The systematic assessment of completeness of public metadata accompanying omics studies

Yu-Ning Huang

Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, California, 90089, USA

yuninghu@usc.edu

ORCID: https://orcid.org/0000-0003-1697-4267

Anushka Rajesh

Department of Pharmacology and Pharmaceutical Sciences, University of Southern California, Los Angeles, CA, 90089, USA

anushkar@usc.edu

ORCID: https://orcid.org/0000-0002-7729-0865

Ram Ayyala

Department of Translational Biomedical Informatics, Keck School of Medicine, University of Southern California, Los Angeles, CA, 90089, USA

rayyala@usc.edu

ORCID: https://orcid.org/0000-0001-7275-271X

Aditya Sarkar

School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, North Campus, Kamand, Mandi, HP 175005, India

asnov2k@gmail.com

ORCID: https://orcid.org/0000-0002-4496-8289

Twitter: SF_AdityaS

Ruiwei Guo

Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, CA, 90089, USA
ruiweigu@usc.edu
ORCID: https://orcid.org/0000-0002-6978-8735

Elizabeth Ling
Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, California, 90089, USA
lingeliz@usc.edu

Irina Nakashidze
Department of Clinical Medicine, Faculty of Natural Science and Health Care, Batumi Shota Rustaveli State University, 6010, Batumi, Georgia
irinanakashidze@yahoo.com
ORCID: https://orcid.org/0000-0001-8934-6312

Man Yee Wong
Department of Translational Genomics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
manyeewo@usc.edu

Jieting Hu
Department of Translational Genomics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
jietingh@usc.edu

Alexey Nosov
Department of Physical Sciences, Santa Monica College, Santa Monica, CA 90405, USA
nosov_alexei@student.smc.edu
ORCID: https://orcid.org/0000-0002-2261-405X

Yutong Chang
Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, CA, 90089, USA
yutongch@usc.edu

Malak S. Abedalthagafi*
Genomics Research Department, Saudi Human Genome Project, King Fahad Medical City and King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia
malthagafi@kacst.edu.sa
ORCID: https://orcid.org/0000-0003-1786-3366
Twitter: malakabed

Serghei Mangul*
Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA
serghei.mangul@gmail.com
ORCID: https://orcid.org/0000-0003-4770-3443
Twitter: smangul1

*These authors jointly supervised this work

Abstract

The scientific community has accumulated enormous amounts of genomic data stored in specialized public repositories. Genomic data is easily accessible and available from public genomic repositories allowing the biomedical community to effectively share the omics datasets. However, improperly annotated or incomplete metadata accompanying the raw omics data can negatively impact the utility of shared for secondary analysis. In this study, we perform a comprehensive analysis under 137 studies over 18,559 samples across six therapeutics fields to

assess the completeness of metadata accompanying omics studies in both publication and its related online repositories across and make observations about how the process of data sharing could be made reliable. This analysis involved an initial literature survey in finding studies based on the seven therapeutic fields, that are Alzheimer's disease, acute myeloid leukemia, cystic fibrosis, cardiovascular diseases, inflammatory bowel disease, sepsis, and tuberculosis. We carefully examined the availability of metadata over nine clinical variables, that included disease condition, age, organism, sex, tissue type, ethnicity, country, mortality, and clinical severity. By comparing the metadata availability in both original publications and online repositories, we observed discrepancies in sharing the metadata. We determine that the overall availability of metadata is 72.8%, where the most complete reported phenotypes are disease condition and organism, and the least is mortality. Additionally, we examined the completeness of metadata reported separately in original publications and online repositories. The completeness of metadata from the original publication across the nine clinical phenotypes is 71.1%. In contrast, the overall completeness of metadata information from the public repositories is 48.6%. Our study is the first one to assess the completeness of metadata accompanying raw data across a large number of studies and phenotypes and opens a crucial discussion about solutions to improve completeness and accessibility of metadata accompanying omics studies.

Introduction

Over the last decade, genomics data has become widely available to the public, allowing researchers to access genomics data across a wide range of diseases and clinical phenotypes. Today, the biomedical community has instant access to various types of genomics data stored at public repositories. The most well-known of these repositories are Gene Expression Omnibus[1] (GEO) and Sequence Read Archive[2](SRA), and ArrayExpress[3]. These repositories allow easy access to the raw data of the genomic studies, which promotes the secondary analysis based on the data and facilitates novel biomedical discoveries.

Although the majority of raw genomics data is made available on the public repositories, the metadata accompanying the raw data is often incomplete[4]. We have previously shown that the

metadata accompanying sepsis-based transcriptomics studies was severely incomplete[4]. The detailed information about the samples, such as their ethnicity, age, disease condition, and sex information, is mostly unavailable when directly searched for on the online repositories[4,5]. The incompleteness of metadata and its improper annotation on the public repositories impede the researchers from reproducing the results reported in the study and further complicate the secondary analysis[6]. The biomedical community has made a tremendous effort to share omics data but little attention is allocated to ensure the completeness of metadata accompanying raw omics data.

However, if metadata accompanying raw data is shared only in the publications, the researchers will require a time-consuming and laborious data-mining process while performing secondary analysis, especially when looking at large scale studies. Using public repositories to share data further simplifies the research process, as one can access all the information though automatic computational methods across a large number of genomic studies. Thus, public repositories have a vital role in ensuring effective data sharing and should have standardized policies in data sharing, including metadata.

In our study, we performed a comprehensive analysis of the completeness of public metadata accompanying omics data on both original publication and online repositories. We examined the completeness of metadata across six therapeutic fields, including Alzheimer's disease, acute myeloid leukemia, cardiovascular diseases, inflammatory bowel disease, tuberculosis, and sepsis, to examine the completeness of metadata in other therapeutic fields. We note that 72.8% of the metadata is available from original publication or public repositories. However, a decreased availability of metadata is observed on public repositories, the availability of metadata information is 48.6% in public repositories, which 22.5% of metadata is lost when metadata is stored in public repositories. Additionally, we note that the most complete reported phenotypes are disease condition and organism, and the least complete phenotypes is mortality

Results

The overall availability of metadata accompanying raw omic data

We have performed a comprehensive analysis of the completeness of the metadata across original scientific publications and public repositories where metadata is accompanying raw omics data for studies relating to numerous disease conditions. To generalize the results obtained from our previous meta-analysis of sepsis cohorts[4], we have extended it to several disease conditions. These include Alzheimer's disease, cardiovascular diseases, inflammatory bowel disease, acute myeloid leukemia, cystic fibrosis, and tuberculosis. We searched the candidate studies in the above field of disease studies with available next-generation sequencing genomics data that has an available GEO accession number on the GEO website. Then we collected all those datasets which were having human samples. We carefully examined 18,559 samples across 137 studies for the availability of the metadata information shared on original publication and public repositories. (Supplementary table 1)

We observed that the overall availability of the samples' metadata information is 72.8% **(Figure 1)**. The most completely shared phenotype disease condition (100%) and organism (100%), followed by the samples' tissue type(99.3%). Conversely, the least complete metadata information is the samples' mortality rate(32.7%), ethnicity information(33.6%), and the clinical severity(48.9%) **(Figure 1, 2)**. Similar trends of completeness of metadata reported can be seen across all studies in six therapeutics fields (**Figure 2**).

Incomplete metadata annotated on public repositories

Next, we investigated the gap between the completeness of metadata reported on original publications and the completeness of metadata reported on public repositories. The completeness of overall metadata from the original publication across the nine clinical phenotypes is 71.1%. In contrast, the overall completeness of metadata information from the public repositories is 48.6%. There are marked inconsistencies in the completeness of metadata reported on public repositories and on original publications, in which 22.5% of the metadata information is lost from storing the metadata information in original publications to online repositories(**Figure 3, 4**).

Most commonly and least commonly reported phenotypes

We found that the clinical phenotypes, organism, tissue types, and disease condition, are the most commonly reported samples' metadata information on the public repositories. Similarly, we observed similar completeness of the metadata information reported in original publications (**Figure 3**). We note that the phenotypes for the organism information is the most completely reported on both original publication and online repositories. On the other hand, the samples' ethnicity(29.9%), mortality of the diseases(31.8%) and clinical severity of the diseases(45.3%) information is the least commonly reported in original publication while the samples' the country of origin(9.5%), ethnicity(12.4%), and mortality of the diseases(10.9%) information is the least commonly reported metadata in public repositories.

Discussion

The effective sharing of metadata on public repositories promises to facilitate the secondary analysis of the omics data to enable novel biological discoveries[4]. Although the samples' metadata can be accessed directly from original publications or from the request to the authors, the process can be time-consuming and error-prone. Additionally, information that is made upon request to the authors has a low success rate[7,8]. Therefore, leveraging previously published data for novel biological discoveries could be facilitated when the metadata accompanying its raw omics data is reported, present in a standardized format, and made available in online repositories. Complete and well annotated metadata shared on public repositories ensure the reproducibility and transparency of biomedical research and enables effective secondary analysis. Improving the accessibility and completeness of metadata are beneficial to the entire scientific community according to the FAIR guidelines, in which the data accompanying the metadata should be Findable, Accessible, Interoperable, and Reusable[9]. It is crucial to ensure metadata is shared on public repositories in addition to original publication since the researchers conducting the secondary analysis will have to make additional efforts to manually examine original publications or contact the authors to request the intended metadata information, which is laborious, time-consuming and error-prone.

According to the results, we note that the overall availability of metadata is 72.8% (**Figure 1**). However, there are great discrepancies in the completeness of metadata reported across different clinical phenotypes. The most complete metadata information is disease condition, organism, and tissue types, which almost 100% of the samples reported the above metadata, while the least complete metadata information is samples' mortality rate, ethnicity, and disease severity. According to figure 5, we can note that all studies have shared some part of information about the samples' metadata, indicating that the authors have some awareness toward sharing the samples' metadata information. However, our results reveal the insufficient information reported, which no studies share complete metadata information accompanying raw data in public repositories (**Figure 4**). Specifically, the samples' metadata information is unevenly available, in which the ethnicity and country information is the least commonly reported metadata information (**Figure 3**). The low availability of the samples' metadata of ethnicity and country information might hinder the future investigations on population-specific omics studies.

There are several challenges in improving the completeness of metadata. First, the limitations in the study design that makes it hard for researchers to collect samples' metadata information. For example, the de-identified samples in the study makes the researchers hard to extract the complete and accurate samples' metadata information. Additionally, if the authors do not include the collection of the clinical phenotypes' information in the study design of the International Review Broad's (IRB) application, the authors would not be allowed to collect complete metadata information during the study design. Second, the participants' privacy concern could also hinder the availability of metadata. It is known that the Health Insurance Portability and Accountability Act of 1996 (HIPAA) protects the study participants' privacy. Hence, it is hard for authors to disclose study participants' clinical phenotypic information due to privacy concerns.

In order to mitigate the incompleteness of metadata reported on public repositories, proper metadata sharing policies should be established, similar to the existing policies for sharing raw omics data. Once the public repositories establish standardized metadata sharing policies, the

reported metadata will be standardized and easy-to-utilize for secondary analyses[13]. In addition, we should increase the education levels of biomedical researchers about the value of rich and properly collected phenotypes, and educate all authors to be aware of the importance of samples' metadata information and include the plans for collecting samples' metadata in the IRB's application process to increase the metadata availability, and reduce the regulatory barriers in metadata-sharing work. In general, well designed metadata sharing policies from the public repositories and increasing the authors' awareness in the collection of complete samples' metadata can be efficient ways to improve the metadata information.

The completeness, unreliableness and correctness of the metadata from online public repositories is not yet fulfilled[14], indicating the need to improve the availability of metadata as well as the correctness of the metadata information. Publicly available, complete, and standardized metadata information shared on public repositories are fundamental elements to enhance the availability, reproducibility and reliability of the samples' metadata accompanying raw data and facilitate the secondary analyses.

Methods

We reviewed published studies for the individual samples' metadata information. We searched the studies in the six therapeutics fields, including Alzheimer's disease, cardiovascular diseases, inflammatory bowel disease, acute myeloid leukemia, cystic fibrosis, and tuberculosis, with available next-generation sequencing genomics data that has an available GEO accession number on the GEO website. Then we collected all those datasets which were having human samples. We examined 18,559 samples across 137 studies for the availability of the metadata information shared on original publication and public repositories (Table 1) and examined the availability of metadata for up to 9 phenotypes, including disease condition, age, organism, sex, tissue type, ethnicity, country, mortality, and clinical severity. We then compared the completeness of both the metadata and checked the metadata information's consistency in original publication and public repositories.

To extract metadata from the public genomic repositories, we downloaded the metadata information from the Gene Expression Omnibus (GEO) repository using our existing Python script. We first scraped GEO datasets' metadata by logging into NCBI FTP using ftp python module and iterated over all the GEO accession numbers listed in a text file, and downloaded it into an Extensible Markup Language (XML) file. XML files are difficult to interpret and work with, so we parsed all the files into readable Comma Separated Values (CSV) files. The CSV sheet was structured in such a way that it contains submission date, release date, last update date, title, accession, type, source, organism, sex, age, BMI, braak stage, brain bank, replicate, molecule, extract protocol and description. Our code can be found in the github repository at https://github.com/Mangul-Lab-USC/metadata_sepsis. Once we have all the CSV files along with the metadata, we examined individual cohorts' metadata information, including disease condition, organism, sex, age, tissue type, cell type, country, ethnicity, clinical severity, and mortality, both in the original publication and the CSV file extracted from the public repository, and analysis the completeness and availability of samples' metadata information accompanying raw data.

Contributions

SM supervised the work. YH analyzes the collected data. AR collected the data from sepsis studies. RG and EL collected the data from tuberculosis studies. IN collected the data from cardiovascular disease studies. RA and MYW collected data from acute myeloid leukemia. JH collected data from inflammatory bowel disease studies. AS and AN collected data from Alzheimer's disease studies. All co-authors contributed to the revision of the manuscript.

Acknowledgements

We acknowledge all authors sharing the information with us.

Code Availability

All code for the study can be found at https://github.com/Mangul-Lab-USC/metadata_sepsis.

Data Availability

All data for the study can be found at https://github.com/Mangul-Lab-USC/metadata_sepsis.

References

1. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Methods Mol. Biol. Clifton NJ* **1418**, 93–110 (2016).
2. Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19-21 (2011).
3. Sarkans, U. *et al.* From ArrayExpress to BioStudies. *Nucleic Acids Res.* **49**, D1502–D1506 (2021).
4. Rajesh, A. *et al.* Improving the completeness of public metadata accompanying omics studies. *Genome Biol.* **22**, 106 (2021).
5. Toczydlowski, R. H. *et al.* Poor data stewardship will hinder global genetic diversity surveillance. *Proc. Natl. Acad. Sci.* **118**, (2021).
6. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **14**, 89–99 (2013).
7. Huang, Y.-N. *et al.* Ancestral diversity is limited in published T cell receptor sequencing studies. *Immunity* **54**, 2177–2179 (2021).
8. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci.* **115**, 2584–2589 (2018).
9. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
10. Bernstein, M. N., Doan, A. & Dewey, C. N. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics* **33**, 2914–2923 (2017).
11. Panahiazar, M., Dumontier, M. & Gevaert, O. Predicting biomedical metadata in CEDAR:

A study of Gene Expression Omnibus (GEO). *J. Biomed. Inform.* **72**, 132–139 (2017).

12. Klie, A. *et al.* Increasing metadata coverage of SRA BioSample entries using deep learning–based named entity recognition. *Database* **2021**, (2021).

13. Pannucci, C. J. & Wilkins, E. G. Identifying and Avoiding Bias in Research. *Plast. Reconstr. Surg.* **126**, 619 (2010).

14. Gozashti, L. & Corbett-Detig, R. Shortcomings of SARS-CoV-2 genomic metadata. *BMC Res. Notes* **14**, 189 (2021).

Figure

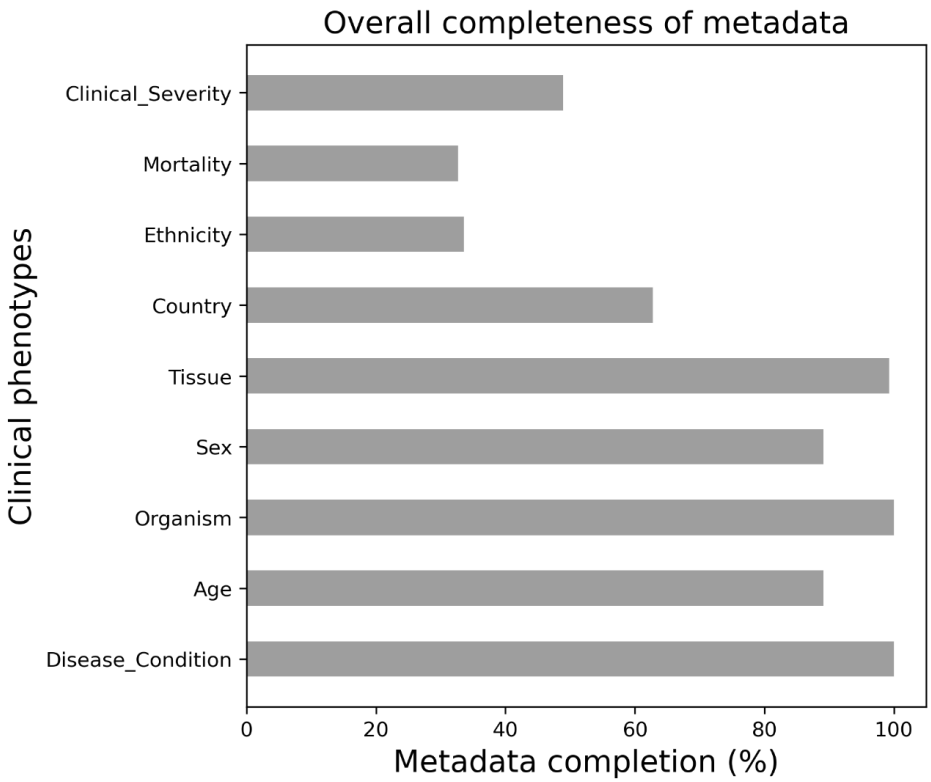

Figure 1. The overall completeness of metadata information that is available from original publications or public repositories.

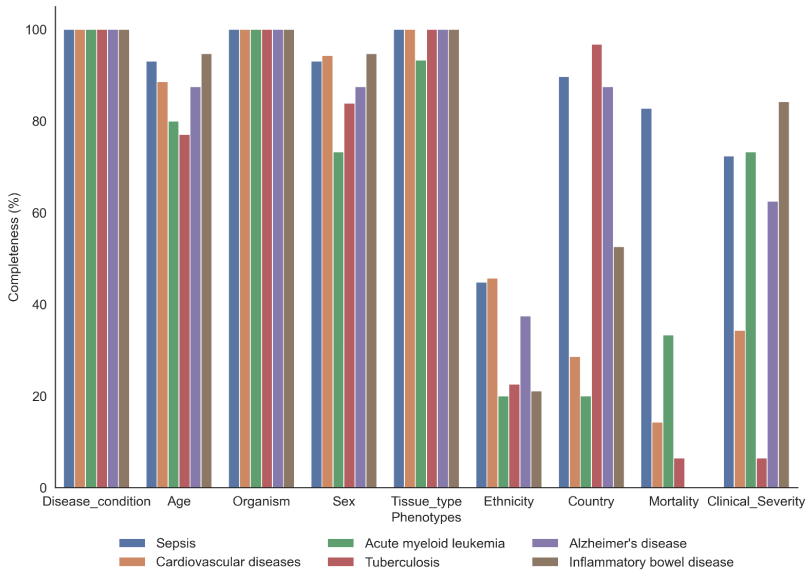

Figure 2. The overall completeness of metadata accompanying raw data across studies in six therapeutic fields, including sepsis, acute myeloid leukemia, Alzheimer's disease, cardiovascular disease, tuberculosis, and inflammatory bowel disease

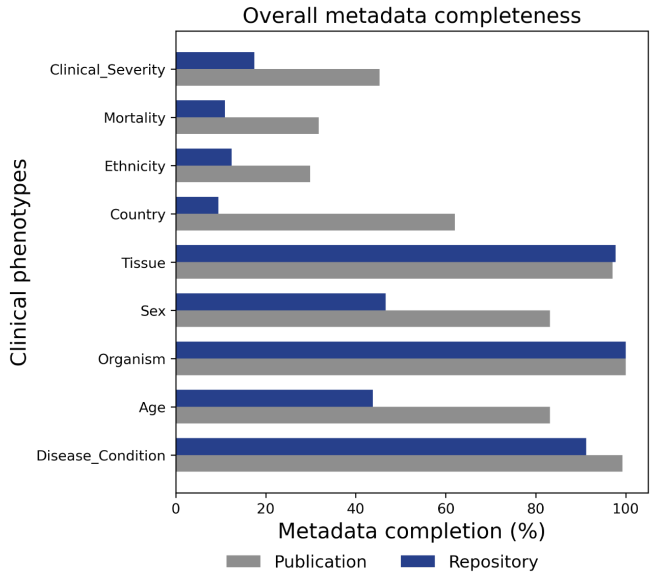

Figure 3. The overall completeness of the samples' metadata across nine clinical variables in orginical publications and public repositories among all studies across six therapeutic fields

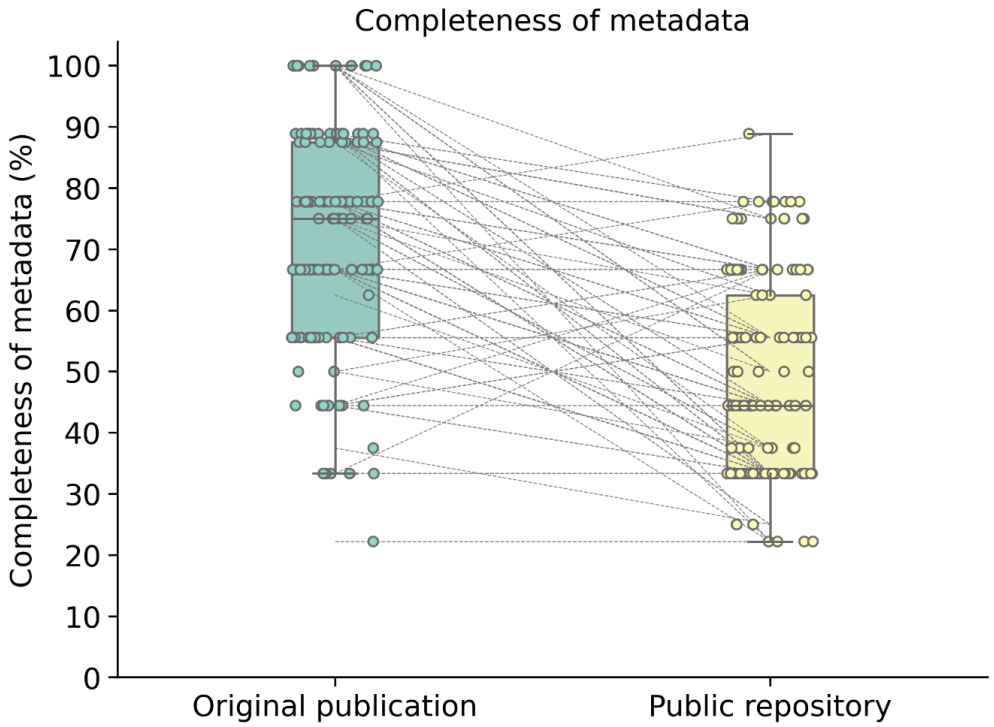

Figure 4. Boxplot for the individual studies' completeness of metadata reported in original publication and public repositories

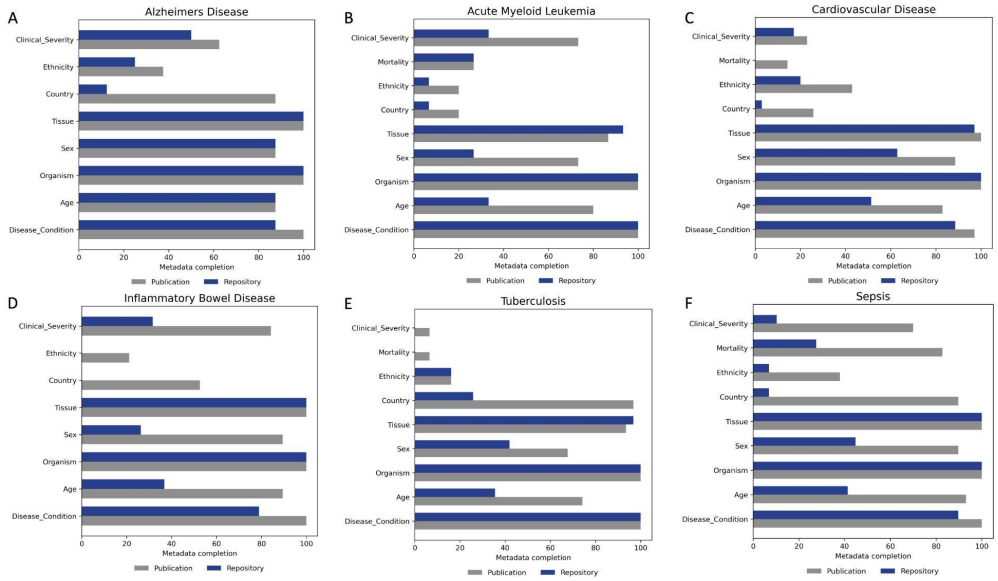

Figure 5. The completeness of metadata across nine clinical phenotypes in six therapeutic fields' studies