Indian Contribution to Medicine Datasets: An Analysis

Nayanthara S[1] and Anila Sulochana[1]

[1]Central University of Tamil Nadu

Abstract

The registry of research data repositories, re3data.org, documents over 1,400 research data repositories worldwide, making it the web's largest and most comprehensive online catalog. The Library and Information Services Department (LIS) of the GFZ German Research Centre for Geosciences, the Computer and Media Service at Humboldt-Universität zu Berlin, the Purdue University Libraries, and the KIT Library at the Karlsruhe Institute of Technology (KIT) are project partners in re3data.org (The Landscape of Research Data Repositories in 2015: A Re3data Analysis, n.d.). After merging with the American DataBib in 2014, re3data.org became a DataCite service in 2016. DataCite is a non-profit international organization that provides a way for researchers to obtain credit and recognition for sharing their research data (The British Library, n.d.). "Research data" refers to a model of raw statistical or visual data gathered from necessary sources during a scientific investigation. A massive amount of data is generated in the medical sciences due to observations, experiments, and clinical investigations. In medicine, there are 961 listed repositories, divided into 48 sub-categories. A few registered repositories are noted as closed, and there are also subjects with no currently listed data repositories. India contributes 2.81% of the total datasets in medicine indexed in re3data.org with 21 datasets. This study aims at analyzing the Indian raw datasets that belong to Medicine in re3.org. The availability and accessibility of research data will aid in discovering new knowledge and developing new approaches for improving research quality. To advance knowledge, researchers affiliated with institutions and organizations should be motivated to deposit and incorporate data gathered through their research into institutional research data repositories or other discipline repositories. There are enormous quantities of organized public health data emerging in various academic, government, or non-commercial disciplines, and making these datasets available, especially in the medical discipline, is extremely necessary to accelerate the research as we live in a world of unforeseeable global medical emergencies. Moreover, re3data.org can potentially improve and promote the best data preservation and management practices.

Keywords: India, Medicine datasets, re3data.org

Corresponding Author: Nayanthara S, Email: nayantharabunglavil@gmail.com and and Anila Sulochana, Email: anila.sulochana@gmail.com

1. Introduction

In today's data-driven world, the field of medicine stands on the precipice of a transformative revolution propelled by the exponential growth of medical datasets and advancements in data analytics. India, with its diverse population, rich medical heritage, and burgeoning technological sector, has the potential to generate valuable medical datasets and shape the future of healthcare on a global scale. The utilization of medical datasets is a cornerstone of modern medical research and practice. These datasets, comprising patient records, clinical trials, imaging studies, genetic profiles, and more, offer a comprehensive view of health trends, disease patterns, treatment outcomes, and population demographics. As the world increasingly embraces evidence-based medicine, the availability of large, diverse, and high-quality datasets becomes pivotal in driving ground-breaking research, enabling personalized treatments, and informing public health policies. The centrality of India in this realm cannot be overstated. With over 1.3 billion people, India is a microcosm of medical diversity, encompassing various genetic, cultural, and socioeconomic factors influencing medical conditions. This diversity translates into a vast repository of medical data that, when properly tapped, can reveal insights into rare diseases, common health issues, and treatment responses that have far-reaching implications. Cooperation between technology companies, research institutions, and healthcare providers has resulted in the development of large databases that record a wide range of medical information, from urban to rural settings, chronic illnesses, to emerging infections. In this article, we carry out a comprehensive review of the Indian contribution to medical datasets. The analysis is carried out regarding fair data principles, citability, source of information, privacy policy, etc., as balancing the power of data and protecting individual rights is a critical discourse underpinning the responsible use of medical datasets in the digital age. We try to discover the data-driven potential of India's healthcare landscape and the collaborative spirit driving transformative shifts through a closer look at these datasets. As we move forward, we must keep in mind that the insights gained from these datasets have the potential to shape a healthier, more informed world.

Research data is vital information gathered from observations, experiments, and clinical studies. The global registry Re3data.org indexes over 2000 repositories, including 568 in medicine. Sharing research data can help with disaster preparedness, disease transmission modelling, tracking health outcomes, and providing reusable data for future research. Re3data.org raises awareness of existing and emerging open datasets in eScience repositories (Greenberg, C.J., & Narang, S. 2020). Researchers need infrastructures for accessibility, stability, and reliability in working with and sharing research data. The re3data.org project indexes research data repositories, and it helps researchers identify appropriate repositories for storage and reuse, categorizing institutional, disciplinary, multidisciplinary, and project-specific RDR. The project's features help researchers find suitable repositories for data storage and search (Pampel et al., 2013). Data sharing is crucial for scholarly research and publishing, improving results and driving discovery. This study examines research data repositories (RDR) and strategies used by countries for efficient organization and optimal use of scientific literature. The re3data registry contains diverse repositories from most countries, with English being the dominant language. Most repositories are open, with over half being disciplinary, and significant data sources include scientific and statistical data and standard office documents (Khan et al., 2023).

2. Objectives

- To figure out India's contribution to the datasets that belong to Medical Sciences in re3data.org.
- To examine the nature and various aspects of the Indian Medicine datasets in re3data.org.
- To understand the collaboration pattern of medicine datasets in re3data.org.

3. Method

The Indian Medicine raw datasets are listed using various filters from re3data.org and then analyzed based on the following aspects:

- Whether these datasets follow fair data principles?
- Are these datasets citable?
- Is the dataset source clearly stated, corroborating the authenticity?
- Who are the major contributors to these datasets?
- Is the privacy policy stated, whether it is of patients or medical data?
- What category of access is provided to these datasets?
- What kind of licensing is used for these datasets?
- What are the keywords used within these datasets?

This data is then analyzed, and a comprehensive picture is drawn regarding the status of the Indian Medicine datasets on re3data.org. These aspects are examined to verify the interoperability, reliability, and authenticity of the datasets.

4. Metadata and Data Standards for Health Domain in India

The "Metadata and Data Standards" initiative taken by the Ministry of Communication and Technologies under the National e-Governance Plan (NeGP) aims to promote the growth of e-Governance within the country by establishing interoperability across e-Governance applications for the seamless sharing of data and services. Under the MDDS initiative, domain-specific committees have been constituted in priority areas (Govt of India, 2013). This initiative focuses on Interoperability. According to the policy, Interoperability at the institutional level would necessitate conversations between public health organizations in order to understand information needs and barriers to better information quality and use—much of which relates to the information collection and recording, patterns of flow and aggregation, and contexts of use of information rather than semantic or technical considerations. The MDDS standard covers a gap in the health domain by providing semantic standardization and a framework for interoperability. Though implementation and adoption will take time, and additional steps must be taken, the effort to get there will be worthwhile. Thus, the MDDS publication is only the beginning of a long journey, not its final destination.

To achieve the goal of Universal Health Coverage, all public and private Health IT systems must eventually merge on a Health Information Exchange. This model addresses MDDS standards to ensure semantic interoperability across all applications, including their data storage, privacy, security, integration, data retrieval, analysis, and information usage. The HIE's core is a registry-based model with disease, facility, and patient registries at the district and state levels. Using individual identifiers, the registry can be indexed and searched. The metadata in the registry will point to the details in the source system. The indicators derived from the state disease registries should be combined and reported to the central disease registry. However, drill-down should be available to obtain granular data on demand.

However, the datasets indexed in re3data.org from India are not exclusively dedicated to medicine. So, this standard is not applied. Though mandates and standards are being developed for curating, preserving, and disseminating medical data, India has yet to implement a uniform code for medical datasets.

The authors analyzed the datasets available based on the mandate proposed by FAIRshare.org, as shown below:

Table 1: Attributes and Conditions by FAIRshare.org

	Data Curation	Data Access Condition	Resource Sustainability	Data Preservation Policy	Data Deposition Condition	Data Versioning	Data Contact Information	Citation To Related Publications	Data Access For Pre Publication Review
Indian Genetic Disease Database	Manual	Open	Developed and maintained at-CSIR: Indian Institute of Chemical Biology	Not clearly mentioned	Open via e-mail. No limit per researcher is mentioned	yes	yes	No	Not clear
Human Proteinpedia	Manual	Open	Developed & maintained by Pandey at Johns Hopkins University, and Institute of Bioinformatics, Bangalore	Not clearly mentioned	Allows research laboratories around the world to contribute and maintain protein annotations.	yes	yes	yes	Not clear
Open Government Data Portal of Surat City	Manual	Open	National Informatics Centre (NIC), Ministry of Electronics, Information Technology, Government of India & Surat Municipal Corporation	Not clearly mentioned	Restricted to GAD-IT	No	yes	No	Not clear

Open Government Data Portal of Odisha	Manual	Open	Government of Odisha & National Informatic Centre of the Government of India	Not clearly mentioned	Restricted to GAD-IT	No	yes	No	Not clear
Open Government Data Portal of Tamil Nadu	Manual	Open	Government of Tamil Nadu	Not clearly mentioned	Restricted to GAD-IT	No	yes	No	Not clear
Maharashtra State Data Bank	Manual	Open	Directorate of Economics and Statistics (DES), Planning Department, Government of Maharastra & Maharashtra State Government	Not clearly mentioned	Restricted to GAD-IT	No	yes	Yes	Not clear

As we can derive from the table, the datasets do not comply satisfactorily with the standards mandated by FAIRshare.

4.1. Indian Medicine Datasets in re3data.org

Table 2: Indian Medicine Datasets in re3data.org

	Repository/Database					
	Maharashtra State Data Bank	**Open Government Data Portal of Odisha**	**Open Government Data Portal of Surat City**	**Human Proteinpedia**	**Open Government Data Portal of Tamil Nadu**	**Indian Genetic Disease Database**
Does the database follow fair data principles?	Yes	Yes	Yes	Yes	Yes	Yes
Is it citable	Yes	Yes	Yes	Yes	Yes	Yes
Type of Database	Open Government Data Portal	Open Government Data Portal	Open Government Data Portal	Community Portal	Open Government Data Portal	Open Government Data Portal
Is the source of the dataset clear?	Yes	Yes	Yes	Yes	Yes	Yes
Major contributors	Govt. of Maharashtra, Mastek Ltd. (for Directorate of Economics and Statistics, Planning Department)	various Departments/ Organizations of Government of Odisha	National Informatics Centre (NIC), Government of India	Pandey at Johns Hopkins University, and Institute of Bioinformatics, Bangalore	Government of Tamil Nadu & National Informatics Centre	Council of Scientific and Industrial Research, Government of India, Ministry of Science and Technology, Department of Biotechnology, Indian Institute of Chemical Biology

Privacy aspects	Access to data is restricted to registered users	Access to data is restricted to registered users	Access to data: Open	Open, Restricted & closed access to data	Access to data: Open	Access to data: Open
Access	Access to research data repository: openAccess to data: restricted	Access to research data repository: openAccess to data: restricted	Access to research data repository: restrictedAccess to data: Open	Access to research data repository: open Access to data: restricted/ closed	Access to research data repository: OpenAccess to data: Open	Access to research data repository: closed Access to data: Open
Licenses	Open Government License	Open Government License/ Copyrights	Copyrights	other	Open Government License	Copyrights
Keywords	Natural Sciences, Medicine, Education Sciences, Public Health, Health Services, Research, Social Medicine etc.	Life Sciences, Medicine, Public Health, Health Services Research, Social Medicine, Public Health Research etc.	Public Health, Social Medicine, Public Health Research etc.	Basic Biological and Medical Research, Human Genetics, Bioinformatics and Theoretical Biology, Biology, Life Sciences, Medicine etc.	Medicine, Public Health, Health Services Research, Social Medicine, Public Health Research etc.	Epidemiology, Medical Biometry, Medical Informatics, Human Genetics, Public Health, Health Services Research, Social Medicine, Medicine, Life Sciences
Is medicine Exclusive?	No	No	No	No	No	No

From Table 1, the following conclusions can be derived:

1. Of the 197 medicine databases with raw data sets, six are from India (3.04%).

- **Maharashtra State Data Bank:** This data platform was initiated by the Government of Maharastra (India) and maintained by Mastek Ltd. (for the Directorate of Economics and Statistics, Planning Department) to consolidate and collate data sets available with the various state departments. Objectively it creates a decision support system that will also serve as a knowledge repository for various information seekers such as researchers, academicians, and the general public (Maharashtra State Data Bank | Re3data.org, n.d.).

- **Open Government Data Portal of Odisha:** It is a platform for supporting the Open Data initiative of the Government of Odisha, which intends to publish datasets collected by them for public use. It also supports widely used file formats suitable for machine processing, thus giving avenues for many more innovative uses of Government Data from different perspectives. This portal has been created under

the Software as A Service (SaaS) model of the Open Government Data (OGD) Platform India of NIC. The data available in the portal are owned by various Departments/Organizations of the Government of Odisha. It follows principles on which data sharing and accessibility must be based, including Openness, Flexibility, Transparency, Quality, Security, and machine-readability (Open Government Data Portal of Odisha | Re3data.org, n.d.).

However, no Health and Family welfare datasets are available now in this repository. The data derived are from other categories available.

- **Open Government Data Portal of Surat City:** It is a platform (designed and developed by the National Informatics Centre (NIC), Government of India) for supporting the Open Data initiative of Surat Municipal Corporation, intended to publish government datasets for public use. The portal has been created under the Software as A Service (SaaS) model of the Open Government Data (OGD) Platform, thus giving avenues for reusing datasets of the City from different perspectives. This Portal has numerous modules: (a) Data Management System (DMS) for contributing data catalogs by various departments for making those available on the front-end website after a due approval process through a defined workflow; (b) Content Management System (CMS) for managing and updating various functionalities and content types of Open Government Data Portal of Surat City; (c) Visitor Relationship Management (VRM) for collating and disseminating viewer feedback on various data catalogs; and (d) Communities module for community users to interact and share their zeal and views with others, who share common interests as that of theirs. This repository is now inaccessible (Open Government Data Portal of Surat City | Re3data.org, n.d.).

- **Human Proteinpedia:** Human Proteinpedia is a community portal for sharing and integrating of human protein data. This is a joint project between Pandey at Johns Hopkins University and the Institute of Bioinformatics, Bangalore. This portal allows research laboratories worldwide to contribute and maintain protein annotations. Human Protein Reference Database (HPRD) integrates data deposited in Human Proteinpedia along with the existing literature curated information in the context of an individual protein. All the public data contributed to Human Proteinpedia can be queried, viewed, and downloaded. Data about post-translational modifications, protein interactions, tissue expression, expression in cell lines, subcellular localization and enzyme-substrate relationships may be deposited (Human Proteinpedia | Re3data.org, n.d.)

- **Open Government Data Portal of Tamil Nadu:** The Open Government Data Portal of Tamil Nadu is a platform (designed by the National Informatics Centre) for the Open Data initiative of the Government of Tamil Nadu. The portal is intended to publish datasets collected by the Tamil Nadu Government for public use from different perspectives. It has been created under Software as A Service (SaaS) model of Open Government Data (OGD) and publishes datasets in open formats like CSV, XLS, ODS/OTS, XML, RDF, KML, GML, etc. This data portal has the following modules, namely (a) Data Management System (DMS) for contributing data catalogs by various state government agencies for making those available

on the front-end website after a due approval process through a defined workflow; (b) Content Management System (CMS) for managing and updating various functionalities and content types; (c) Visitor Relationship Management (VRM) for collating and disseminating viewer feedback on various data catalogs; and (d) Communities module for community users to interact and share their views and common interests with others. It includes different types of datasets generated both in geospatial and non-spatial data classified as shareable data and non-shareable data. Geospatial data consists primarily of satellite data, maps, etc.; and non-spatial data derived from national accounts statistics, price index, census, and surveys produced by a statistical mechanism. It follows the principle of data sharing and accessibility via Openness, Flexibility, Transparency, Quality, Security, and Machine-readable (Open Government Data Portal of Tamil Nadu | Re3data.org, n.d.).

- **Indian Genetic Disease Database:** Indian Genetic Disease Database (IGDD) is an initiative of CSIR Indian Institute of Chemical Biology. It is supported by the Council of Scientific and Industrial Research (CSIR) and the Department of Biotechnology (DBT) of India. The Indian people represent one-sixth of the world's population and consist of an ethnically, geographically, and genetically diverse population. In some communities, the ratio of genetic disorders is relatively high due to consanguineous marriage practiced in the community. This database has been created to keep track of mutations in the causal genes for genetic diseases common in India and help physicians, geneticists, and other professionals retrieve and use the information for the benefit of the public. The database includes scientific information about these genetic **diseases** and disabilities and statistical information about these diseases in today's society (Indian Genetic Disease Database | Re3data.org, n.d.). Data are categorized by body part affected and then by the title of the disease. IGDD release 1.0 holds entries for 52 genetic diseases and 63 related genes collated from 123 reports published during 1993–2010. Currently, 2394 patients and 3366 carriers (resident or non-resident Indian individuals) are enlisted in the database, harbouring 6647 mutations, of which 780 are unique in nature (Pradhan et al., 2010).

2. The major contributors to these Indian datasets are Non-Profit Organizations collaborating with the state government, namely:

Govt. of Maharashtra, Mastek Ltd. (for Directorate of Economics and Statistics, Planning Department), various Departments/Organizations of Government of Odisha, National Informatics Centre (NIC), Government of India, Pandey at Johns Hopkins University, and Institute of Bioinformatics, Bangalore, Government of Tamil Nadu & National Informatics Centre, Council of Scientific and Industrial Research, Government of India, Ministry of Science and Technology, Department of Biotechnology, Indian Institute of Chemical Biology.

3. All six of these databases follow fair data principles, though not to a complete extent.

- **Findable:** Data and resources should be easy for humans and machines to find. This involves providing clear and accurate metadata, assigning persistent identifiers, and ensuring data can be located through standardized search mechanisms.

All six of these databases are findable through the Digital Object Identifiers of their respective repositories.

- **Accessible:** Data should be accessible to a wide range of users. This includes providing open access or controlled access based on well-defined conditions. Access procedures should be clearly documented, and obstacles to accessing the data should be minimized.

  All of them grant access to people though some are restricted through authentication. Even though access is restricted to some of them, researchers can still seek access.

- **Interoperable:** Data should be structured in a way that allows it to be combined and used effectively with other datasets. This involves using standardized formats, data models, and vocabularies and providing clear and unambiguous relationships between data elements.

  All six data repositories are interoperable to a certain extent, where they list and link other related datasets. However, uniformity regarding the terms used cannot be identified, especially as these six data repositories that host the medicine-related datasets are not dedicated to the medical domain. As it is multidisciplinary, a specific medical dataset mandate is not followed. This affects the vocabularies and standardized formats.

- **Reusable**: Data should be well-documented and available in a clear and understandable format, enabling others to use, reproduce, and build upon it. This requires providing rich metadata, clear licensing terms, and guidance on adequately citing and attributing the data.

All six datasets use clear licensing terms. Most of them use Open Government licenses and other various copyrights. As these repositories indexed in re3.org are not exclusively medicine datasets, no uniform code can be found in data formatting or metadata. Still, data are available to download in various formats like XML for Maharashtra State Data Bank and Human Proteinpedia, csv for Open Government Data Portal of Odisha, Open Government Data Portal of Tamil Nadu & Open Government Data Portal of Surat City, and text, CSV, and JSON for the Indian Genetic Disease Database. All six repositories provide information on how to cite and attribute data properly.

The FAIR principles are particularly relevant in the context of modern data-driven research, where collaboration and data sharing play a critical role in advancing knowledge and innovation. By adhering to these principles, organizations and researchers can make their data more valuable, accessible, and impactful to a broader community, fostering scientific progress and societal development.

4. The source of data is authentic in all six data sets as various state governments and the assigned departments or legitimate non-profit organizations carefully curate all these datasets.

5. Most of these data sets are open.Access to the datasets is open for three of the databases, with one partially open and the other two closed. However, when it comes to accessing the databases that host these datasets, four are open, one restricted, and the other closed.

6. Most of them use either any of the various copyrights or Open Government Licenses.

7. All six databases are citable, and the citation is given at each repository's profiles with DataCite citations. DataCite is an international not-for-profit organization that aims to improve data citation in order to: establish easier access to research data on the Internet, increase acceptance of research data as legitimate, citable contributions to the scholarly record, support data archiving that will permit results to be verified and re-purposed for future study (Team, n.d.-b).

8. The Major keywords used are:

- Public Health
- Medicine
- Biology
- Social Medicine
- Public Health Research
- Life Sciences etc.

5. Conclusion

The availability and accessibility of research data will aid in discovering new knowledge and developing new approaches for improving research quality. To advance knowledge, researchers affiliated with institutions and organizations should be motivated to deposit and incorporate data gathered through their research into institutional research data repositories or other discipline repositories. There are enormous quantities of organized public health data emerging in various academic, government, or non-commercial disciplines, and making these datasets available, especially in the medical discipline, is extremely necessary to accelerate the research as we live in a world of unforeseeable global medical emergencies. Furthermore, re3data.org can potentially improve public health data identification, access, and reuse and promote the best data preservation and management practices.

It is essential to recognize that India has not yet become a significant global contributor to medical datasets. The six datasets currently indexed in re3data.org encompass a range of subjects beyond medicine. Prominent medical and research institutions must undertake additional trials to propel the field forward. Given India's vast subcontinental expanse and its diverse genetic and health profiles, there is considerable potential for further advancements. The absence of standardized data formats, coding systems, and data exchange protocols can hinder the integration of different healthcare systems and lead to inefficiencies in managing medical data. Standards such as HL7 (Health Level Seven) and SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) are essential for the smooth exchange of medical information. The availability of high-quality medical datasets is crucial for medical research, drug discovery, and treatment development.

The lack of standardized datasets can hinder the progress of research and innovation in the healthcare sector.

References

Bedford, E., Vitale, C.H., Adair, A., Marshall, B., Buys, C.M., Skeem, D.M., Myntti, J., Paterson, J., Cain, J., & Luschek, K. (2016). Potential data sources from re3data.org.

Govt of India. (2013). Meta Data and Data Standards for Health Domain: Part I Overview Report of the National Committee. In Ministry of Health and Family Welfare. Ministry of Health and Family Welfare, Govt of India.

Greenberg, C.J., & Narang, S. (2020). Re3data.org: Discovery Tool for Reusable Research Data. SocArXiv.

Gundlach, J., Schirmbacher, P., & Dierolf, U. (2013). Making Research Data Repositories Visible: The re3data.org Registry. PLoS ONE, 8.

Khan, A.M., Loan, F.A., Parray, U.Y., & Rashid, S. (2023). Global overview of research data repositories: an analysis of re3data registry. Information Discovery and Delivery.

Mondal, T. (2018). Indian Data Repositories in re3data: A Study. SRELS Journal of Information Management.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.,

Pradhan, S., Sengupta, M., Dutta, A., Bhattacharyya, K., Bag, S. K., Dutta, C., & Ray, K. (2010). Indian genetic disease database. Nucleic Acids Research, 39(Database), D933–D938. https://doi.org/10.1093/nar/gkq1025

Team, D. (n.d.). DataCite's Value. Retrieved June 7, 2023, from https://datacite.org/value.html