Development Standards for Online Repositories for Neglected Infectious Diseases Research Data: Systematic Review

Final Report, Version 1.0

23 October 2017

YOLARX Consultants Inc. | Headquarter: 4540 Circle Road, Montreal, QC, H3W 1Y7, CANADA
Phone: +1 514 903 3389 | Fax: +1 514 316 4912 | www.yolarx.com | Incorporated in Quebec, Canada: 1163777114

SUBMITTED TO:	PRESENTED BY:
Elizabeth Pisani Director Email: pisani@ternyata.org **Ternyata Ltd** 28, Smalley Close London N16 7LE, UK Phone: +44 207 2541654	**Genaro Castillon MD MSc** Medical Advisor Email: genaro.castillon@yolarx.com **Anne-Marie Castilloux MSc** Senior Biostatistician Email: castilloux@yolarx.com **Yola Moride PhD FISPE** President Email: moride@yolarx.com Mobile: +1 (514) 996-1548 **Yolarx Consultants, Inc** 4540 Circle Road Montreal, QC, H3W 1Y7, CANADA Phone: +1 (514) 903-3389 Fax: +1 (514) 316-4912 contact@yolarx.com www.yolarx.com

CONFIDENTIAL

Table of Contents

CONFIDENTIAL

CONFIDENTIAL

CONFIDENTIAL

List of Tables

CONFIDENTIAL

CONFIDENTIAL

List of Figures

CONFIDENTIAL

Abbreviations

CDISC	Clinical Data Interchange Standards Consortium
EMBASE	Excerpta Medica dataBASE
EMR	Electronic Medical Record
LILACS	Latin American and Caribbean Literature on Health Sciences
MEDLINE	Medical Literature Analysis and Retrieval System Online
PCR	Polymerase chain reaction
SDTM	Study Data Tabulation Model
UK	United Kingdom
US	United States

CONFIDENTIAL

1 Executive summary

Background	Data accessibility, sharing, and reuse are essential for the timely translation of research results into knowledge and best practices for improving the global health of humanity. Many public (e.g., NIH) and private organisations (e.g., Project Data Sphere) have developed policies governing research data accessibility and sharing. One of the overarching contributions of data sharing is the generation of new discovery that no single study could provide on its own. This is especially true for neglected infectious diseases that may be rare and/or insufficiently recorded. Data sharing involves the creation of a secure platform where patient-level data obtained from multiple studies is made more accessible to researchers under specific conditions, or in a controlled environment. Through the implementation of repositories, raw data can be transformed into useful codified information, leading to new knowledge that may improve public health and patient care. Standards which become widely adopted can help scientists and data analysts to better utilise, share, and archive the ever-growing amount of health care data. Quality control/assurance and reference standards which maximise comparability of data across different studies and sources are of particular interest as data sharing tools and analytics mature and start to play an expanded role in health care research. While several data sharing initiatives are currently in place in a variety of disease areas (e.g., oncology, tuberculosis), to our knowledge there is no recognized set of international standards for data sharing practices. Standardization may therefore be implemented at the time of data collection or data transmission to the repository.
Objectives	1. To identify existing repositories which house health research data for the four diseases of interest: dengue, leprosy, malaria and tuberculosis; 2. To assess the standards which are currently in place in each of these repositories.
Methods	The systematic review aimed at identifying the existing health data repositories for the four diseases of interest as well as at identifying standards and quality assessment processes currently in use. Using the Cochrane Group recommendations, systematic review included a literature search using MEDLINE, Embase, and LILACS electronic bibliographical databases, and grey literature sources were searched. In order to supplement results from the literature searches and identify additional data repositories not published in the scientific literature or indexed in MEDLINE, Embase or LILACS, repositories were searched using Google and Google Scholar search engines.
Results	A total of 149 data repositories identified in the systematic review include data for all diseases of interest (i.e., dengue, leprosy, malaria, and tuberculosis). This

CONFIDENTIAL

	includes duplicate data repositories which are used for more than one disease of interest. Among these, malaria account for the most data repositories (n=64, 43.0%) followed by tuberculosis (n=57, 38.3%), dengue (n=23, 15.4%) and leprosy (n=5, 3.4%).
Strengths and Limitations	Although these literature searches followed the Cochrane Group recommendations for systematic literature reviews and terms used were broad, the results obtained are limited by the keywords used on the search strategies. Therefore, not all data repositories available for the study of the diseases of interest might have been found.
Conclusion	Following this descriptive study, several repositories were described. Identification of data repositories is crucial to develop agreements and to harmonise data in order help data input, sharing, analysis and reuse. Repositories distribution varies according to the disease of interest. The countries with the most number of data repositories for the four diseases of interest are the US followed by the UK. Most of the data repositories included in this study include aggregate data, which is crucial for planning and guidance of the performance of health systems. However, aggregate data cannot provide the type of detailed information which patient level data can. Mostly, data repositories were owned by a private entity followed by universities and governments. In most cases, data is hosted on websites. Web-based data repositories ease data sharing as its content is available to anyone with internet access. Most of data repositories were created with the purpose of research, which the majority have an open access policy and just a few are restricted and required authorization for the use of data. Open access eliminates the economic and physical barriers that stop access to research data and improves the way researchers conduct and share research.

CONFIDENTIAL

2 Introduction

2.1 Background

Data accessibility, sharing, and reuse are essential for the timely translation of research results into knowledge and best practices for improving the global health of humanity. Many public (e.g., NIH) and private organisations (e.g., Project Data Sphere) have developed policies governing research data accessibility and sharing. One of the overarching contributions of data sharing is the generation of new discovery that no single study could provide on its own. This is especially true for neglected infectious diseases that may be rare and/or insufficiently recorded.

Data sharing involves the creation of a secure platform where patient-level data obtained from multiple studies is made more accessible to researchers under specific conditions, or in a controlled environment. Through the implementation of repositories, raw data can be transformed into useful codified information, leading to new knowledge that may improve public health and patient care.

Data sharing enhances the study of the natural history development and patterns followed by specific diseases over time, risk factors, quality and availability of or gaps in healthcare provided to the populace, and treatment effectiveness. In practice, the origin and types of data are increasingly heterogeneous. For example, data may originate from clinical studies, observational studies, hospitals, routine clinical practice, registries (disease, drug, pregnancy registries), or, as part of the administration of health care programs (e.g., administrative claims database). Such heterogeneity is associated with a lack of standardization in the types of data collected. For example, claims databases are transactional databases which collect data on diagnostic codes (e.g., International Classification of Diseases) or drug dispensing, while clinical databases record disease-specific data such as laboratory test results, genotype, etc. Data sources are categorized into primary (i.e., collected for the purpose of a specific study) or secondary (i.e., collected for other purposes). Typically, primary data collection requires individual patient informed consent while the secondary use of existing data sources does not, although this may vary according to the legislation in place in a given country.

Several models exist for the pooling of these heterogeneous data. In some cases, it may be possible to pool raw data while in most cases, common data models are used, whereby data from each data source must be transformed and/or analysed prior to pooling (the latter with pooling of results subsequently

CONFIDENTIAL

conducted through meta-analysis). Standards which become widely adopted can help scientists and data analysts to better utilise, share, and archive the ever-growing amount of health care data. Quality control/assurance and reference standards which maximise comparability of data across different studies and sources are of particular interest as data sharing tools and analytics mature and start to play an expanded role in health care research.

2.2 Rationale

While several data sharing initiatives are currently in place in a variety of disease areas (e.g., oncology, tuberculosis), to our knowledge there is no recognized set of international standards for data sharing practices. For example, the standards used by the Centres for Disease Control and Prevention in the United States (US) rely predominantly on privacy and security concerns. However, standards used by the Clinical Trials Network at the National Institute on Drug Abuse, aim at defining uniform data elements and tools which can then be mapped into Study Data Tabulation Model (SDTM) to facilitate pooled analyses and cross-product comparisons. Standardization may therefore be implemented at the time of data collection or data transmission to the repository.

Repository designs consist of web interfaces based on a robust security framework which includes role-based data access, data encryption, and digital certification. Authorized users are able to input data directly into the central repository. Alternatively, for centres which are already using the repository data dictionary, data may be transmitted periodically for uploading into the pooled repository. Descriptions of repositories tend to focus on the technology, security, and access policies. However, considerations of the data recorded, data structure, analytical processes and quality assurance often prove to be challenging in building repositories due to the heterogeneity of data sources. Equally important as the hardware and software, is a system that is "user-friendly". As shown in the literature, a major barrier to the implementation of Databases of Prescription Monitoring Programs for Opioids in the US was the difficulty of access by health care providers. Increased end-user involvement in the design of the repository has been shown to increase its usability.

CONFIDENTIAL

3 Objectives

3.1 Main objectives

1. To identify existing repositories which house health research data for the four diseases of interest: dengue, leprosy, malaria and tuberculosis;
2. To assess the standards which are currently in place in each of these repositories;

4 Methods

4.1 Clinical Data Repositories Literature Search

4.1.1 Literature Search Strategy

The systematic review aimed at identifying the existing health data repositories for the four diseases of interest as well as at identifying standards and quality assessment processes currently in use. Using the Cochrane Group recommendations, systematic review included a literature search using MEDLINE, Embase, and LILACS electronic bibliographical databases, and grey literature sources were searched. Full search strategies for each disease and bibliographical database are presented in Appendix 1.1.

4.1.2 Pragmatic Web-based Searches

In order to supplement results from the literature searches and identify additional data repositories not published in the scientific literature or indexed in MEDLINE, Embase or LILACS, repositories were searched using Google and Google Scholar search engines.

A summary of the relevant keywords used for the grey literature search is presented in Table 1 below.

Table 1: Keywords for the Grey Literature Search on Web Sources on Clinical Data Repositories for the Four Diseases of Interest

Data source	Disease of Interest
Repository Data Repository Platform Electronic Registry Database	*Dengue:* Dengue Dengue Virus *Leprosy:*

CONFIDENTIAL

Electronic Medical Record EMR Electronic Health Record EHR	Leprosy *Mycobacterium leprae* *<u>Malaria</u>:* Malaria *Plasmodium falciparum* *Plasmodium knowlesi* *Plasmodium malariae* *Plasmodium ovale* *Plasmodium vivax* *<u>Tuberculosis</u>:* TB Tuberculosis *Mycobacterium tuberculosis*

4.1.3 Study Selection Process and Eligibility Criteria

For each disease of interest and following the exclusion of duplicates, records were screened based on titles and abstracts in order to determine eligibility according to the criteria presented in Table 2 below. In order to avoid the retrieval of obsolete data repositories that are no longer in existence and enhance efficiency of the review process, screening was limited to articles published since 1st January 2010.

Table 2: Inclusion and Exclusion Criteria for Eligibility Assessment of Clinical Literature Sources

Inclusion Criteria	Exclusion Criteria
- Studies written in English, French or Spanish - Studies conducted in humans; - Studies including patients diagnosed with one of the four diseases of interest; - Studies which used data sources available online; - Publications indexed in MEDLINE, Embase and LILACS between January 1st, 2010 and July 5th, 2017 (date last searched)	- Literature reviews - Editorials or Opinions

4.2 Genomics Literature Search

We conducted an additional literature review, which focused on Genomics with the purpose of identifying data repositories that store and share genetic data.

4.2.1 Literature Search Strategy

This additional systematic review aimed at identifying the existing genomics data repositories for the four diseases of interest as well as at identifying standards and quality assessment processes currently in use. We also used the Cochrane Group recommendations for this systematic review using the same databases and sources of information used in the clinical data repository literature search. Full search strategies for each disease of interest and results are presented in Appendix 1.1.

4.2.2 Pragmatic Web-based Searches

Using the same search engines used for clinical data repositories literature search, a summary of the relevant keywords is presented in Table 3 below.

Table 3: Keywords for the Grey Literature Search on Web Sources on Genomic Data Repositories for the 4 Diseases of Interest

Data source	Disease of Interest
Genomics Repository Genomics Data Repository Genomics Platform Genomics Database	*Dengue:* Dengue Dengue Virus *Leprosy:* Leprosy *Mycobacterium leprae* *Malaria:* Malaria *Plasmodium falciparum* *Plasmodium knowlesi* *Plasmodium malariae* *Plasmodium ovale* *Plasmodium vivax* *Tuberculosis:* TB Tuberculosis *Mycobacterium tuberculosis*

4.2.3 Study Selection Process and Eligibility Criteria

After exclusion of duplicates, publications were screened based on titles and abstracts in order to determine eligibility according to the criteria presented in Table 4 below. Screening was also limited to articles published since 1st January 2010.

Table 4: Inclusion and Exclusion Criteria for Eligibility Assessment of Genomics Literature Sources

Inclusion Criteria	Exclusion Criteria
- Data repositories in English, French or Spanish - Data repositories that store data or to conduct studies on one of the four diseases of interest - Studies which used data sources are available online - Publications indexed in MEDLINE, Embase and LILACS between January 1st, 2010 and August 15th, 2017 (date last searched)	- Literature reviews - Editorials or opinions

4.3 Data Extraction Process

Using full text articles, all data sources retained following screening of abstracts as well as data repositories identified through pragmatic web-based searches were reviewed in depth in order to verify their eligibility. At this stage, further exclusions of sources have occurred and a broad list of potential data repositories was developed.

For each data source included in the broad lists, key characteristics, i.e., general characteristics of data repositories, available essential data elements for the specific disease, governance, data accessibility, data management, and ethics, were extracted into a standardized data extraction form that is searchable (see Appendix 1.3 to Appendix 1.6, respectively for: 1) Dengue, 2) Leprosy 3) Malaria, and 4) Tuberculosis.

Data were entered in the extraction matrices using a drop-down list in order to standardize the information. A list of response options and their definitions are presented in the respective disease data extraction form Appendix 1.3 to Appendix 1.6 in the Glossary tab.

CONFIDENTIAL

5 Results

5.1 Systematic review results

Literature and web searches led to the identification of 149 data repositories across the four diseases of interest.

5.1.1 Dengue

Applying the literature search strategy presented in Appendix 1.1 and Appendix 1.2, 574 data repositories were identified: 313 and 261 retrieved from the clinical and genomics literature searches, respectively. Of these, 42 duplicates were removed yielding 532 abstracts screened for eligibility based on titles and abstracts.

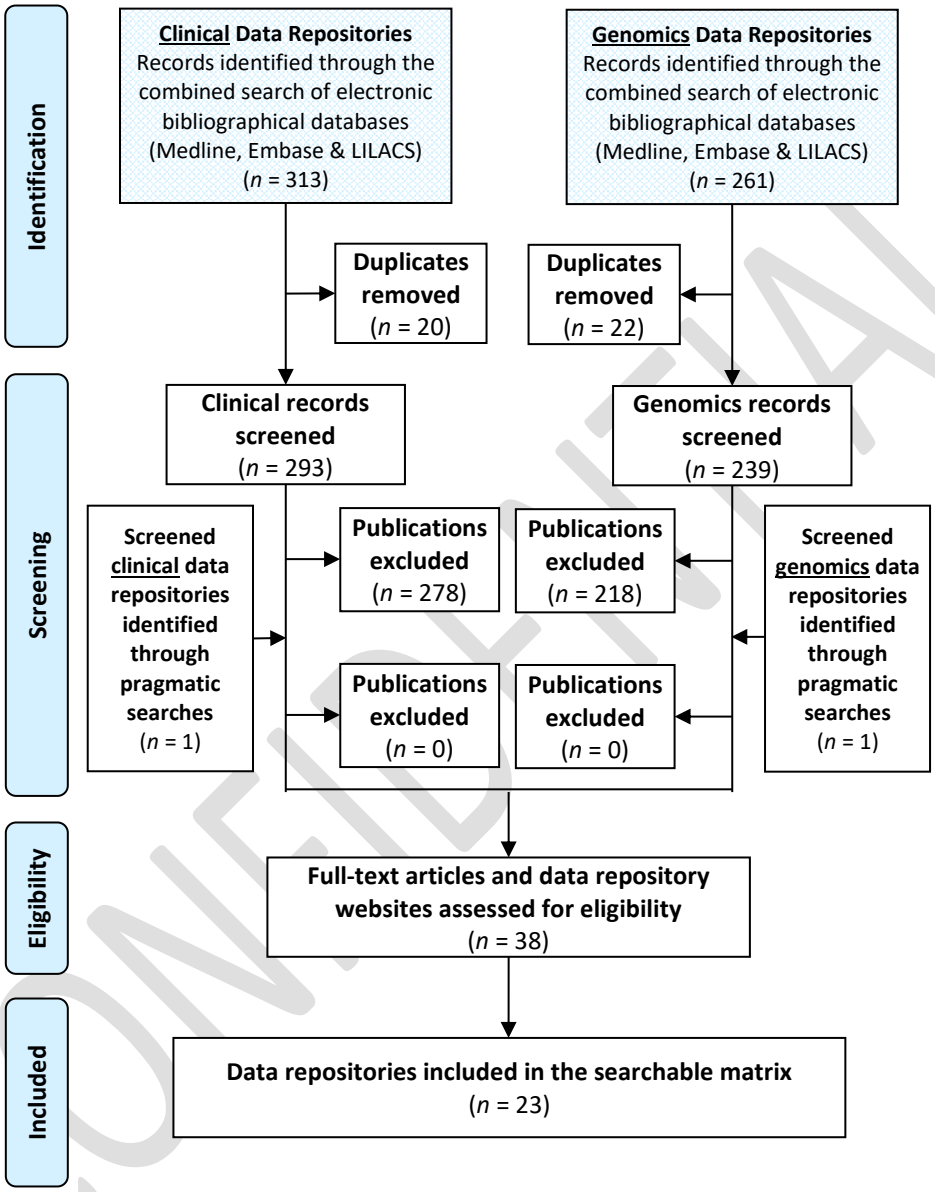
A total of 496 records were excluded during the screening process leaving 36 references retained for an in-depth review of full-text articles. Pragmatic searches led to the identification of 2 additional data repositories for dengue. As a result, 38 full-text articles and data repository websites were assessed for eligibility.

Consequently, 23 data depositories were considered relevant for obtaining information on their key characteristics and available data elements were subsequently abstracted in the standardized information form.

Search results are summarized in Figure 1 below.

CONFIDENTIAL

Figure 1: Quorum Chart of the Literature and Pragmatic Web-based Searches for Dengue Since 2010



CONFIDENTIAL

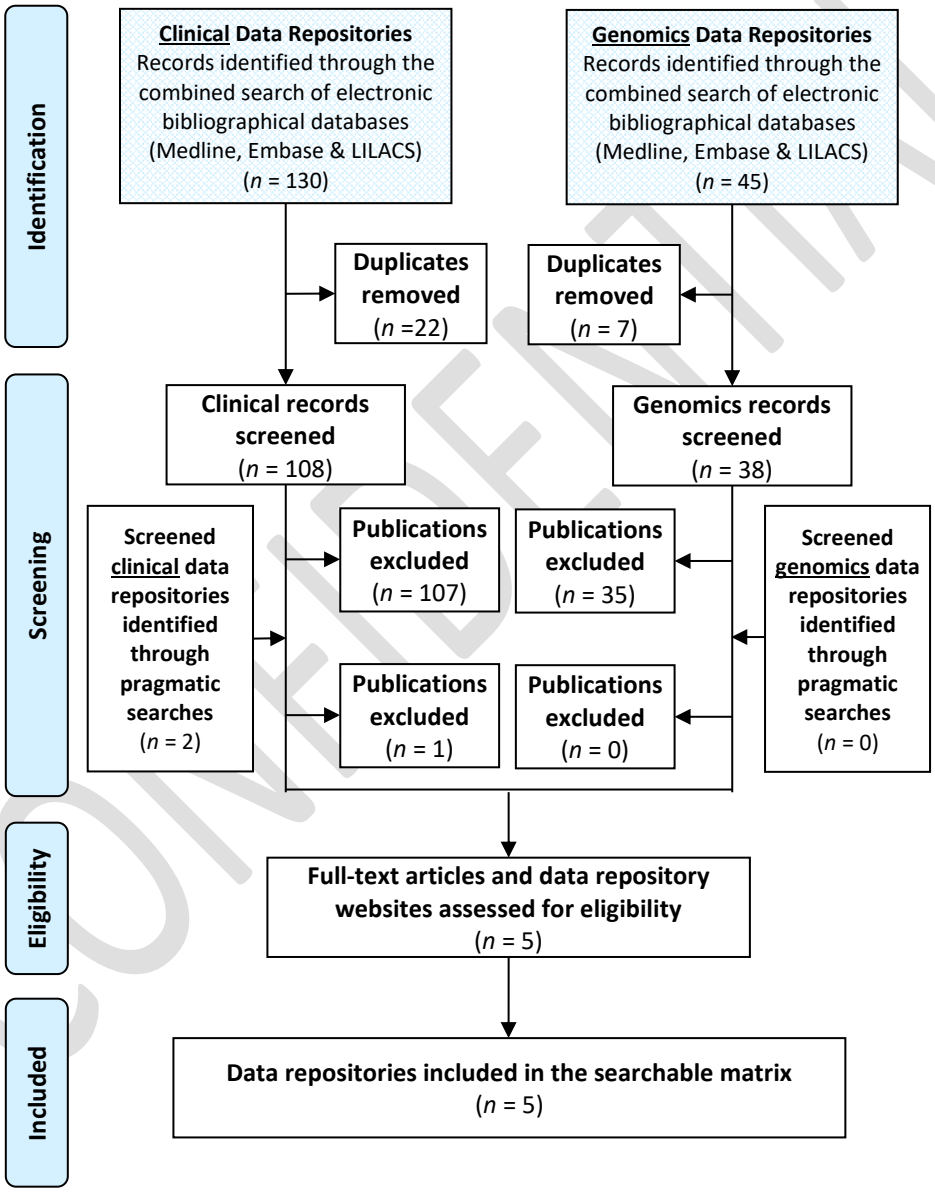
5.1.2 Leprosy

For leprosy, 175 references were identified (130 and 45 for clinical and genomics data repositories, respectively). Of these, 29 duplicates were excluded. After having screened 146 publications, 142 references were excluded yielding four articles corresponding to four data repositories. Additionally, only

1 data repository was found on grey literature searches, yielding a total of 5 data repositories for data extraction.

A summary of these findings is presented in the Figure 2 below.

Figure 2: Quorum Chart of the Literature and Pragmatic Web-based Searches for Leprosy since 2010



CONFIDENTIAL

5.1.3 Malaria

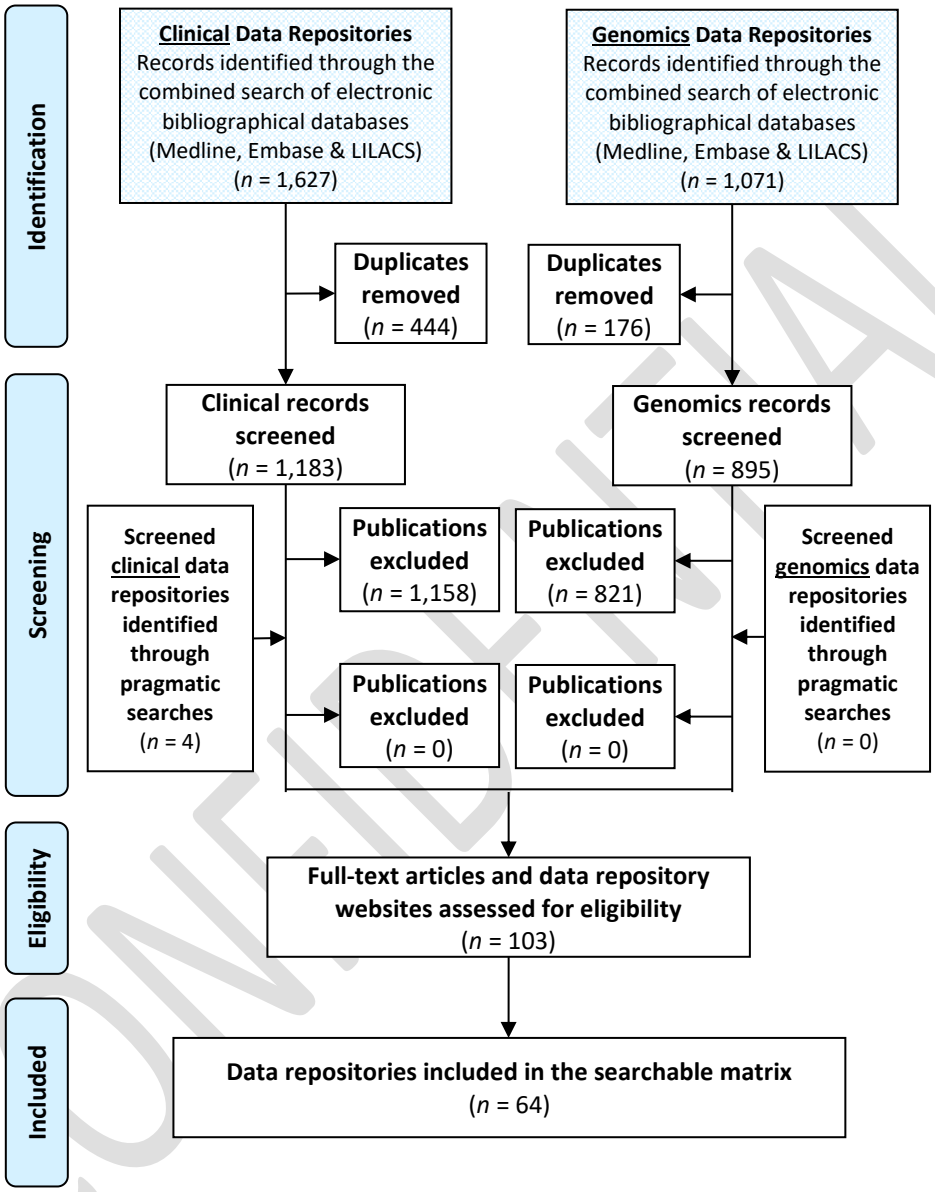
A total of 2,698 abstracts were found (1,627 and 1,071 from the clinical and genomics strategies, respectively). Of these, 620 duplicates were removed yielding 2,078 for the screening process.

A total of 1,979 records were excluded during screening. In addition, grey literature searches for clinical data repositories led to the identification of four data repositories, whereas there were none sources found for genomics that were not already detected from the genomics literature search strategy. Consequently, a total of 103 full-text articles and data repository websites where eligible, which yielded 64 data repositories included for data extraction.

A summary of these findings is presented in Figure 3 below.

CONFIDENTIAL

Figure 3: Quorum Chart of the Literature and Pragmatic Web-based Searches for Malaria since 2010



CONFIDENTIAL

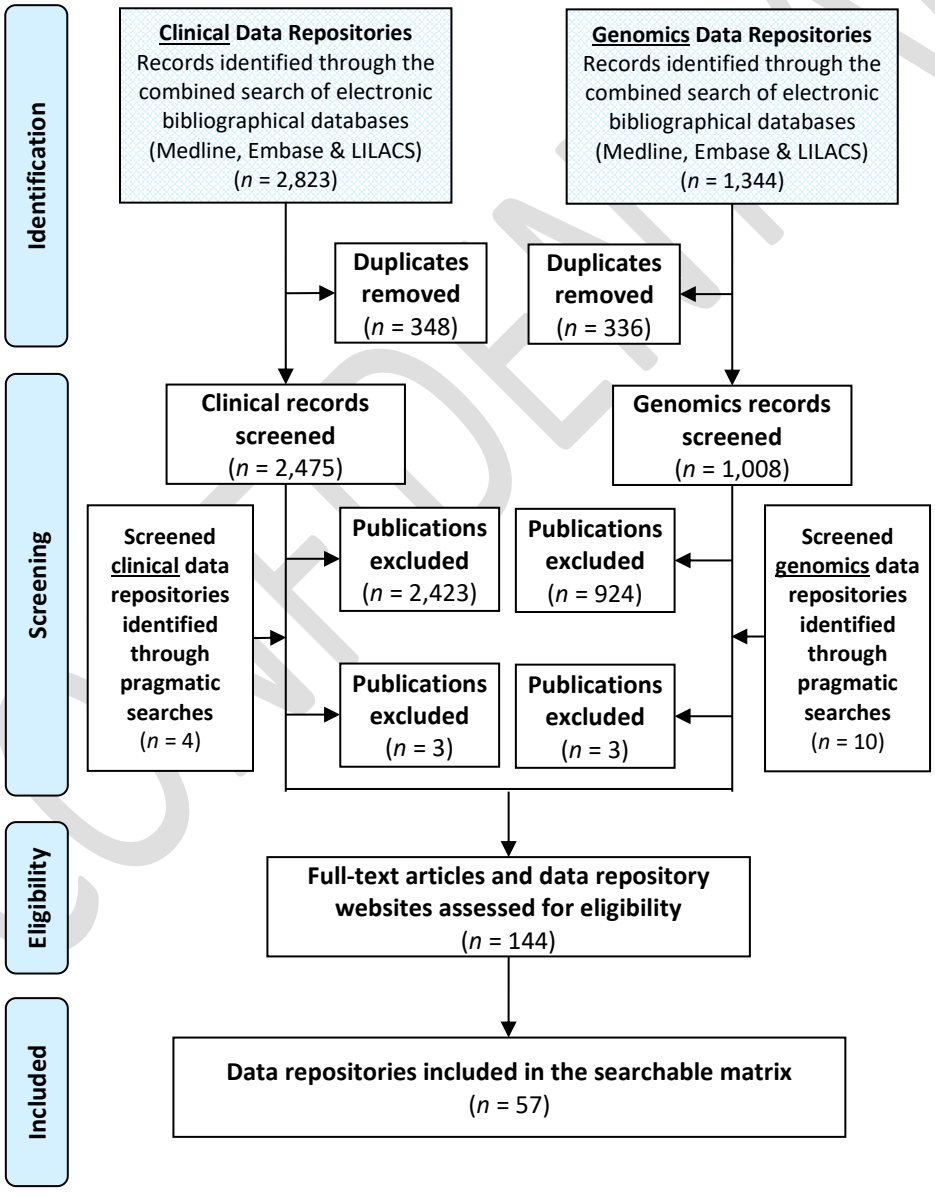
5.1.4 Tuberculosis

A total of 4,167 references were identified (2,823 and 1,344 from the clinical and genomics strategies, respectively). Of these, 684 duplicates were removed.

After screening 3,483 publications, 3,347 were excluded yielding 136 references with tuberculosis data repositories. Grey literature led to the identification of another 14 references, of which six were excluded yielding eight data repositories.

A total of 144 full-text articles and data repository websites were assessed for eligibility, which yielded 57 data repositories related to the study of tuberculosis (see Figure 4).

Figure 4: Quorum Chart of the Literature and Pragmatic Web-based Searches for Tuberculosis since 2010



CONFIDENTIAL

5.2 Description of Data Sources Included in the Broad Lists

This section details the characteristics of the data repositories included in the broad list. Table 5 provides a summary of the number of data sources included in each broad list.

Table 5: Summary Table of Number of Data Repositories Included in the Broad Lists

Number of data repositories included in the broad lists, n(%)	
Dengue	23 (15.4)
Leprosy	5 (3.4)
Malaria	64 (43.0)
Tuberculosis	57 (38.3)
Total	149 (100)

5.2.1 Overview of General Data Sources

A total of 149 data repositories identified in the systematic review include data for all diseases of interest (i.e., dengue, leprosy, malaria, and tuberculosis). This includes duplicate data repositories which are used for more than one disease of interest. Among these, malaria account for the most data repositories (n=64, 43.0%) followed by tuberculosis (n=57, 38.3%), dengue (n=23, 15.4%) and leprosy (n=5, 3.4%).

5.2.2 Dengue

5.2.2.1 Coverage of Data Repositories

A total of 23 data repositories included data for dengue. The majority were developed in America (n=13, 56.5%), followed by Europe (n=4, 17.5%), Asia and International (n=3, 13.0% each). Further details are presented in Table 6 below. Appendix 1.3 presents a list of selected data repositories involving dengue.

Table 6: Data Repositories for Dengue by Continent and Country

Continent	Country	n (%) (N=23)
America	US	8 (34.8)
	Canada	1 (4.3)
	Brazil	2 (8.7)

CONFIDENTIAL

	Colombia	2 (8.7)
Asia	India	1 (4.3)
	Taiwan	1 (4.3)
	Thailand	1 (4.3)
Europe	Switzerland	1 (4.3)
	UK	3 (13.0)
International	International	3 (13.0)

A list of selected data repositories involving this disease is presented in Appendix 1.3.

5.2.2.2 Characteristics of Dengue Data Repositories

Most of data repositories (n=15, 65.2%) were considered platforms of aggregated data. Five (21.7%) databases involved patients at hospital and ambulatory level, whereas 2 (8.7%) were in a hospital setting. A total of 17 (73.9%) included the general population and 26.1% covered disease specific populations. Most of data repositories (n=15, 65.2%) were developed for research. More details are shown in Table 7 below.

Table 7: Distribution of Characteristics of Dengue Data Repositories

Characteristic	n (%) (N=23)
Category of data repository	
Metadata repository	1 (4.3)
Platform of aggregated data	15 (65.2)
Curated data	7 (30.4)
Setting	
Hospital	2 (8.7)
Hospital & ambulatory	5 (21.7)
Not applicable	10 (43.5)
Unknown	6 (26.1)
Population covered	
General population	17 (73.9)
Disease specific	6 (26.1)
Purpose of data repository	
Research	15 (65.2)
Surveillance	5 (21.7)
Patient care	3 (13.0)

CONFIDENTIAL

5.2.2.3 Availability of Essential Data for Dengue

The distribution of available data on dengue data repositories is presented in Table 8 below. Overall, most of data repositories included genomics data (n=15, 65.2%). The patients profile information (e.g., age, gender) was available only in 34.8% (n=8). For clinical information, such as symptoms, haemorrhagic manifestations, and fever, data was only available for 8.7% (n=2). Diagnostic tests (e.g., PCR, serology identification exams) was obtainable for 21.7% (n=5) whereas it was not available in 60.9% (n=14). Laboratory data (e.g., platelets, hematocrit) and pharmacological treatments (e.g., inpatient and outpatient treatments, drug names) and other treatments data (e.g., blood transfusion, platelet transfusion) was found in only 8.7% (n=2), while 69.6% (n=16) was not available. As for outcomes (e.g., death, hospitalization, recovery), data was available in 26.1% (n=6). Safety data was not available in any data repositories.

Table 8: Distribution of Available Data on Dengue

Data	n (%) (N=23)
Genomics data	
Yes	15 (65.2)
No	1 (4.3)
Unknown	7 (30.4)
Patient profile	
Yes	8 (34.8)
No	14 (60.9)
Unknown	1 (4.3)
Clinical information	
Yes	2 (8.7)
No	16 (69.6)
Unknown	5 (21.7)
Diagnostic tests	
Yes	5 (21.7)
No	14 (60.9)
Unknown	4 (17.4)
Laboratory data	
Yes	2 (8.7)
No	16 (69.6)
Unknown	5 (21.7)
Pharmacological treatments	

CONFIDENTIAL

Yes	2 (8.7)
No	16 (69.6)
Unknown	5 (21.7)
Other treatments	
Yes	2 (8.7)
No	16 (69.6)
Unknown	5 (21.7)
Outcomes	
Yes	6 (26.1)
No	15 (65.2)
Unknown	2 (8.7)
Safety data	
Yes	0 (0.0)
No	16 (69.6)
Unknown	7 (30.4)

5.2.2.4 Data Governance, Curation and Sustainability

A total of 19 (82.6%) data repositories store their data on a website. More than half repositories were owned by a governmental entity. Permanent public funding was the most common financial source (n=12, 52.2%). None of the data repositories stated the presence or absence of a succession plan nor the use of a back-up or migration standards. Further information is shown in Table 9 below.

Table 9: Distribution of Governance, Curation and Sustainability Information of Dengue Data Repositories

Data governance, curation and sustainability	n (%) (N=23)
Infrastructure / hosting location	
Website	19 (82.6)
Unknown	4 (17.4)
Ownership	
University	4 (17.4)
Private	4 (17.4)
Government	13 (56.5)
Hospital	2 (8.7)
Funding	
Permanent public funding	12 (52.2)
Grants	3 (13.0)

CONFIDENTIAL

Unknown	8 (34.8)
Succession plan	
Unknown	23 (100.0)
Back-up and migration standards	
Unknown	23 (100.0)

5.2.2.5 Data Accessibility

The standard to ensure discoverability of databases was not obtained in any of the repositories, however, as these repositories were found either on literature search or grey literature, we can assume that all have a discoverability standard. Almost 70% (n=16) of data depositories have their data open for everybody. In order to get access to data, there were 30.4% (n=7) data repositories that did not require any procedure. However, a presence of a data access procedure was unknown in 43.5% (n=10) of repositories. Linkage capabilities was available in only 13.0% (n=3). However, this information was not available in 82.6% (n=19) of data repositories for this disease (see Table 10).

Table 10: Distribution of Data Accessibility Information of Dengue Data Repositories

Data accessibility	n (%) (N=23)
Discoverability	
Unknown	23 (100.0)
Access policy	
Open access	16 (69.6)
On request	1 (4.3)
Restricted	4 (17.4)
Unknown	2 (8.7)
Who can access?	
Everybody	16 (69.6)
Designated research units	3 (13.0)
External researchers	2 (8.7)
Unknown	2 (8.7)
Access procedure	
Data access committee	1 (4.3)
Authorization by government regulatory agencies	1 (4.3)
Website registration	4 (17.4)
None	7 (30.4)

CONFIDENTIAL

Unknown	10 (43.5)
Linkage capacities	
Yes	3 (13.0)
No	1 (4.3)
Unknown	19 (82.6)

5.2.2.6 Data Management

Most of data repositories (n=22, 95.7%) had data entered by curators. The information is being coded in only 26.1% (n=6) data repositories. However, the information is unknown in 65.2% (n=15). Information on coding or standardization procedure was unknown in 82.6% (n=19) of data repositories. Only 13.0% (n=3) of data sources indicated the use of a pharmaceutical, disease or procedures coding. Quality control and data verification procedures are used in only 13.0% (n=3). However, this information was not available in 87.0% (n=20) of repositories. Four (17.4%) repositories update their information in real time, 8.7% (n=2) weekly and 4.3% (n=1) daily, monthly, every two months each. One database had no data update since 2010. More than 60% offer support for users and data depositors. Further details are presented in Table 11 below.

Table 11: Distribution of Data Management Information of Dengue Data Repositories

Data management	n (%) (N=23)
Data entry	
Curator	22 (95.7)
Unknown	1 (4.3)
Coding / Data standardization	
Yes	6 (26.1)
No	2 (8.7)
Unknown	15 (65.2)
Coding or standardization procedure	
Standardized data collection tools	2 (8.7)
No	2 (8.7)
Unknown	19 (82.6)
Pharmaceutical coding / Diseases coding / Procedures coding	
Yes	3 (13.0)
No	6 (26.1)

CONFIDENTIAL

Unknown	14 (60.9)
Data verification / Quality control procedures	
Yes	3 (13.0)
Unknown	20 (87.0)
Frequency of update	
Real time	4 (17.4)
Daily	1 (4.3)
Weekly	2 (8.7)
Monthly	1 (4.3)
Two months	1 (4.3)
No more updates since 2010	1 (4.3)
Unknown	13 (56.5)
Availability of follow-up data	
No	5 (21.7)
Unknown	18 (78.3)
Support for data depositors	
Yes	14 (60.9)
Unknown	9 (39.1)
Support for data users	
Yes	15 (65.2)
Unknown	8 (34.8)

5.2.2.7 Ethics

In order to access a database, it is not required to get the approval of an ethics committee in 21.7% (n=5) of the databases, whereas 13.0% (n=3) do require approval. However, the need for an ethics committee approval is unknown for 65.2% (n=15) of the databases (see Table 12 below).

Table 12: Distribution of Ethics Data of Dengue Data Repositories

Ethics data	n (%) (N=23)
Patient informed consent	
No	6 (26.1)
Unknown	17 (73.9)
Requires ethics committee approval	
Yes	3 (13.0)
No	5 (21.7)
Unknown	15 (65.2)

CONFIDENTIAL

Standards for data anonymization	
Yes	3 (13.0)
No	2 (8.7)
Unknown	18 (78.3)

5.2.3 Leprosy

5.2.3.1 Coverage of Data Repositories

Only 5 data repositories for the study of leprosy were identified. Three (60.0%) were developed in America, followed by Europe and International with 1 (20.0%) data repository each. Additional information is shown in Table 13 below. Appendix 1.4 presents a list of selected data repositories involving leprosy.

Table 13: Data Repositories for Leprosy by Continent and Country

Continent	Country	n (%) (N=5)
America	US	2 (40.0)
	Brazil	1 (20.0)
Europe	Switzerland	1 (20.0)
International	International	1 (20.0)

5.2.3.2 Characteristics of Leprosy Data Repositories

A total of 3 (60.0%) data repositories were considered of curated data type. Two (40.0%) repositories were developed in a hospital or ambulatory setting. Three (60.0%) and 2 (40.0%) were based on general and specific population respectively. A total of 3 (60.0%) data repositories were developed for research purposes. Further details are shown in Table 14 below.

Table 14: Distribution of Characteristics of Leprosy Data Repositories

Characteristic	n (%) (N=5)
Category of data repository	
Platform of aggregated data	2 (40.0)
Curated data	3 (60.0)
Setting	
Hospital & ambulatory	2 (40.0)
Not applicable	3 (60.0)

CONFIDENTIAL

Population covered	
General population	3 (60.0)
Disease specific	2 (40.0)
Purpose of data repository	
Research	3 (60.0)
Surveillance	1 (20.0)
Patient care	1 (20.0)

5.2.3.3 Availability of Essential Data for Leprosy

A total of 3 (60.0%) data repositories include genomics data. Patient profile data, such as age and gender, was available in only 20.0% (n=1) of the repositories. A small proportion (20.0%, n=1) of repositories includes clinical information (e.g., clinical presentation, degree of disability, spectrum of disease [paucibacillary, multibacillary or borderline]). Data on diagnostic tests (e.g., skin biopsy, skin smear, PCR), laboratory exams, and pharmacological treatments (e.g., inpatient and outpatient treatments, drug name) was available in only 20.0% (n=1) of data repositories. A proportion of 40.0% (n=2) of data repositories included outcome data (e.g., death, recovery, hospitalization). Table 15 presents further information.

Table 15: Distribution of Available Data on Leprosy

Data	n (%) (N=5)
Genomics data	
Yes	3 (60.0)
No	1 (20.0)
Unknown	1 (20.0)
Patient profile	
Yes	1 (20.0)
No	3 (60.0)
Unknown	1 (20.0)
Clinical information	
Yes	1 (20.0)
No	3 (60.0)
Unknown	1 (20.0)
Diagnostic tests	
Yes	1 (20.0)
No	3 (60.0)
Unknown	1 (20.0)

CONFIDENTIAL

Laboratory data	
Yes	1 (20.0)
No	3 (60.0)
Unknown	1 (20.0)
Pharmacological treatments	
Yes	1 (20.0)
No	3 (60.0)
Unknown	1 (20.0)
Other treatments	
No	3 (60.0)
Unknown	2 (40.0)
Outcomes	
Yes	2 (40.0)
No	3 (60.0)
Safety data	
No	3 (60.0)
Unknown	2 (40.0)

5.2.3.4 Data Governance, Curation and Sustainability

All identified data repositories for leprosy store data on their website. Data repositories were mostly owned by government (n=2, 40.0%) and universities (n=2, 40.0%). Two (40.0%) repositories have a permanent public funding. No data repository provided information on a succession plan or a back-up and migration standards. More detailed information is shown in Table 16.

Table 16: Distribution of Governance, Curation and Sustainability Information of Leprosy Data Repositories

Data governance, curation and sustainability	n (%) (N=5)
Infrastructure / hosting location	
Website	5 (100.0)
Ownership	
Government	2 (40.0)
University	2 (40.0)
Public-private	1 (20.0)
Funding	
Permanent public funding	2 (40.0)
Private initiative	1 (20.0)

CONFIDENTIAL

Public-private funding	1 (20.0)
Unknown	1 (20.0)
Succession plan	
Unknown	5 (100.0)
Back-up and migration standards	
Unknown	5 (100.0)

5.2.3.5 Data Accessibility

The standard to ensure discoverability of databases was not obtained in any of the repositories, however, as these repositories were found either on literature search or grey literature, we can assume that all have a discoverability standard. Three (60.0%) data depositories have their data open for everybody. Linkage capabilities was available in only 1 repository (see Table 17).

Table 17: Distribution of Data Accessibility Information of Leprosy Data Repositories

Data accessibility	n (%) (N=5)
Discoverability	
Unknown	5 (100.0)
Access policy	
Open access	3 (60.0)
Restricted	2 (40.0)
Who can access?	
Everybody	3 (60.0)
Designated research units	1 (20.0)
External researchers	1 (20.0)
Access procedure	
Data access committee	1 (20.0)
Authorization by Government Regulatory Agencies	1 (20.0)
None	2 (40.0)
Unknown	1 (20.0)
Linkage capacities	
Yes	1 (20.0)
No	1 (20.0)
Unknown	3 (60.0)

CONFIDENTIAL

5.2.3.6 Data Management

Most information on data management was not provided. See details presented in Table 18 below.

Table 18: Distribution of Data Management Information of Leprosy Data Repositories

Data management	n (%) (N=5)
Data entry	
Curator	2 (40.0)
Unknown	3 (60.0)
Coding / data standardization	
Yes	1 (20.0)
Unknown	4 (80.0)
Coding or standardization procedure	
Standardized data collection tools	1 (20.0)
Unknown	4 (80.0)
Pharmaceutical coding / Diseases coding / Procedures coding	
Yes	1 (20.0)
Unknown	4 (80.0)
Data verification / Quality control procedures	
Unknown	5 (100.0)
Frequency of update	
Real time	1 (20.0)
Unknown	4 (80.0)
Availability of follow-up data	
Unknown	5 (100.0)
Support for data depositors	
Yes	1 (20.0)
Unknown	4 (80.0)
Support for data users	
Yes	1 (20.0)
Unknown	4 (80.0)

CONFIDENTIAL

5.2.3.7 Ethics

In order to access a database, it is not required to get the approval of an ethics committee for 1 database, 1 does require approval, and 3 did not have any information (see Table 19 below).

Table 19: Distribution of Ethics Data of Leprosy Data Repositories

Ethics data	n (%) (N=5)
Patient informed consent	
No	1 (20.0)
Unknown	4 (80.0)
Requires ethics committee approval	
Yes	1 (20.0)
No	1 (20.0)
Unknown	3 (60.0)
Standards for data anonymization	
Yes	1 (20.0)
Unknown	4 (80.0)

5.2.4 Malaria

5.2.4.1 Coverage of Data Repositories

A total of 64 data repositories for the study of malaria were identified. Most repositories were developed in Europe (n=21, 31.8%), followed by America (n=17, 26.6%). Additional information is shown in Table 20 below. Appendix 1.5 presents a list of selected data repositories involving malaria.

Table 20: Data Repositories for Malaria by Continent and Country

Continent	Country	n (%) (N=64)
Africa	Burkina Faso	1 (1.6)
	Ethiopia	1 (1.6)
	Malawi	1 (1.6)
	Mozambique	1 (1.6)
	Rwanda	1 (1.6)
America	Brazil	1 (1.6)
	Canada	2 (3.1)
	US	14 (21.9)

CONFIDENTIAL

Asia	Japan	5 (7.8)
	Singapore	1 (1.6)
	Thailand	1 (1.6)
	United Arab Emirates	1 (1.6)
Europe	Europe	5 (7.8)
	France	1 (1.6)
	Italy	1 (1.6)
	Netherlands	1 (1.6)
	UK	13 (20.3)
Oceania	Australia	1 (1.6)
International	International	12 (18.8)

5.2.4.2 Characteristics of Malaria Data Repositories

A total of 36 (56.3%) data repositories were considered as a platform of aggregated data. A total of 42 (65.6%) and 17 (26.6%) were based on general and specific population respectively. Most repositories were developed for research purposes (n=47, 73.4%). Further details are shown in Table 21 below.

Table 21: Distribution of Characteristics of Malaria Data Repositories

Characteristic	n (%) (N=64)
Category of data repository	
Metadata repository	3 (4.7)
Platform of aggregated data	36 (56.3)
Curated data	25 (39.1)
Type of data repository	
Census records	4 (6.3)
Claims database	1 (1.6)
Drug resistance database	1 (1.6)
Genomics	40 (62.5)
Medical records (electronic)	1 (1.6)
Medical records (electronic) & Genomics	2 (3.1)
Pathways database	1 (1.6)
Registry (electronic)	1 (1.6)
Surveillance System	12 (18.8)
Website	1 (1.6)
Setting	

CONFIDENTIAL

Hospital	0 (0.0)
Ambulatory	4 (6.3)
Mixed	9 (14.1)
Not applicable	39 (60.9)
Unknown	12 (18.8)
Population covered	
General population	42 (65.6)
Disease specific	17 (26.6)
Unknown	5 (7.8)
Purpose of data repository	
Research	47 (73.4)
Surveillance	15 (23.4)
Patient care	2 (3.1)

5.2.4.3 Availability of Essential Data for Malaria

A total of 15 (23.4%) repositories include genomics data. Patient profile data, such as age and gender, was only available in 23.4% (n=15) of repositories. A small proportion (10.9%, n=7) of repositories includes clinical information. Few repositories had information on diagnostic tests, laboratory exams, and pharmacological treatments. Table 22 presents further information.

Table 22: Distribution of Available Data on Malaria

Data	n (%) (N=64)
Genomics data	
Yes	15 (23.4)
No	40 (62.5)
Unknown	9 (14.1)
Patient profile	
Yes	15 (23.4)
No	40 (62.5)
Unknown	9 (14.1)
Clinical information	
Yes	7 (10.9)
No	43 (67.2)
Unknown	14 (21.9)
Diagnostic tests	
Yes	5 (7.8)

CONFIDENTIAL

No	44 (68.8)
Unknown	15 (23.4)
Laboratory data	
Yes	3 (4.7)
No	44 (68.8)
Unknown	17 (26.6)
Pharmacological treatments	
Yes	7 (10.9)
No	42 (65.6)
Unknown	15 (23.4)
Other treatments	
Yes	1 (1.6)
No	44 (68.8)
Unknown	19 (29.7)
Outcomes	
Yes	10 (15.6)
No	41 (64.1)
Unknown	13 (20.3)
Safety data	
Yes	3 (4.7)
No	44 (68.8)
Unknown	17 (26.6)

5.2.4.4 Data Governance, Curation and Sustainability

Most data repositories store data on their website (n=52, 81.3%). Data repositories were mostly private (n=27, 42.2%). A total of 25 (39.1%) repositories have a permanent public funding. No data repository provided information on a succession plan or a back-up and migration standards. More detailed information is shown in Table 23.

Table 23: Distribution of Governance, Curation and Sustainability Information of Malaria Data Repositories

Data governance, curation and sustainability	n (%) (N=64)
Infrastructure / Hosting location	
Website	52 (81.3)
Unknown	12 (18.8)
Ownership	

CONFIDENTIAL

University	15 (23.4)
Private	27 (42.2)
Government	17 (26.6)
Public-private	1 (1.6)
University/Private	1 (1.6)
Unknown	3 (4.7)
Funding	
Permanent public funding	25 (39.1)
Grants	9 (14.1)
Private initiative	7 (10.9)
Federal-State-Industry	1 (1.6)
Public-private funding	1 (1.6)
Unknown	21 (32.8)
Succession plan	
Not applicable	1 (1.6)
Unknown	63 (98.4)
Back-up and migration standards	
Unknown	64 (100.0)

5.2.4.5 Data Accessibility

The standard to ensure discoverability of databases was not obtained in any of the repositories, however, as these repositories were found either on literature search or grey literature, we can assume that all have a discoverability standard. A total of 51 (79.9%) data depositories have their data open for everybody. Linkage capabilities was available in 5 (7.8%) repositories. Table 24 provides further information.

Table 24: Distribution of Data Accessibility Information of Malaria Data Repositories

Data accessibility	n (%) (N=64)
Discoverability	
Unknown	64 (100.0)
Access policy	
Open access	49 (76.6)
On request	6 (9.4)
Restricted	5 (7.8)
Unknown	4 (6.3)
Who can access?	
Everybody	51 (79.7)
Designated research units	3 (4.7)

CONFIDENTIAL

External researchers	4 (6.3)
Unknown	6 (9.4)
Access procedure	
Authorization by government regulatory agencies	2 (3.1)
Data access committee	2 (3.1)
None	21 (32.8)
Unknown	16 (25.0)
Website registration	23 (35.9)
Linkage capacities	
Yes	5 (7.8)
No	6 (9.4)
Unknown	53 (82.8)

5.2.4.6 Data Management

Most of data repositories (n=31, 48.4%) had data entered by curators. The information is being coded in only 21.9% (n=14) data repositories. However, the information is unknown in 75.0% (n=48). Information on coding or standardization procedure was unknown in 85.9% (n=55) of data repositories. Only 6.3% (n=4) of repositories indicated the use of a pharmaceutical, disease or procedures coding. Quality control and data verification procedures are used in only 15.6% (n=10). However, this information was not available in 82.8% (n=53) of repositories. More than 45% offer support to users but only 23.4 to data depositors. However, the percentage of unknowns is high. Further details are presented in Table 25 below.

Table 25: Distribution of Data Management Information of Malaria Data Repositories

Data management	n (%) (N=64)
Data entry	
Researcher	2 (3.1)
Curator	31 (48.4)
Unknown	31 (48.4)
Coding / Data standardization	
Yes	14 (21.9)
No	2 (3.1)
Unknown	48 (75.0)
Coding or standardization procedure	

CONFIDENTIAL

Standardized data collection tools	7 (10.9)
No	2 (3.1)
Unknown	55 (85.9)
Pharmaceutical coding / Diseases coding / Procedures coding	
Yes	4 (6.3)
No	11 (17.2)
Unknown	49 (76.6)
Data verification / Quality control procedures	
Yes	10 (15.6)
No	1 (1.6)
Unknown	53 (82.8)
Frequency of update	
Real time	3 (4.7)
Daily	1 (1.6)
Weekly	1 (1.6)
Monthly	1 (1.6)
Two months	1 (1.6)
Yearly	1 (1.6)
Not applicable	1 (1.6)
Unknown	55 (85.9)
Availability of follow-up data	
Yes	3 (4.7)
No	3 (4.7)
Not applicable	1 (1.6)
Unknown	57 (89.1)
Support for data depositors	
Yes	15 (23.4)
No	2 (3.1)
Unknown	47 (73.4)
Support for data users	
Yes	29 (45.3)
No	2 (3.1)
Unknown	33 (51.6)

CONFIDENTIAL

5.2.4.7 Ethics

In order to access a database, it is required to get approval of an ethics committee for 3 (4.7%) repositories. More information on ethics is provided in Table 26.

Table 26: Distribution of Ethics Data of Malaria Data Repositories

Ethics data	n (%) (N=64)
Patient informed consent	
Yes	1 (1.6)
No	13 (20.3)
Not applicable	15 (23.4)
Unknown	35 (54.7)
Requires ethics committee approval	
Yes	3 (4.7)
No	13 (20.3)
Not applicable	15 (23.4)
Unknown	33 (51.6)
Standards for data anonymization	
Yes	10 (15.6)
No	0 (0.0)
Not applicable	15 (23.4)
Unknown	39 (60.9)

5.2.5 Tuberculosis

5.2.5.1 Coverage of Data Repositories

A total of 57 repositories included data for tuberculosis. The majority were developed in America (n=21, 36.8%), followed by Europe (n=15, 26.3%). Further details are presented in Table 27 below. Appendix 1.6 presents a list of selected data repositories involving tuberculosis.

Table 27: Data Repositories for Tuberculosis by Continent and Country

Continent	Country	n (%) (N=57)
Africa	South Africa	2 (3.5)
America	Argentina	1 (1.8)
	Brazil	2 (3.5)

CONFIDENTIAL

	Canada	3 (5.3)
	US	15 (26.3)
Asia	China	2 (3.5)
	India	6 (10.5)
	Japan	1 (1.8)
	Taiwan	1 (1.8)
Europe	France	2 (3.5)
	Germany	1 (1.8)
	Sweden	1 (1.8)
	Switzerland	1 (1.8)
	UK	10 (17.5)
Oceania	Australia	1 (1.8)
International	International	5 (8.8)
Unknown	Unknown	3 (5.3)

5.2.5.2 Characteristics of Tuberculosis Data Repositories

A total of 29 (50.9%) data repositories were considered as a platform of aggregated data. A total of 28 (49.1%) and 27 (47.4%) were based on general and specific population respectively. Most repositories were developed for research purposes (n=50, 87.7%). Further details are shown in Table 28 below.

Table 28: Distribution of Characteristics of Tuberculosis Data Repositories

Characteristic	n (%) (N=57)
Category of data repository	
Metadata repository	2 (3.5)
Platform of aggregated data	29 (50.9)
Curated data	26 (45.6)
Type of data repository	
Medical records (electronic)	2 (3.5)
Registry (electronic)	2 (3.5)
Surveillance System	2 (3.5)
Claims database	2 (3.5)
Study database	1 (1.8)
Pathways database	1 (1.8)
Drug database	1 (1.8)
Genomics	44 (77.2)
Chemical database	2 (3.5)

CONFIDENTIAL

Setting	
Mixed	5 (8.8)
Not applicable	52 (91.2)
Population covered	
General population	28 (49.1)
Disease specific	27 (47.4)
Unknown	2 (3.5)
Purpose of data repository	
Research	50 (87.7)
Surveillance	6 (10.5)
Patient care	1 (1.8)

5.2.5.3 Availability of Essential Data for Tuberculosis

A total of 9 (15.8%) repositories include genomics data. Patient profile data was available in only 15.8% (n=9) of repositories. A small proportion (8.8%, n=5) of repositories includes clinical information. Few repositories had information on diagnostic tests, laboratory exams, and pharmacological treatments. Table 29 presents further information.

Table 29: Distribution of Available Data on Tuberculosis

Data	n (%) (N=57)
Genomics data	
Yes	9 (15.8)
No	43 (75.4)
Unknown	5 (8.8)
Patient profile	
Yes	9 (15.8)
No	43 (75.4)
Unknown	5 (8.8)
Clinical information	
Yes	5 (8.8)
No	44 (77.2)
Unknown	8 (14.0)
Screening / Diagnostic tests	
Yes	5 (8.8)
No	45 (78.9)
Unknown	7 (12.3)

CONFIDENTIAL

Laboratory data	
Yes	4 (7.0)
No	45 (78.9)
Unknown	8 (14.1)
Pharmacological treatments	
Yes	11 (19.3)
No	35 (61.4)
Unknown	11 (19.3)
Other treatments	
No	45 (78.9)
Unknown	12 (21.1)
Outcomes	
Yes	7 (12.3)
No	45 (78.9)
Unknown	5 (8.8)
Safety data	
Yes	1 (1.8)
No	44 (77.2)
Unknown	12 (21.1)

5.2.5.4 Data Governance, Curation and Sustainability

Most data repositories store data on their website (n=50, 87.7%). Data repositories were mostly private (n=22, 38.6%) or owned by universities (n=20, 35.1%). A total of 15 (26.3%) repositories come from private initiatives. More detailed information is shown in Table 30.

Table 30: Distribution of Governance, Curation and Sustainability Information of Tuberculosis Data Repositories

Data governance, curation and sustainability	n (%) (N=57)
Infrastructure / Hosting location	
Website	50 (87.7)
Local servers	1 (1.8)
Unknown	6 (10.5)
Ownership	
University	20 (35.1)
Private	22 (38.6)
Government	8 (14.0)

CONFIDENTIAL

Unknown	7 (12.3)
Funding	
Permanent public funding	10 (17.5)
Grants	13 (22.8)
Private initiative	15 (26.3)
Researcher fees	1 (1.8)
Unknown	18 (31.6)
Succession plan	
Yes	1 (1.8)
Unknown	56 (98.2)
Back-up and migration standards	
Yes	2 (3.5)
Unknown	55 (96.5)

5.2.5.5 Data Accessibility

The standard to ensure discoverability of databases was not obtained in any of the repositories, however, as these repositories were found either on literature search or grey literature, we can assume that all have a discoverability standard. A total of 45 (78.9%) data depositories have their data open for everybody. Linkage capabilities was available in only 2 repositories. Table 31 provides further information.

Table 31: Distribution of Data Accessibility Information of Tuberculosis Data Repositories

Data accessibility	n (%) (N=57)
Discoverability	
Unknown	57 (100.0)
Access policy	
Open access	45 (78.9)
On request	6 (10.5)
Restricted	2 (3.5)
Unknown	4 (7.0)
Who can access?	
Everybody	43 (75.4)
Designated research units	2 (3.5)
External researchers	3 (5.3)
External researchers (fees applicable)	3 (5.3)
Unknown	6 (10.5)
Access procedure	

CONFIDENTIAL

Data access committee	3 (5.3)
Authorization by Government Regulatory Agencies	3 (5.3)
Website registration	4 (7.0)
None	39 (68.4)
Unknown	8 (14.0)
Linkage capacities	
Yes	2 (3.5)
Unknown	55 (96.5)

5.2.5.6 Data Management

Most of data repositories (n=48, 84.2%) had data entered by curators. The information is being coded in only 12.3% (n=7) data repositories. However, the information is unknown in 87.7% (n=50). Information on coding or standardization procedure was unknown in 91.2% (n=52) of data repositories. Quality control and data verification procedures are used in only 17.5% (n=10). However, this information was not available in 82.5% (n=47) of repositories. More than 59% offer support to users but only 21.1 to data depositors. However, the percentage of unknowns is high. Further details are presented in Table 32 below.

Table 32: Distribution of Data Management Information of Tuberculosis Data Repositories

Data management	n (%) (N=57)
Data entry	
Researcher	6 (10.5)
Curator	48 (84.2)
Unknown	3 (5.3)
Coding / Data standardization	
Yes	7 (12.3)
Unknown	50 (87.7)
Coding or standardization procedure	
Standardized data collection tools	4 (7.0)
CDISC SDTM	1 (1.8)
Unknown	52 (91.2)
Pharmaceutical coding / Diseases coding / Procedures coding	
Yes	4 (7.0)

CONFIDENTIAL

No	36 (63.2)
Unknown	17 (29.8)
Data verification / Quality control procedures	
Yes	10 (17.5)
Unknown	47 (82.5)
Frequency of update	
Real time	2 (3.5)
Daily	2 (3.5)
Weekly	1 (1.8)
Monthly	1 (1.8)
Every 2 months	1 (1.8)
Every 3-4 months	1 (1.8)
Every 6 months	1 (1.8)
Project ended	1 (1.8)
Unknown	47 (82.5)
Availability of follow-up data	
No	2 (3.5)
Unknown	55 (96.5)
Support for data depositors	
Yes	12 (21.1)
No	1 (1.8)
Unknown	44 (77.2)
Support for data users	
Yes	34 (59.6)
No	1 (1.8)
Unknown	22 (38.6)

5.2.5.7 Ethics

In order to access a database, it is required to get approval of an ethics committee for 4 (7.0%) repositories only. More information on ethics is provided in Table 33.

Table 33: Distribution of Ethics Data of Tuberculosis Data Repositories

Ethics data	n (%) (N=57)
Patient informed consent	
No	48 (84.2)
Unknown	9 (15.8)

CONFIDENTIAL

Requires ethics committee approval	
Yes	4 (7.0)
No	45 (78.9)
Unknown	8 (14.0)
Standards for data anonymization	
Yes	6 (10.5)
No	45 (78.9)
Unknown	6 (10.5)

6 Study Strengths and Limitations

Although these literature searches followed the Cochrane Group recommendations for systematic literature reviews and terms used were broad, the results obtained are limited by the keywords used on the search strategies. Therefore, not all data repositories available for the study of the diseases of interest might have been found.

7 Conclusion

Following this descriptive study, several repositories were described. Identification of data repositories is crucial to develop agreements and to harmonise data in order help data input, sharing, analysis and reuse.

Repositories distribution varies according to the disease of interest. The countries with the most number of data repositories for the four diseases of interest are the US followed by the UK. Most of the data repositories included in this study include aggregate data, which is crucial for planning and guidance of the performance of health systems. However, aggregate data cannot provide the type of detailed information which patient level data can [1]. Mostly, data repositories were owned by a private entity followed by universities and governments. In most cases, data is hosted on websites. Web-based data repositories ease data sharing as its content is available to anyone with internet access.

Most of data repositories were created with the purpose of research, which the majority have an open access policy and just a few are restricted and required authorization for the use of data. Open access eliminates the economic and physical barriers that stop access to research data and improves the way researchers conduct and share research [2].

CONFIDENTIAL

8 References

List of retained articles is included in Appendix 1.3 to Appendix 1.6 under the tab called "References".

[1] District Health Information Software (DHIS), 23 November 2016. [En ligne]. Available: https://docs.dhis2.org/2.22/en/user/html/ch01s05.html.

[2] Lwoga ET, Questier F, «Open access behaviours and perceptions of health sciences faculty and roles of information professionals,» *Health Information & Libraries Journal,* vol. 32, n° %11, pp. 37-49, 1 March 2015.

CONFIDENTIAL

Appendix 1.1: Literature Search Strategy

See Word document attached.

CONFIDENTIAL

Appendix 1.2: Additional Literature Search Strategy

See Word document attached.

CONFIDENTIAL

Appendix 1.3: Selected Data Repositories Involving Dengue

See Excel spreadsheet attached.

CONFIDENTIAL

Appendix 1.4: Selected Data Repositories Involving Leprosy

See Excel spreadsheet attached.

CONFIDENTIAL

Appendix 1.5: Selected Data Repositories Involving Malaria

See Excel spreadsheet attached.

CONFIDENTIAL

Appendix 1.6: Selected Data Repositories Involving Tuberculosis

See Excel spreadsheet attached.

CONFIDENTIAL