significant importance, the final results of the processing may have a significant influence because of the potential to touch much more people.

Data accuracy is a quality attribute that could be closely related to the context of use, so the new views on data may insist on new levels of accuracy. Data completeness is also quite sensitive to the specific objectives of the processing and should be assessed again.

In the primary use of data, where the focus is on individual care for patients, the data is validated by a physician, respectively the inaccuracy and incompleteness of the data are compensated by the expertise of the therapist. In addition, the human factor can easily deal with inconsistencies in data obtained through different channels (consistency and integrity issues). The secondary use of medical data relies much more on algorithmic and machine processing, where the results of the analysis are much more sensitive to the quality of data. In the secondary

use scenarios the problems related to the integration of data from various sources, as well as the validity of data across relationships, emerge in full force and hinder the effective use of data. The lack of sufficient details on the context in which the data was collected is a major obstacle to identifying the reasons for the inconsistency in the information and how to use it reliably. Without this context, in the presence of contradictions, even the human factor would find it difficult to determine which information is reliable.

Data time characteristic is important, too, as some data may be outdated because of the requirements of the new data processing. On the other side, the research on data can be provided over different time intervals and in these cases, the accumulated data with time characteristics could be extremely useful.

4. Supporting data quality in the main activities of the secondary use of health data

Evaluating and improving the quality of data through its secondary use is a multi-component task. The major factor here is that it depends on the data set, accumulated for the purposes of its primary use. The secondary use is based on the data volume extracted from existing applications, where data is presented in data structures, corresponding to the goals of the primary use. The quality of collected data also depends on the requirements of the primary use.

To reach appropriate levels of data quality in secondary use, new requirements are enforced. This leads to a transformation of organizational processes and changes the groups of data consumers and data providers and the relationships between the stakeholders. New procedures and rules may be enforced. As data will be used for new purposes, new competencies and skills may be required from the participants in data processing, for example, related to data analytics.

Figure 1 summarizes the main components that need to be considered and reorganized in case of secondary use of data.

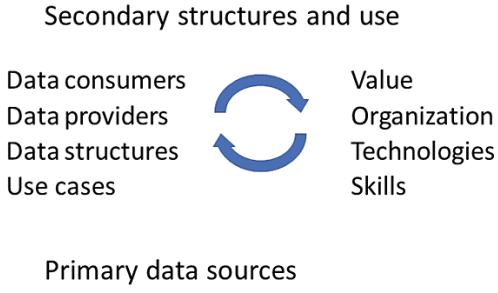

Figure 1: Secondary use of data – areas of changes

4.1. Data reformatting and quality criteria

Evaluating existing data volumes

The sources of health data are clinical trials, electronic health records, wearable technologies, health-insurance claims data, health registry data, etc. that are accumulated gradually. In the case of secondary use mainly sets of summarized health data are processed. This data should be consistent, trustable, and shared across different organizations. It should also be considered that data must be clean and compatible after being processed or coming from other systems but these quality aspects should be reviewed in the context of new uses of data.

This raises two main questions about data interoperability and the use of standards.

When data from different sources is collected, the main obstacle is data compatibility. The use of common data models is a big challenge even on a national level in most domains. The efforts of many organizations and committees in Europe are now focused on these issues and many initiatives are recently presented, especially for health data. Unfortunately, not all existing software applications follow these standards. The latest EC initiatives could foster the European countries to resolve this problem, both on technical and legislative levels.

Discovering new use cases

The goals of secondary use usually differ from the purposes of the initial data collection and use. The initial data collecting purposes could make a strong influence on data entities and their characteristics. Data that are extracted from operational systems and other software applications for routine activities should be carefully checked and validated again. The level of granularity is essential in determining new use cases and it could be different for the data reuse.

Urgency, usefulness, and relevancy could be considered not only as important quality characteristics but also to initiate new use cases, particularly for clinicians.

Collecting the right (quality) data for reuse

Not all gathered data will be valuable for the new use cases. Therefore, not all data will be used in the new environment. The adequacy, regarding the scope of the new data processing, should lead to the criteria for data extraction.

Consistency is a high-level quality dimension and should be addressed in any particular case. Here, consistency can also be considered in the terms of how the extracted data sets are logically compatible with the new scenarios.

Modeling data in new structures

After the data extraction, the new volume of data should be organized into a new structure and processed with new tools.

When data is used for specific research and analysis, in many cases the volume of data will be significantly smaller, so the quality attributes could be supported easily. Traditional relational databases usually fit these purposes.

In other cases, data from different health data sets can be combined for larger studies – statistics or descriptive analysis. Then other, non-relational decisions could be applied.

4.2. Work organization restructuring

The group of stakeholders in the health data processing usually includes patients, healthcare professionals, healthcare regulators, healthcare service providers, policy and lawmakers, information regulators, health system administrators, and others. Among them, data producers and data consumers are most closely involved in data quality aspects. The full engagement of the stakeholders in data entry and data processing is important for reaching high levels of data quality.

Among healthcare workers, clinicians, and managers of health organizations are the most active users of the health software applications. But the list of the stakeholders, as well as their prioritization, may change during the reuse of health data, as new, revised or specific requirements will be set. Particular barriers to data sharing may arise. This can also affect some work procedures and change the roles and responsibilities of the participants. New rules need to be considered.

Relevance, usefulness, completeness, comparability, and conciseness should be considered as key quality attributes, related to the reorganization of data.

4.3. Application reengineering

Building the new infrastructure that corresponds to the new use cases and goals and the adequate technologies are critical for the efficiency of data rework. Several considerations can be helpful here:

- In some cases, only a part of the data needs to be used. This reflects on the size of the applications and the technologies used.
- When new applications are developed, a part of the functionality that supports daily operations on data or different user management could be avoided as research purposes do not require complete administration.

However, the newly developed applications need to provide an appropriate level of usability, a clear understandable interface, and good visualization of the results.

5. Conclusion

The paper highlights the importance of data quality for the secondary use of health data, as successfully resolving data quality issues is a key prerequisite for

significant results in many directions. The secondary use of health data would lead to positive results not only in improving patient treatment but also in optimizing health system organization and spreading out innovations. Recent EC initiatives will help in the establishment of a common environment for health data sharing and will support the reuse of health data among the Member States.

6. Acknowledgments

This research is supported by Project BG05M2P001-1.001-0004 "Universities for Science, Informatics and Technologies in the e-Society (UNITe)" financed by Operational Program "Science and Education for Smart Growth", co-financed by the European Regional Development Fund and National Scientific Program "eHealth" in Bulgaria.

7. References

[1] Act on the Secondary Use of Social Welfare and Health Care Data, URL: https://stm.fi/en/secondary-use-of-health-and-social-data.

[2] COM(2018) 232 final, Towards a common European data space, Brussels, 2018, URL: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52018DC0232.

[3] D. R Schlegel and G. Ficheur, Secondary use of patient data: review of the literature in Yearbook of medical informatics, 26(01), 2016, pp. 68–71.

[4] EU policy on secondary use of health data, Open Data Institute, July 2021, URL: https://theodi.org/article/white-paper-eu-policy-on-secondary-use-of-health-data.

[5] European Health Data Space, URL: https://ec.europa.eu/health/ehealth-digital-health-and-care/european-health-data-space_en.

[6] GDPR, General Data Protection Regulation – Official Legal Text, URL:https://gdpr-info.eu.

[7] G. Fortuna and L.a Bertuzzi, LEAK: The EU Commission's data space for unleashing health data, URL: https://www.euractiv.com/section/digital/news/leak-the-eu-commissions-data-space-for-unleashing-health-data.

[8] ISO/IEC 25012 Software and Data Quality, URL: https://iso25000.com/index.php/en/iso-25000-standards/iso-25012.

[9] K. Kaloyanova, I. Naydenova, Z. Kovacheva, Addressing Data Quality in Healthcare, Proc. of the 14-th conference on Information Systems and Grid Technologies, ISGT 2021, Sofia, Bulgaria, May 28–29, 2021, CEUR-WS.org, vol-2933, pp. 155–164.

[10] K. Van Quathem, S. Choi & A. de Meneses, Leaked: Draft Version of the European Health Data Space Regulation, URL: https://www.insideprivacy.

com/international/european-union/leaked-draft-version-of-the-european-health-data-space-regulation.
[11] Open Data Institute, “Discover which European countries are ready for the secondary use of health data”, 2021, URL: https://theodi.org/project/discover-how-ready-your-country-is-for-the-secondary-use-of-health-data.
[12] P. R. Burton et al., Policies and strategies to facilitate secondary use of research data in the health sciences, International Journal of Epidemiology, 2017, pp. 1729–1733.
[13] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0197.
[14] S. Fox, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc., 2007 Jan 1; 14(1): 1–9.
[15] T. Botsis, G.Hartvigsen, F. Chen, C. Weng, (2010). Secondary use of EHR: Data quality issues and informatics opportunities. Summit on Translational Bioinformatics, 2010: 1–5.
[16] World Health Organization, 2020: Overview of the Data Quality Review (DQR) Framework and Methodology, URL: https://cdn.who.int/media/docs/default-source/data-quality-pages/who-dqrframework-v1-0-overview.pdf.