BBMRI.at
Biobanking and BioMolecular resources Research Infrastructure Austria

Publication by BBMRI.at partner Alpen Adria University

Data quality for federated medical data lakelands

July 2021

Johann Eder, Vladimir A. Shekhovtsov

International Journal of Web Information Systems Vol. 17 No. 5, 2021 pp. 407-426
Emerald Publishing Limited 1744-0084, DOI 10.1108/IJWIS-03-2021-0026

Medical research requires biological material and data of high quality as they are associated with / in biobanks. Medical studies based on data with unknown or questionable quality are useless or even dangerous

The authors of this paper propose an IT architecture to support researchers to efficiently and effectively identify relevant collections of material and data with documented quality for their research projects while observing strict privacy rules.

They describe the landscape of biobanks as federated medical data lakes such as the collections of samples and their annotations in the European federation of biobanks BBMRI-ERIC and developed a conceptual model capturing schema information with quality annotation.

Read original article>>

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/1744-0084.htm

Data quality for federated medical data lakes

Federated medical data lakes

Johann Eder and Vladimir A. Shekhovtsov
University of Klagenfurt, Klagenfurt, Austria

407

Received 19 March 2021
Revised 11 May 2021
Accepted 17 May 2021

Abstract

Purpose – Medical research requires biological material and data collected through biobanks in reliable processes with quality assurance. Medical studies based on data with unknown or questionable quality are useless or even dangerous, as evidenced by recent examples of withdrawn studies. Medical data sets consist of highly sensitive personal data, which has to be protected carefully and is available for research only after the approval of ethics committees. The purpose of this research is to propose an architecture to support researchers to efficiently and effectively identify relevant collections of material and data with documented quality for their research projects while observing strict privacy rules.

Design/methodology/approach – Following a design science approach, this paper develops a conceptual model for capturing and relating metadata of medical data in biobanks to support medical research.

Findings – This study describes the landscape of biobanks as federated medical data lakes such as the collections of samples and their annotations in the European federation of biobanks (Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium, BBMRI-ERIC) and develops a conceptual model capturing schema information with quality annotation. This paper discusses the quality dimensions for data sets for medical research in-depth and proposes representations of both the metadata and data quality documentation with the aim to support researchers to effectively and efficiently identify suitable data sets for medical studies.

Originality/value – This novel conceptual model for metadata for medical data lakes has a unique focus on the high privacy requirements of the data sets contained in medical data lakes and also stands out in the detailed representation of data quality and metadata quality of medical data sets.

Keywords Biobank, Metadata, Data quality, Data lake, Privacy, LOINC, Metadata and ontologies

Paper type Research paper

1. Introduction

Data lakes are architectures for the storage of data for further use (Inmon, 2016; Giebler *et al.*, 2019; Sawadogo and Darmont, 2020). The data lake concept arose with the advent of *big data* as organizations were not able to keep up with the ever-increasing possibilities for collecting and storing data and to integrate all these data in structured data repositories. Data warehouses (Golfarelli and Rizzi, 2018; Vaisman and Zimányi, 2014) require that data, which should be stored in a data warehouse or a data mart, is structured, cleaned, harmonized and integrated, before it is entered into the data warehouse – usually through a carefully designed process of extracting data from the sources, transforming the data into

This work has been supported by the Austrian Bundesministerium für Bildung, Wissenschaft und Forschung within the project BBMRI.AT (GZ 10.470/0010-V/3c/2018).

Emerald Publishing Limited

Bundesministerium Bildung, Wissenschaft und Forschung

Funded by
GZ 10.470/0016-II/3/2013 (2013-2018)
BMBWF-10.470/0010-V/3c/2018 (2018-2023)