Development of Standards for Online Repositories

Final Report, Version 2.0

12 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

Table of Contents

List of Tables

List of Figures

Abbreviations

ADaM	Analysis Data Model
AIP	Archival Information Package
BC	British Columbia
C-Path	Critical Path Institute
CC0	Creative Commons Zero
CDISC	Clinical Data Interchange Standards Consortium
CDM	Common Data Model
CODR	Critical Path Institute Online Data Repository
CRF	Case report form
CSDR	ClinicalStudyDataRequest.com
DOI	Digital Object Identifier
EMR	Electronic medical record
GB	Gigabyte
H3Africa	Human Heredity and Health in Africa
ICSU	International Council of Scientific Unions
ICTRP	International Clinical Trials Registry Platform
IDDO	Infectious Diseases Data Observatory
IP	Intellectual property
ISARIC	International Severe Acute Respiratory Emerging Infection Consortium
IPD	Individual participant-level data
ISO	International Organization for Standardization
LSC	Life Sciences Consortium
MSF	Médecins sans frontières (Doctors Without Borders)

NCBI	National Center for Biotechnology Information
NDA	National Institute of Mental Health Data Archive
NIH	National Institute of Health
OHDSI	Observational Health Data Sciences and Informatics
PDS	Project Data Sphere
PopData	Population Data BC
QA	Quality assurance
QRS	Questionnaires, rating and scales
RCSB	Research Collaboratory for Structural Bioinformatics
SEND	Standard for Exchange of Nonclinical Data
SDTM	Study Data Tabulation Model
SOP	Standard operating procedure
TB	Tuberculosis
TB-PACTS	Platform for Aggregation of Clinical Tuberculosis Studies
TRAC	Trustworthy Repositories Audit & Certification
TRDS	Trial Registration Data Set
US	United States
WHO	World Health Organization
wwPDB	Worldwide Protein Data Bank
YODA	Yale University Open Data Access

1 Introduction

1.1 Background

Data accessibility, sharing, and reuse are essential for the timely translation of research results into knowledge and best practices, including policies, for improving global health. Many public (e.g., National Institute of Health [NIH]) and private organisations (e.g., Project Data Sphere) have developed governance principles and processes regarding research data accessibility and sharing. One of the overarching contributions of data sharing is the generation of new discovery that no single study or organisation could provide on its own.

Data sharing involves the creation of a secure platform where patient-level data obtained from multiple studies or sources are 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.

Usefulness of data sharing resides, amongst others, in the study of the natural history of diseases or conditions, patterns of treatments, risk factors, quality and availability of or gaps in healthcare provided, 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 (e.g., electronic medical records (EMRs)), registries (disease, drug, pregnancy registries), or as part of the administration of health care programs (e.g., administrative claims database). While the breadth of data has expanded, resulting heterogeneity in sources is associated with a lack of standardisation in the types of data collected. For example, claims databases are transactional databases that collect data on diagnostic codes (e.g., International Classification of Diseases) or drug dispensings, while clinical databases may 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., secondary use of existing data that have been collected for other purposes). Typically, primary data collection requires a case report form (CRF) and individual patient informed consent, while 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 heterogeneous data. The first is the pooling of raw data, which requires a standardised and common platform across all sources. EMR systems typically follow this model, whereby data entry in the common platform is conducted in real-time at the individual sites.

The second model is referred to as a Common Data model (CDM), whereby, at individual sites, data are extracted, transformed and loaded into a common platform. Analysis is conducted using the pooled data set. Examples of CDM include Observational Health Data Sciences and Informatics (OHDSI) or Sentinel in the United States (US). The third model consists of a distributed network whereby patient-level data remain at the individual sites and are analysed using a common protocol. Results are then sent to the coordinating site for pooling, using mainly 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.

1.2 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 US address predominantly privacy and security concerns, while those of 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. Standardisation may therefore be implemented at the time of data collection or data transmission to the repository.

Repository designs consist of web interfaces that are based on a robust security framework including role-based data access, data encryption, and digital certification. Authorised users are able to input data directly into the central repository. Alternatively, for centres that 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 described in the literature, a major barrier to implementation of Databases of Prescription Monitoring Programs for Opioids in the US has been the difficulty of access by health care providers (Moride Y., et al., 2017). Increased end-user involvement in the design of the repository has been shown to increase its usability.

Governance and access policies for internal and external researchers are fundamental to achieve the overarching aim data repositories, which is the sharing of data. Anonymization of personal data through de-identification is fundamental for data sharing.

To support the development of standards for trusted data repositories, a review of best practices and governance requirements used by existing data repositories was undertaken.

2 Objectives

2.1 Main Objective

To review existing standards, best practices, and governance requirements to establish and run trusted repositories that house health research data.

2.2 Specific Objectives

1) To identify existing standards related to data repositories;
2) To assess the standards that are currently in place in selected repositories that house health data;
3) To conduct a gap analysis of governance standards used in existing repositories.

3 Methods

3.1 Search and Review of Repository Standards

A pragmatic search of web sources was conducted in order to identify existing repository standards. We utilised the key words “repository standard”, “data repository”, “data repositories”, “research consortium”, “data sharing consortium”.

We retained only repository standards that cover the components shown in Table 1.

Table 1: Repository Standards Selection Criteria

Components
Data curation (metadata standards, discoverability, degree of data standardisation, data verification & quality assurance [QA] procedures)
Security and longevity (encryption, access control and other security measures, storage, backup, migration standards, sustainability of funding, wind-down plan),
Governance (legal status of the repository, terms of use, data access, user licences, benefit sharing and intellectual property [IP] issues, hosting location, and audit procedures)

Review consisted of determining whether each component listed in Table 1 was addressed in the standards and of documenting the available components.

3.2 Review of Data Repositories

We first identified repositories from a systematic review conducted in the first phase of the project as well as from a pragmatic search. Pragmatic search of Web sources was conducted using Google and Google Scholar search engines, for four diseases of interest (malaria, tuberculosis, dengue, leprosy). Repositories that were either datasets, a single database, or tools were not of interest for this component of the project and were therefore excluded. In addition, we also excluded repositories that were already covered by Ternyata Ltd or by organisations that owned multiple data repositories.

For the pragmatic search, key words such as “research data repository”, “data repository” AND “clinical research”, “research consortium”, “data sharing consortium” were used in order to identify additional repositories related to general populations, or to specific diseases. Only repositories that mention in their website at least one of the components of interest (listed in Table 1) were retained. Repositories identified by the project lead (Elizabeth Pisani, from Ternyata Ltd) were also included.

For each repository, information on the following elements was obtained, using publicly available sources (mainly website or related documents available on the web): Mission, data type, population covered, repository type, governance (who governs, ownership, funding), data accessibility (discoverability, intellectual property, access policy, who has access, access policy, terms and license for reuse), and fees.

4 Results

4.1 Search Results - Repository Standards

Search strategy yielded a total of 11 repository standards, out of which 2 were retained for the review because they covered either data curation, security and longevity, or governance. In addition, 3 other standards previously identified by Elizabeth Pisani, from Ternyata Ltd, were also retained, yielding a total of 5 repository standards, listed below:

1. Trustworthy Repositories Audit & Certification (TRAC)
2. World Health Organization (WHO) – International Standards for Clinical Trial Registries

3. ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories
4. ICSU World Data System
5. The Human Heredity and Health in Africa (H3Africa) Initiative

For several components, TRAC and ISO 16363 are nearly identical, therefore one can assume that TRAC based their standards on ISO 16363. In addition to these standards, we reviewed in-depth the Clinical Data Interchange Standards Consortium (CDISC), which develops and supports global, platform-independent data standards. For this reason, CDISC standards cover metadata and data standardisation only.

4.1.1 Overview of Data Repository Standards

Each of the elements of interest addressed in the 6 (5 repository standards and CDISC) standards are described in the following sections and summarised in Table 2 below.

Table 2: Elements of Interest Addressed by the Data Repository Standards

	TRAC	WHO	ISO 16363	ICSU	H3Africa	CDISC
Metadata Standards	+	-	+	-	+	+
Discoverability	-	-	+	+	+	-
Data Standardisation	-	-	-	-	-	+
Data Verification and Quality Assurance Procedures	+	+	-	+	+	-
Access Control & Other Security Measures	+	-	+	+	+	-
Storage Standards	+	-	-	+	+	-
Data Backup Standards	+	+	+	-	-	-
Data Migration Standards	+	-	+	-	-	-
Sustainability of Funding Standards	+	+	+	+	-	-
Data Preservation Standards	-	-	+	+	-	-
Succession Plan Standards	+	+	+	-	-	-
Legal Status	+	+	+	-	-	-
Data Access Standards	+	+	+	+	+	-
Data Sharing & Intellectual Property Issues	+	-	+	-	+	-
Audit Procedures	+	+	+	-	-	-

+ Stated in publicly available information
- Information was not mentioned in the publicly available information
TRAC Trustworthy Repositories Audit & Certification
WHO World Health Organization – International Standards for Clinical Trial Registries
ISO 16363 Space data and information transfer systems — Audit and certification of trustworthy digital repositories
CDISC Clinical Data Interchange Standards Consortium
ICSU International Council of Scientific Unions World Data System

The data curation standards that included the most elements of interest were the TRAC and ISO 16363. Both of these standards often complement each other, except for the item related to Data Standardisation which is only addressed by CDISC. A detailed description of standards is provided in the following sections.

4.1.2 Description of Data Curation Standards

4.1.2.1 Metadata Standards

An Archival Information Package (AIP) is the set of content and metadata managed by a preservation repository, and organized in a way that allows the repository to perform preservation services (Society of American Archivists, 2012). Four of the five standards provide information on how metadata should be handled. TRAC and ISO 16363 are aligned in their standards, while the others tend to diverge. H3Africa recommends the use of recognised data standards and metadata formats, however they do not specify which standards should be used. A description of each of the metadata standards is provided in Table 3.

Table 3: Description of Metadata Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository articulates minimum metadata requirements to enable the designated community(ies) to discover and identify material of interest.
	Repository captures or creates minimum descriptive metadata and ensures that it is associated with the archived object (i.e., AIP).
	Repository can demonstrate that referential integrity is created between all archived objects (i.e., archival information packages (AIPs)) and associated descriptive information.
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall capture or create minimum descriptive information and ensure that it is associated with the AIP. Minimum descriptive information that was either received from the producer or created by the repository should be shown.
	The repository shall maintain bi-directional linkage between each AIP and its descriptive information.
Clinical Data Interchange Standards Consortium (CDISC)	Analysis Data Model (ADaM) defines dataset and metadata standards that support: 1. efficient generation, replication, and review of clinical trial statistical analyses, and 2. traceability between analysis results, analysis data, and data represented in the Study Data Tabulation Model (SDTM).
H3Africa (Human Heredity and Health in Africa)	Recognised data standards and metadata formats should be used wherever possible.

4.1.2.2 Discoverability Standards

ISO 16363, ICSU, and H3Africa provide standards on discoverability, which consists in assigning a unique and permanent identifier (see Table 4).

Table 4: Description of Discoverability Standards

Standard Name	Description
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall have and use a convention that generates persistent, unique identifiers for all AIPs.
	The repository shall have a system of reliable linking/ resolution services in order to find the uniquely identified object, regardless of its physical location.
ICSU World Data System	The repository enables users to discover the data and refer to them in a persistent way through proper citation.
H3Africa (Human Heredity and Health in Africa)	Research networks and programs should establish clear and transparent policies based on these agreed mechanisms, which should be readily discoverable by potential users.

4.1.2.3 Data Standardisation

CDISC provides standards on clinical and nonclinical data, through STDM, SEND (Standard for Exchange of Nonclinical Data), and ADaM (Analysis Data Model), as described in Table 5.

Table 5: Description of Standards for Data Standardisation

Standard Name	Description
Clinical Data Interchange Standards Consortium (CDISC)	SDTM provides a standard for organising and formatting data to streamline processes in collection, management, analysis and reporting.
	Standard for Exchange of Nonclinical Data (SEND) is an implementation of the SDTM standard for nonclinical studies. SEND specifies a way to collect and present nonclinical data in a consistent format.
	ADaM defines dataset and metadata standards that support: 1. efficient generation, replication, and review of clinical trial statistical analyses, and 2. traceability between analysis results, analysis data, and data represented in the Study Data Tabulation Model (SDTM).
	Questionnaires, rating and scales (QRS) - Each QRS instrument is a series of questions, tasks or assessments used in clinical research to provide a qualitative or quantitative assessment of a clinical concept or task-based observation.

4.1.2.4 Data Verification and Quality Assurance Procedures

TRAC, WHO, ICSU, and H3Africa provide standards on data verification and quality assurance (QA). In summary, procedures, either automated or manual, should be in place to assure accuracy of data. Such procedures are further described in Table 6.

Table 6: Description of Data Verification and Quality Assurance Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository has procedures and policies in place, and mechanisms for their review, update, and development as the repository grows and as technology and community practice evolve.
	Repository has policies and procedures to ensure that feedback from producers and users is sought and addressed over time.
World Health Organization (WHO) - International Standards for Clinical Trial Registries	Registry staff must routinely check all data submitted about a trial for completeness and meaningfulness to ensure that all trial registry data set (TRDS) fields are populated and comply with the minimum standards.
	Registry database systems must apply automated checking procedures (e.g. range checks, logic rules) to data items to facilitate validity checking.
	Registries must undertake regular internal quality control audits to assess the level of completeness and accuracy of the data collected.
	Registries must have Standard Operating Procedures (SOP).
ICSU World Data System	The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
H3Africa (Human Heredity and Health in Africa)	Any data that are shared must be of an appropriate quality and in a format that enables them to be used by others, with sufficient metadata provided.

4.1.3 Security and Longevity Standards

4.1.3.1 Access Control & Other Security Measures

TRAC, ISO 16363, ICSU, and H3Africa provide standards on security and longevity of repositories. Amongst others, access failures should be reviewed and risks should be assessed. Features of access control and security measures are further described in Table 7.

Table 7: Description of Standards for Access Control & Other Security Measures

Standard Name	Description

Trustworthy Repositories Audit & Certification (TRAC)	Repository has implemented a policy for recording all access actions (includes requests, orders, etc.) that meet the requirements of the repository and information producers/depositors.
	Repository ensures that agreements applicable to access conditions are adhered to.
	Repository access management system fully implements access policy.
	Repository logs all access management failures, and staff review inappropriate "access denial" incidents.
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall log and review all access management failures and anomalies.
	The repository shall have a process to record and react to the availability of new security updates based on a risk-benefit assessment.
	The repository shall maintain a systematic analysis of security risk factors associated with data, systems, personnel, and physical plant.
ICSU World Data System	The repository should analyse potential threats, assess risks, and create a consistent security system. It should describe damage scenarios based on malicious actions, human error, or technical failure that pose a threat to the repository and its data, products, services, and users. It should measure the likelihood and impact of such scenarios, decide which risk levels are acceptable, and determine which measures should be taken to counter the threats to the repository and its Designated Community.
H3Africa (Human Heredity and Health in Africa)	The processes through which users can obtain access to the data should be proportionate, transparent and not unduly delay requests from legitimate users. It may be appropriate to apply different access processes for different data types.
	Data must be held and shared in a safe and secure manner that provides adequate protections for research participants, and in a way that is fully consistent with the consent obtained from research participants.

4.1.3.2 Storage

Only TRAC, ICSU, and H3Africa provide standards on data storage, as described in Table 8.

Table 8: Description of Storage Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository implements/responds to strategies for archival object (i.e., AIP) storage.
	Repository has contemporaneous records of actions and

	administration processes that are relevant to preservation (Archival Storage).
	Repository has defined processes for storage media.
ICSU World Data System	The repository applies documented processes and procedures in managing archival storage of the data. Repositories need to store data and metadata from the point of deposit, through the ingest process, to the point of access. Repositories with a preservation remit must also offer 'archival storage' in OAIS terms.
H3Africa (Human Heredity and Health in Africa)	Genomic data should be deposited in existing community data repositories wherever possible.

4.1.3.3 Data Backup

All 3 main repository standards provide similar standards on data backup, as shown in Table 9. The other 2 standards (CDISC and ICSU) do not address data backup.

Table 9: Description of Data Backup Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository ensures that it has adequate hardware and software support for backup functionality sufficient for the repository's services and for the data held.
	Repository manages the number and location of copies of all digital objects.
World Health Organization (WHO) - International Standards for Clinical Trial Registries	Registries must have documented procedures for ensuring adequate data security and other provisions to prevent data corruption and loss. This will include regular database replication and/or back-up (minimum 500 GB data backup capability).
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall have adequate hardware and software support for backup functionality sufficient for preserving the repository content and tracking repository functions. Simple backup mechanisms must preserve not only the repository main content, but also the system metadata generated by the preservation functions. Repositories need to develop backup plans that ensure their continuity of operations across all failure modes.
	The repository shall have effective mechanisms to detect bit corruption or loss. The objective is a comprehensive treatment of the sources of data loss and their real-world complexity. Any data or metadata that is (temporarily) lost should be recoverable from backups.
	The repository shall have suitable written disaster preparedness and recovery plan(s), including at least one off-site backup of all preserved information together with an

	offsite copy of the recovery plan(s).

4.1.3.4 Data Migration

Only TRAC and ISO 16363 provide migration standards; both being similar as shown in Table 10.

Table 10: Description of Data Migration Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository implements/responds to strategies for archival object (i.e., AIP) migration.
	Repository has defined processes for hardware change (e.g., refreshing, migration).
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall have defined processes for storage media and/or hardware change (e.g., refreshing, migration). This is necessary in order to ensure that data is not lost when either the media fail or the supporting hardware can no longer be used to access the data.

4.1.3.5 Sustainability of Funding

Four repository standards provide standards on sustainability of funding. Overall, all data repositories should maintain business plans to sustain the repository. Some, such as the WHO, is less directive than the others to ensure monitoring of funding and to bridge gaps (detailed in Table 11).

Table 11: Description of Sustainability of Funding Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository has short- and long-term business planning processes in place to sustain the repository over time.
	Repository has in place processes to review and adjust business plans at least annually.
	Repository has ongoing commitment to analyse and report on risk, benefit, investment, and expenditure (including assets, licenses, and liabilities).
	Repository commits to monitoring for and bridging gaps in funding.
World Health Organization (WHO) - International Standards for Clinical Trial Registries	Registries must have a documented business plan that addresses the strategies the registry has in place to address its medium to long term sustainability.
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall have short- and long-term business planning processes in place to sustain the repository over time.
	The repository shall have an ongoing commitment to analyse and report on financial risk, benefit, investment, and

	expenditure (including assets, licenses, and liabilities).
ICSU World Data System	The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

4.1.3.6 Data Preservation

ISO 16363 and ICSU provides standards on data preservation, described in Table 12.

Table 12: Description of Data Preservation Standards

Standard Name	Description
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall have a mission statement that reflects a commitment to the preservation of, long term retention of, management of, and access to digital information.
	The repository shall have a Preservation Strategic Plan that defines the approach the repository will take in the long-term support of its mission.
	The repository shall have Preservation Policies in place to ensure its Preservation Strategic Plan will be met.
	The repository shall have documented preservation strategies relevant to its holdings. These preservation strategies and the preservation strategic plan typically address the degradation of storage media, the obsolescence of media drives, and the obsolescence or inadequacy of Representation Information (including formats) as the knowledge base of the Designated Community changes, and safeguards against accidental or intentional digital corruption.
ICSU World Data System	Repositories take responsibility for stewardship of digital objects, and for ensuring that materials are held in the appropriate environment for appropriate periods of time. Depositors and users must be clear that preservation of, and continued access to, the data is an explicit role of the repository.
	The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

4.1.3.7 Succession Plan (Wind-down plan)

Three of the five repository standards provide standards on succession plan, described in Table 13.

Table 13: Description of Standards for Succession Plan

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository has an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place.
	Repository has mechanisms in place for monitoring and notification when Representation Information (including formats) approaches obsolescence or is no longer viable.
World Health Organization (WHO) - International Standards for Clinical	Should a registry cease to function, the registry will transfer at least the WHO TRDS (original and updated) for all trial records

Trial Registries	to another Primary Registry in the WHO Registry Network.
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall have an appropriate succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.
	The repository shall monitor its organizational environment to determine when to execute its succession plan, contingency plans, and/or escrow arrangements.

4.1.4 Governance Standards

4.1.4.1 Legal Status of Repository

Three of the five repository standards provide standards on legal aspects of repositories. However, as shown in Table 14, the WHO standards are not as specific as the others.

Table 14: Description of Legal Status Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository maintains written policies that specify the nature of any legal permissions required to preserve digital content over time, and repository can demonstrate that these permissions have been acquired when needed.
World Health Organization (WHO) - International Standards for Clinical Trial Registries	The Registry will publicly disclose ownership, governance structure and not-for-profit status.
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall have and maintain appropriate contracts or deposit agreements for digital materials that it manages, preserves, and/or to which it provides access.
	The repository shall have contracts or deposit agreements which specify and transfer all necessary preservation rights, and those rights transferred shall be documented.

4.1.4.2 Data Access

Data access standards are described for 5 of the repository standards. As shown in Table 15, the standards should specify who and under which conditions access is allowed.

Table 15: Description of Data Access Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository documents and communicates to its designated community(ies) what access and delivery options are available.
World Health Organization (WHO) - International Standards for Clinical Trial Registries	Registries must make the WHO TRDS items for all studies in their register (i.e., the registry database) accessible online at no charge to the end user.
	Registries must enable online electronic searches of text words and phrases via a simple, single search box. As a minimum, it must be possible to search data in both the condition and intervention fields.
	Access to the register (i.e., the registry's database) to search for registered trials will be available 24 hours a day, 7 days a week, subject to a reasonable minimal period of planned downtime for routine maintenance requirements.
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall comply with Access Policies. Depending on the nature of the repository, the Access Policies may cover statements of what is accessible to which community, and on what conditions.
ICSU World Data System	Repositories must maintain all applicable licenses covering data access and use, communicate about them with users, and monitor compliance.
	Repositories must ensure that data can be understood and used effectively into the future despite changes in technology. This Requirement evaluates the measures taken to ensure that data are reusable.
H3Africa (Human Heredity and Health in Africa)	The processes through which users can obtain access to the data should be proportionate, transparent and not unduly delay requests from legitimate users. It may be appropriate to apply different access processes for different data types.

4.1.4.3 Benefit Sharing & Intellectual Property Issues

As described in Table 16, TRAC and ISO 16363 provide standard on intellectual property (IP) rights.

Table 16: Description of Standards for Data Sharing & Intellectual Property Issues

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository tracks and manages intellectual property rights and restrictions on use of repository content as required by deposit agreement, contract, or license.
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall track and manage intellectual property rights and restrictions on use of repository content as required by deposit agreement, contract, or license.
H3Africa (Human Heredity and Health in Africa)	In line with Wellcome Trust and NIH policies, intellectual property should be developed and used in a way that maximizes global health benefit.
	Research Networks and programs should seek to manage intellectual property, and develop appropriate licensing terms, in ways that help to ensure equitable access to resulting health products and technologies for low and middle-income countries.
	Where appropriate, Research Networks and programs should establish mechanisms to ensure the equitable sharing of benefits with the communities participating in the research – this might take the form of contributions to capacity and skills development, or the specification of appropriate licensing terms that ensure access to healthcare benefits derived from the research.

4.1.4.4 Audit Procedures

WHO is more specific than TRAC and ISO 16363 in terms of audit procedures, as it provides standards on self-audit and site visits (Table 17).

Table 17: Description of Audit Procedures Standards

Standard Name	Description
Trustworthy Repositories Audit & Certification (TRAC)	Repository provides an independent mechanism for audit of the integrity of the repository collection/content.
World Health Organization (WHO) - International Standards for Clinical Trial Registries	*Self-audit:* Registries conduct their own, internal audits to determine if processes and procedures are being complied with, and adjustments made if necessary
	Self-report: Registries will be asked to update their Registry Profile on an annual basis and return it to the International Clinical Trials Registry Platform (ICTRP) Secretariat.
	Site visit: A small audit team will visit a registry and examine

	all processes and procedures. These audits will have structured programmes (yet to be developed) that will include an element of peer review (that is, an Administrator from a Primary Registry in the WHO Registry Network will be part of each audit team).
ISO 16363: Space data and information transfer systems — Audit and certification of trustworthy digital repositories	The repository shall commit to a regular schedule of self-assessment and external certification.

4.2 Search Results - Data Repositories

Systematic review and pragmatic search led to the identification of 161 existing data repositories out of which the following 17 were retained:

1. Population Data British Columbia (BC)
2. Platform for Aggregation of Clinical TB Studies (TB-PACTS)
3. Reactome
4. Worldwide Protein Data Bank (wwPDB)
5. Infectious Diseases Data Observatory (IDDO)
6. Mendeley Data
7. International Severe Acute Respiratory Emerging Infection Consortium (ISARIC)
8. Harvard Dataverse
9. Dryad
10. Project Data Sphere
11. National Center for Biotechnology Information (NCBI)
12. National Institute of Mental Health Data Archive (NDA)
13. ClinicalStudyDataRequest.com (CSDR)
14. YODA Project
15. Médecins sans frontières data sharing policy
16. Figshare
17. Vivli

Repositories 1 to 5 were identified from the previous phase of the project (the search on malaria, tuberculosis, dengue, and leprosy diseases). Repositories 6 to 9 were identified from the report “Data sharing in public health emergencies: What we want, what we have, what we need” presented by

Ternyata Ltd for GloPID-R. Repositories 10 to 12 were identified from a Google search, while repositories 13 to 17 were suggested by Elizabeth Pisani, Director of Ternyata Ltd. A selection flow chart is shown in Figure 1 bellow.

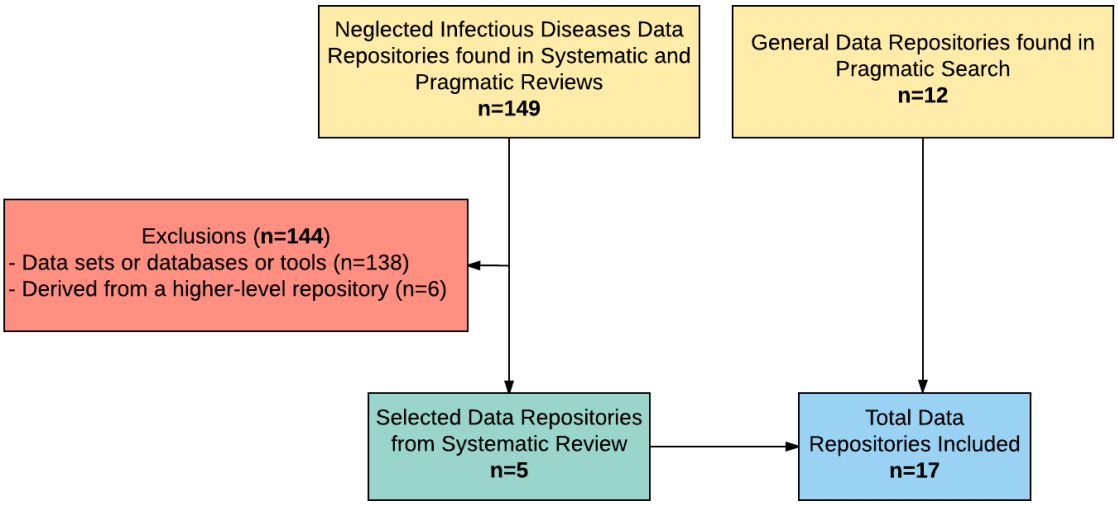

Figure 1: Data Repository Flow Chart

Repositories that were either datasets, a single database, or tools were of not relevant for this study and were therefore excluded. We also excluded repositories that were already covered by Ternyata Ltd (Malariagen, MRCT Center, Harvard Brigham, EMBL-EBI, EBI/ELIXIR, INDEPTH, Biobank, ALPHA, Agincourt, APHRC, Janus, LSHTM, UKDA, DANS, Zenodo, ICTRP, re3data, OpenTrials, Datacite) or under organisations that owned multiple data repositories.

4.2.1 Overview of Standards in existing data repositories

A summary of the searched information from the 17 data repositories and MSF data sharing policy (later called *repository* for convenience) is provided in Table 18 and Table 19 below.

Table 18: Components of Standards Found in Data Repositories (Part 1)

Repository	Mission Statement	Population Covered	Who Governs	Ownership	Funding	Discovera-bility
PopData	+	+	+	+	+	-
TB-PACTS	+	+	+	+	+	-
Reactome	-	-	+	+	-	+
wwPDB	+	-	+	-	+	-
IDDO	-	+	+	+	-	-
Mendeley Data	+	-	-	+	-	+
ISARIC	-	+	+	+	-	-
Harvard Dataverse	-	-	-	-	+	+
Dryad	+	+	+	-	+	+
PDS	-	+	+	+	+	-
NCBI	+	+	+	+	-	-
NDA	-	+	+	+	+	+
CSDR	-	+	-	-	-	-
YODA	+	+	+	+	-	-
MSF	-	+	+	-	+	-
Figshare	+	-	-	+	-	+
Vivli	+	+	+	-	+	+

+ Stated in publicly available information
– Information was not mentioned in the publicly available information
PopData Population Data of British Columbia
PDS Project Data Sphere
NCBI National Center for Biotechnology Information
NDA National Institute of Mental Health Data Archive
TB-PACTS Platform for Aggregation of Clinical Tuberculosis Studies
CSDR ClinicalStudyDataRequest.com
YODA The Yale University Open Data Access Project
ISARIC International Severe Acute Respiratory and Emerging Infection Consortium
wwPDB Worldwide Protein Data Bank
MSF Médecins sans frontières (Doctors Without Borders)

Table 19: Components of Standards Found in Data Repositories (Part 2)

Repository	Intellectual Property	Fees[a]	Access Policy	Who can access?	Access Procedure	Terms of Reuse	Licenses for Reuse
PopData	-	+	+	+	+	-	-
TB-PACTS	-	-	-	+	+	+	-
Reactome	+	-	-	+	-	-	+
wwPDB	+	-	-	+	-	-	-
IDDO	-	-[b]	-	+	+	-	-
Mendeley Data	+	-	-	+	+	-	+
ISARIC	-	-	+	+	+	-	+
Harvard Dataverse	-	-	-	+	+	-	+
Dryad	+	-	-	+	-	-	+
PDS	-	-	+	+	+	+	+
NCBI	-	-	-	+	-	+	-
NDA	+	-	+	+	+	-	-
CSDR	-	-	+	+	+	-	-
YODA	-	-	-	-	+	-	-
MSF	+	+	+	+	-	-	-
Figshare	+	-	+	+	+	-	-
Vivli	-	-[b]	+	+	-	-	-

+ Stated in publicly available information
– Information was not mentioned in the publicly available information
a The positive symbol (+) represents the presence of fees to access information. The negative symbol (–) represents that no fees are required to access information.
b Information not publicly available
PopData Population Data of British Columbia
PDS Project Data Sphere
NCBI National Center for Biotechnology Information
NDA National Institute of Mental Health Data Archive
TB-PACTS Platform for Aggregation of Clinical Tuberculosis Studies
CSDR ClinicalStudyDataRequest.com
YODA The Yale University Open Data Access Project
ISARIC International Severe Acute Respiratory and Emerging Infection Consortium
wwPDB Worldwide Protein Data Bank
MSF Médecins sans frontières (Doctors Without Borders)

The data repositories that disclosed the most complete information in publicly available sources are PDS and NDA followed by PopData, TB-PACTS, Dryad and Figshare, which may be useful as models in general. Conversely, CSDR is the repository with the least available information. The information that was available the most was related to who can access the information, followed by who governs and what is the population covered by the data repository. On the other hand, terms of reuse and information about discoverability were the elements least mentioned.

4.2.2 Data Repositories Characteristics

4.2.2.1 Mission Statement

All 13 repositories provided either a clear mission statement or a description about the repository, or both. Please refer to Table 20 for more details.

Table 20: About the Repository and Mission Statement

Repository	About the repository	Mission
Population Data BC	Population Data BC (PopData) is a multi-university, data and education resource facilitating interdisciplinary research on the determinants of human health, well-being and development. The repository supports research access to individual-level, de-identified longitudinal data on British Columbia's 4.6 million residents. PopData is authorized to receive, store, manage, manipulate and further disclose data through Information Sharing Agreements with the Public Bodies that provide the data, as outlined in PopData's Privacy Impact Assessment.	To foster insights into human health, well-being, and development by advancing research through data and education.
Platform for Aggregation of Clinical TB Studies (TB-PACTS)	Platform for Aggregation of Clinical TB Studies (TB-PACTS) hosted by the Critical Path Institute Online Data Repository (CODR).	The TB-PACTS data platform is designed to catalyse and accelerate TB research by curating and standardizing Phase III tuberculosis (TB) clinical trial data and making this data publicly available to qualified researchers. Researches can access and analyse data in aggregate, or filter and view individual patient-level data from the REMoxTB, RIFAQUIN and OFLUTUB clinical trials. Additional trial data may be available in the future.
Reactome	Reactome is a free, open-source, curated and peer reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to	No information available

	support basic research, genome analysis, modeling, systems biology and education.	
Worldwide Protein Data Bank (wwPDB)	Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies. The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community. The RCSB PDB, PDBe, and PDBj serve as deposition, data processing and distribution sites of the PDB Archive.	The mission of the wwPDB is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community.
Infectious Diseases Data Observatory (IDDO)	IDDO brings together members of the global infectious disease community across the research and humanitarian sectors to collaborate in the generation, analysis and application of data to improve outcomes for patients.	No information available
Mendeley Data	Mendeley Data is a secure cloud-based repository where you can store your data, ensuring it is easy to share, access and cite, wherever you are. Researchers can upload and share their research data for free. Datasets can be shared privately amongst individuals, as well as published to share with the world.	Our mission is to facilitate data sharing. We believe that when research data is made publicly available, science benefits: - the findings can be verified and reproduced - the data can be reused - discovery of relevant research is facilitated - funders get more value from their funding investment - authors receive more citations
International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC)	ISARIC is a global initiative aiming at ensuring that clinical researchers have the open access protocols and data-sharing processes needed to facilitate a rapid response to emerging diseases that may turn into epidemics or pandemics.	No information available

Harvard Dataverse	Dataverse is an open source web application to share, preserve, cite, explore, and analyse research data. It facilitates making data available to others, and allows you to replicate others' work more easily. Researchers, data authors, publishers, data distributors, and affiliated institutions all receive academic credit and web visibility. A Dataverse repository is the software installation, which then hosts multiple dataverses. Each dataverse contains datasets, and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data). As an organizing method, dataverses may also contain other dataverses.	No information available
Dryad	The Dryad Digital Repository is a curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of datatypes.	Our mission is to provide the infrastructure for, and promote the re-use of, data underlying the scholarly literature.
Project Data Sphere	Project Data Sphere, LLC (PDS), an independent, not-for-profit initiative of the CEO Roundtable on Cancer's Life Sciences Consortium (LSC), operates the Project Data Sphere platform, a free digital library-laboratory that provides one place where the research community can broadly share, integrate and analyse historical, patient-level data from academic and industry phase III cancer clinical trials.	No information available

National Center for Biotechnology Information (NCBI)	No information available	As a national resource for molecular biology information, NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. More specifically, the NCBI has been charged with creating automated systems for storing and analysing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analysing the structure and function of biologically important molecules.
National Institute of Mental Health Data Archive (NDA)	The National Institute of Mental Health Data Archive (NDA) makes available human subjects data collected from hundreds of research projects across many scientific domains. The NDA provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science and discovery. De-identified human subjects data, harmonized to a common standard, are available to qualified researchers. Summary data is available to all.	No information available
ClinicalStudyDataRequest.com (CSDR)	ClinicalStudyDataRequest.com (CSDR) is a consortium of clinical study data providers. It is a leader in the data sharing community inspired to drive scientific innovation and improve medical care by facilitating access to patient-level data from clinical studies.	No information available

The YODA Project	No information available	The Yale University Open Data Access (YODA) Project's mission is to advocate for the responsible sharing of clinical research data, open science, and research transparency. The Project is committed to supporting research focused on improving the health of patients and informing science and public health. The YODA Project can only improve with your feedback. Please share your comments and ideas.
Médecins sans frontières (MSF)	MSF and Epicentre, its research affiliate, place a high value on monitoring and documenting MSF's medical interventions to improve the quality of care delivered. This results in the production of a large amount of routinely collected data. This collection of routine and research data can potentially be of value to researchers working in public health.	No information available
Figshare	Figshare is a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner. It was originally created as a solution to keep research outputs in one tidy place whilst allowing it to be discovered by likeminded individuals, the academic community. It quickly became apparent that others, too, sought such a resource and figshare opened its doors, allowing academics to upload, share, cite and importantly discover all manner of research outputs with the security of knowing our hosting options and platform support long term preservation of data.	Figshare mision is to change the face of academic publishing with the improved dissemination and discoverability and reusability of all scholarly research.

Vivli	Vivli will establish an independent general access electronic data repository and search engine through which individual participant-level data (IPD) and metadata from clinical trials conducted by researchers in academic, industry, foundation, and non-profit entities can be identified, hosted, shared and analyzed. Vivli will be agnostic to disease, country, sponsor, funder, and investigator, seeking to serve all elements of the international research community.	Promote, coordinate, and facilitate clinical research data sharing through the creation and implementation of a sustainable global data-sharing enterprise.

4.2.2.2 Data Type, Population Covered, Repository Type

Most repositories consist of patient-level data, and the majority collect clinical trials data. It is observed that data repositories with patient-level data are located in North America. All patient-level data repositories are curated and can only be accessed by researchers who have provided a research proposal. More information on the type of repository is provided in Table 21.

Table 21: Data Type, Population Covered, and Repository Type

Repository	Data Type	Population covered	Repository type
Population Data BC	Administrative data collected by the government agencies and departments (birth, health, education, early childhood development, workplace and the environment...). Data from different sources are linkable. Patient-level data.	General population of British Columbia	Curated data
Platform for Aggregation of Clinical TB Studies (TB-PACTS)	Phase III tuberculosis (TB) clinical trial data. Patient-level data.	Human subject research	Curated and standardized (CDISC-based) data
Reactome	Genomics	No information available	Curated data
Worldwide Protein Data Bank (wwPDB)	Protein	No information available	Curated data
Infectious Diseases Data Observatory (IDDO)	Ebola: Clinical & laboratory data; epidemiological data to	Patients with emergent and neglected infections: ebola,	Curated data

	come. VS: clinical trials. Malaria: not explicitly mentioned	malaria, visceral leishmaniasis (VS)	
Mendeley Data	Any published research data. Datasets seem to come in majority from biochemistry or genomics.	No information available	Platform of aggregated data
International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC)	Metadata	Human subject research	Metadata
Harvard Dataverse	Any published research data.	No information available	Platform of aggregated data
Dryad	From scientific and medical publications.	Could be patient level data	Platform of aggregated data
Project Data Sphere	Clinical trials datasets from different depositors (pharmaceutical industry, cancer research groups…). Patient level data.	Cancer patients from the comparator arm data of clinical trials	Curated and standardized (CDISC-based) data
National Center for Biotechnology Information (NCBI)	Sequence data, microarray data, bioassay data, substance or sequence-based reagents, human clinical data and genetic tests	Genomics	Curated data
National Institute of Mental Health Data Archive (NDA)	Autism research, clinical trials related to mental illness. Patient level data.	Human subject research	Curated data
ClinicalStudyDataRequest.com (CSDR)	Clinical trials. Patient level data.	Human subject research	Curated data
The YODA Project	Clinical trials. Patient level data.	Human subject research	Curated data
Médecins sans frontières (MSF)	Routinely collected data	General population	Platform of aggregated data
Figshare	Posters, code, figures, papers, videos, and datasets	No information available	Metadata
Vivli	Clinical trials. Patient level data and metadata	Human subject research	Platform of aggregated data

4.2.2.3 Governance

Out of the 17 repositories reviewed, four are owned by universities, three by private companies, two by the National Institute of Health (NIH), two are a public/private partnership, and information was not found for the remaining six repositories. Of the 4 data repositories that are governed by a Board of Directors, three are located in the US and one in the UK. More details on governance and funding may be found in Table 22.

Table 22: Governance

Repository	Who governs	Ownership	Funding
Population Data BC	Data Stewards Working Group; Advisory Board; Scientific Director; Managing Director and Unit Leads.	University owned. Simon Fraser University (SFU), University of Victoria (UVic), and University of British Columbia (UBC).	Eight (8) funders: The BC Cancer Agency, The BC Ministry of Health, The BC SUPPORT Unit, The Human Early Learning Partnership, The Michael Smith Foundation for Health Research, WorkSafeBC.
Platform for Aggregation of Clinical TB Studies (TB-PACTS)	Board of Directors	C-Path	C-Path is a public/private partnership funded by sponsors such as the FDA: grants from foundation partners such as the Bill & Melinda Gates Foundation, and fees from industry members.
Reactome	Reactome's Scientific Advisory Board members are internationally recognized researchers. The SAB meets annually to discuss Reactome's scientific agenda, explore ways to expand its research efforts and the critical review of our database, curation	Collaboration between the Ontario Institute for Cancer Research, European Bioinformatics Institute, New York University Medical Center, Oregon Science and Health University	No information available

	practices and software development program.		
Worldwide Protein Data Bank (wwPDB)	The wwPDB Advisory Committee is made up of an international team of experts in X-ray crystallography, cryoEM, NMR, and bioinformatics. The team meets annually.	No information available	- The RCSB PDB is supported by the National Science Foundation, the National Institutes of Health, and the Department of Energy. - PDBe is supported by the European Molecular Biology Laboratory, Wellcome Trust, Biotechnology and Biological Sciences Research Council, the National Institutes of Health and the European Union. - PDBj is supported by National Bioscience Database Center-Japan Science and Technology Agency.
Infectious Diseases Data Observatory (IDDO)	Board of Directors. Members are drawn from endemic regions, academia and the public health sector, and are selected based on their international expertise and standing. Board members provide advice on IDDO's strategic direction and positioning.	University of Oxford	No information available
Mendeley Data	No information available	Mendeley Data is a service provided by Mendeley Limited.	No information available

International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC)	Executive responsibility for ISARIC activities lies with the Executive Committee. The Executive Committee steers the Consortium and provides general supervision of its activities, reports to the Council, and is advised by the Stakeholder Advisory Board. The Executive Committee meets to discuss ISARIC activities once every month by teleconference.	University	No information available
Harvard Dataverse	No information available	No information available	Funded by Harvard University with additional support from the Alfred P. Sloan Foundation, National Science Foundation, National Institutes of Health, Helmsley Charitable Trust, IQSS's Henry A. Murray Research Archive, and many others.
Dryad	Dryad Members nominate and elect the Board of Directors, twelve individuals from the stakeholder community who provide strategic planning, fiscal oversight, and oversee the position of the Executive Director.	No information available	U.S. National Science Foundation, 2016-2019. European Commission under the Horizon 2020 programme, 2015-2018. Dryad-Open Science Framework integration grant, 2015.

Project Data Sphere	Officers include: President, 1st Vice President, Treasurer, Assistant Treasurer. Governance includes the Executive Committee, which functions with an advisory capacity.	Project Data Sphere, LLC, a Delaware non-profit, limited liability company.	Project Data Sphere is supported by the members of the CEO Roundtable on Cancer's Life Sciences Consortium through voluntary, in-kind contributions.
National Center for Biotechnology Information (NCBI)	The Board of Scientific Counsellors of the National Center for Biotechnology Information (NCBI) advises the NIH Director, Deputy Director for Intramural Research, the NLM Director, and the NCBI Director about the intramural research and development programs of the NCBI. Regularly scheduled visits are held by the Board members for assessment of NCBI research and development programs in progress, assessment of proposed programs, and evaluation of the productivity and performance of staff scientists. Meetings are held twice a year at the National Library of Medicine.	National Institute of Health (NIH)	No information available

National Institute of Mental Health Data Archive (NDA)	The Director of the National Institute of Mental Health (NIMH) oversees the NDA, its policy, and implementation. In carrying out this responsibility, the NIMH Director participates on a Governing Committee, with several other NIH Institute and Center Directors, or their designees that fund the NDA. The Governing Committee is responsible for the on-going management and stewardship of NDA Policy and procedures.	NIH	Funded in part by National Institute of Mental Health (NIMH), National Institute of Neurological Disorders and Stroke (NINDS), National Institute of Environmental Health Sciences (NIEHS), The Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD).
ClinicalStudyDataRequest.com (CSDR)	No information available	No information available	No information available
The YODA Project	The Steering Committee provides guidance to the YODA Project, specifically in reference to developing processes for making clinical trial data available to external investigators, as well as processes for reviewing requests for these data.	Yale University, Connecticut	No information available
Médecins sans frontières (MSF)	No information available	No information available	No information available
Figshare	No information available	Figshare LLP	No information available
Vivli	Board of Directors	No information available	Laura and John Arnold Foundation contributes funding to launch Vivli

4.2.2.4 Data Accessibility

Most data are discoverable with a digital object identifier (DOI) or Handle®. However, 9 repositories did not provide information regarding discoverability. Fees are applicable to use the data in only 2 repositories. Only 6 repositories provide information on intellectual property. See Table 23 for more details.

Table 23: Discoverability, Intellectual Property, and Fees

Repository	Discoverability	Intellectual Property	Fees
Population Data BC	No information available	No information available	Yes
Platform for Aggregation of Clinical TB Studies (TB-PACTS)	No information available	No information available	No
Reactome	DOI	The contents of Reactome are copyright (c) 2003-2016 Cold Spring Harbor Laboratory (CSHL), Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI).	No
Worldwide Protein Data Bank (wwPDB)	No information available	Data files contained in the PDB archive are free of all copyright restrictions and made fully and freely available for both non-commercial and commercial use.	No
Infectious Diseases Data Observatory (IDDO)	No information available	No information available	No informa-tion available
Mendeley Data	DOI	Mendeley is the owner of all intellectual property rights on their Site and on the material published on it.	No
International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC)	No information available	No information available	No
Harvard Dataverse	DOI or Handle®	No information available	No
Dryad	DOI	Screenshots, Dryad Digital Repository name, Dryad name and other Dryad trademarks to be used for informational purposes only. Content other than Data Packages on the Dryad Website and wiki is available for reuse with attribution under a Creative Commons CC-BY 3.0 license.	No
Project Data Sphere	No information available	No information available	No
National Center for Biotechnology Information (NCBI)	No information available	No information available	No

National Institute of Mental Health Data Archive (NDA)	DOI	Most information at this site is in the public domain. Unless stated otherwise, documents and files on NIH Web servers can be freely downloaded and reproduced. Most documents on this server are sponsored by the CIT; however, you may encounter documents that were sponsored along with private companies and other organizations. Other parties may retain all rights to publish or reproduce these documents or to allow others to do so. Some documents available from this server may be protected under the U.S. and foreign copyright laws. Permission to reproduce these documents may be required.	No
ClinicalStudyDataRequest.com (CSDR)	No information available	No information available	No
The YODA Project	No information available	No information available	No
Médecins sans frontières (MSF)	No information available	Recipients shall not seek any Intellectual Property rights of any kind in respect of Results generated or arising out of the use of MSF Datasets without prior written consent.	No informa-tion available
Figshare	DOI	CC-BY (figures, media, posters, papers, filesets) By licensing research outputs under CC-BY, figshare ensures that the research is openly available, but requires that others should give credit to the author, in the form of a citation, should they use or refer to the research object. This licence lets others distribute, remix, tweak, and build upon published work, even commercially, as long as they credit to the author for the original creation. This is the most accommodating of licences offered. It is recommended for maximum dissemination and use of licensed materials. CC0 (datasets and metadata) CC0 can be particularly important for the sharing of data and databases, since it otherwise may be unclear whether highly factual data and databases are restricted by copyright or other rights. Databases may contain facts that, in and of themselves, are not protected by copyright law. CC0 is recommended for data and databases and is used by hundreds of organisations. It is especially recommended for scientific data. Although CC0 doesn't legally require users of	No

		the data to cite the source, it does not take away the moral responsibility to give attribution, as is common in scientific research.	
Vivli	DOI	No information available	No informa-tion available

4.2.2.5 Data Access

Seven repositories restrict access to researchers only, while data is available to anyone in 4 repositories. Six repositories have disclosed details on their access policies; ranging from the creation of an account to the development of a research proposal. Data repositories that do not have an access policy, meaning that technically anyone could have access to the data, do not store human data (i.e., genomics, proteins sequences). Table 24 provides more details on data access.

Table 24: Data Access

Repository	Access Policy	Who can access?	Access Procedure
Population Data BC	PopData enters into separate Information Sharing Agreements, Data Directives or other data sharing agreements with government ministries and public agencies (collectively, the "Data Stewards") for health information and other Personal Information on the population of British Columbia relating to human health, well-being and development. Personal Information which PopData holds from the Data Stewards under these agreements are referred to as "Data". Each agency retains ownership of	Eligible researchers. Only researchers who will conduct their analyses in Canada are eligible to apply for access to data pursuant to FIPPA section 33.2(k).	1. Data access request form filled and submitted 2. Data Steward review 3. Contracts and account set up 4. Data preparation and delivery

	its particular Data set(s) and reviews and approves requests for access to its Data.		
Platform for Aggregation of Clinical TB Studies (TB-PACTS)	No information available	Researchers	Web registration. The TB-PACTS steering committee will review all user access applications in a timely manner and this may take up to 4 weeks to process.
Reactome	No information available	Everybody	None
Worldwide Protein Data Bank (wwPDB)	None	Everybody	None
Infectious Diseases Data Observatory (IDDO)	No information available	Researchers	An application form including the researcher's information and a research plan of the study must be approved. Only Ebola data can be accessed so far.
Mendeley Data	No information available	Datasets can be shared privately amongst individuals, as well as published to share with the world.	Web registration
International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC)	If required by investigators' institutional requirements, an agreement should be signed between involved parties for sharing samples and data in compliance with local and international laws (if applicable).	Researchers	Create account
Harvard Dataverse	No information available	Depositor can restrict access to data, but not metadata.	For gaining full Use of Service, one must register for and be logged into an Account on Harvard

			Dataverse.
Dryad	No information available	Everybody	No information available
Project Data Sphere	Agreement document. Access is granted for 1 year, unless renewed before expiration.	Researchers affiliated with life science companies, hospitals and institutions, as well as independent researchers.	Access to the Platform will be provided to bona fide researchers who submit a completed online application and agree to the terms, covenants and conditions found in the Agreement document, upon receiving access authorization from Project Data Sphere. A summary of initial research goals is requested during the user application process.
National Center for Biotechnology Information (NCBI)	None	Anybody	None
National Institute of Mental Health Data Archive (NDA)	Access to subject level datasets submitted and stored in the NDA will only be provided for research purposes through the completion of the NDA Data Use Certification. For the majority of the data available in the NDA, Data Use Certifications will only be accepted from researchers who are sponsored by an institution registered in the NIH's eRA Commons with an active Federal-wide Assurance issued through the Office for	Qualified researchers who have completed a Data Use Certification and received approval.	The established Data Access Committees or its designees will review requests for access to determine whether the proposed use of the dataset is scientifically and ethically appropriate.

	Human Research Protections (OHRP).		
ClinicalStudyDataRequest.com (CSDR)	Researcher must sign a Data Sharing Agreement	Researchers	Researchers can submit research proposals and request anonymised data from clinical studies listed on this site. Following approval and after the relevant study sponsor or sponsors receive a signed Data Sharing Agreement, access to the data needed for the research is provided on a password protected website.
The YODA Project	No information available	No information available	Must create an account and request access with a research proposal
Médecins sans frontières (MSF)	Once transfer of requested MSF Dataset(s) has been accepted, and before transfer is effectively granted, Requestors must agree to the conditions of access and return a signed material transfer agreement (MTA) to MSF.	Access to the Collection is limited to all appropriately qualified researchers from academia, charitable organizations and private companies, such as drug companies. Details of requestor and research must be provided.	No information available
Figshare	All public content hosted on figshare can be downloaded by anyone, with no need to log in. The content can be mass downloaded or mined using the figshare API, also available to anyone.	Anybody	None

Vivli	Vivli will also provide a secure environment for data requestors to combine anonymized individual patient-level data (IPD) from different hosts and data generators, including industry, academia, and biotech. In this controlled-access environment, strict data security models and export controls can be applied according to the specifications of each data contributor.	Researchers	No information available

4.2.2.6 Terms and Licences for Reuse

Four repositories use the Creative Common license for reuse. Other repositories have different terms or licenses for reuse. Please refer to Table 25 for details on terms and licenses for reuse.

Table 25: Terms and Licences for Reuse

Repository	Terms of Reuse	Licenses for Reuse
Population Data BC	No information available	No information available
Platform for Aggregation of Clinical TB Studies (TB-PACTS)	By posting, uploading, inputting, providing or submitting a Submission, depositors are granting C-Path, its affiliated companies and necessary sublicensees permission to use their Submission in connection with the operation of their Internet businesses including, without limitation, the rights to: copy, distribute, transmit, publicly display, publicly perform, reproduce, edit, translate and reformat your Submission; and to publish your name in connection with their Submission.	No information available
Reactome	No information available	The software and information contents are distributed under the terms of the Creative Commons Attribution 4.0 International License, which grants parties the non-exclusive right to use, distribute and create derivative works based on Reactome, provided that the software and information is correctly attributed to CSHL, OICR and EBI.
Worldwide Protein Data Bank (wwPDB)	No information available	No information available
Infectious Diseases Data Observatory (IDDO)	No information available	No information available
Mendeley Data	No information available	Creative Commons and open software licences. This means depositors retain control of the data, and choose the terms under which others may consume and reuse it. They may delete their dataset at any time, by contacting Mendeley Data.
International Severe Acute	No information available	Depends on investigator

Respiratory and Emerging Infection Consortium (ISARIC)		
Harvard Dataverse	No information available	Default license is a Creative Commons Zero ("CC0") Public Domain Dedication Waiver. Users also have the option of drafting a custom data usage license agreement. Users also have the option of choosing to use Harvard Dataverse's restricted data usage license agreement ("Data Use Agreement").
Dryad	No information available	Reusable under Creative Commons Zero (CC0) waiver.
Project Data Sphere	You acknowledge that each Data Provider retains ownership of, and all intellectual property rights in, the Data it has made available in the Database and that you acquire no rights in the Data other than those limited rights set forth in this Agreement. You do not obtain any rights under any patents, copyrights or other intellectual property rights of any Data Provider – whether implicitly, by estoppel or any other legal theory – other than the limited license to use the Data as expressly permitted by this Agreement and subject to all terms, covenants and conditions set forth herein and in any applicable Supplemental Terms.	User Contributions You retain your intellectual property rights to any User Contributions you post on or through the Community Tools. By posting User Contributions on or through the Community Tools, you grant to Project Data Sphere and to all Authorized Users to whom you make the User Contributions available a worldwide, non-exclusive, transferable, sub-licensable, royalty-free, perpetual and irrevocable license to use, copy, reproduce, process, adapt, modify, publish, transmit, perform, display and distribute such User Contributions in any and all media and distribution methods (now known or hereafter developed) for the Purpose. You hereby waive your moral rights in any User Contributions.
National Center for Biotechnology Information (NCBI)	NCBI itself places no restrictions on the use or distribution of the data contained therein. Nor do we accept data when the submitter has requested restrictions on reuse or redistribution. However, some submitters of the original data (or the country of origin of such data) may claim patent, copyright, or other intellectual property rights in	No information available

	all or a portion of the data (that has been submitted). NCBI is not in a position to assess the validity of such claims and since there is no transfer of rights from submitters to NCBI, NCBI has no rights to transfer to a third party. Therefore, NCBI cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in the molecular databases.	
National Institute of Mental Health Data Archive (NDA)	No information available	No information available
ClinicalStudyDataRequest.com (CSDR)	No information available	No information available
The YODA Project	No information available	No information available
Médecins sans frontières (MSF)	No information available	No information available
Figshare	No information available	No information available
Vivli	No information available	No information available

5 Discussion - Gap Analysis

Data sharing is crucial for the development of new evidence, which its availability in a timely manner is critical in an epidemic outbreak. Many data sharing platforms have been created from different initiatives and with different standards. Harmonisation may simplify storage, sharing, reuse and analysis of new data.

5.1 Data Curation Standards

Based on the results of this study, data repositories that contain patient level data should be curated.

Data should have a unique and persistent identifier in order to ensure the discoverability and referential integrity between the information associated in the AIP. Only 7 data repositories disclosed the use of DOI for ensure discoverability.

Data should be organized and formatted using the Study Data Tabulation Model (SDTM), which defines a standard structure for study data tabulations and concentrate the data in one file minimizing the need of multiple files that contain the same data and streamline processes in collection, management, analysis and reporting (Clinical Data Interchange Standards Consortium (CDISC), 2016).

5.2 Security and Longevity Standards

It was observed that when patient-level data are stored in a repository, access should be restricted to qualified researchers. On the other hand, data repositories with no Human data have a free access policy.

An access control policy should be implemented to record all access actions, including requests, access denial incidents, etc. This control should also help to create a log and to review all access failures and anomalies, which will be useful to record, analyse and react to any potential threat. Repositories should have a risk management plan and describe damage scenarios based on malicious actions, human error, or technical failure that pose any risk. There should be also a continuous analysis and assessment of security risk factors associated with data, products, services, and users.

Repositories should implement strategies and procedures for managing archival storage of the data, which should be stored in contemporaneous formats. Data repositories must have a mission statement reflecting the commitment to preservation, retention, management and access to data.

Sustainability of funding should be assured according to the standards shown in Table 11. Repositories should prepare a short- to long-term business plan and should analyse and report on risk, benefit, investment, and expenditure. From the information obtained from repository websites, none of them mentioned the existence of a business plan. Therefore, this information should be gathered during the interview. Funding sources should be well stablished. Only half of the data repositories disclose this information.

5.3 Governance Standards

Data repositories must publicly disclose registry ownership, governance structure and not-for-profit status. Two out of the 17 repositories reviewed in this project do not disclose this information. There should be contracts or deposit agreements for digital materials which specify all necessary preservation rights of the data it manages, preserves, and/or to which provides access. In a similar vein, repositories should have written policies that specify the nature of any legal permissions required to preserve digital content over time.

Data access policies, conditions and all applicable licenses covering data access and use must be communicated to all users, whom must be monitored to ensure compliance. Data access is restrained to researchers on half of the repositories in this study, the other half is open to any individual. Similarly, an online registration form must be completed in half of data repositories in order to obtain access to

information, 3 and 4 need a simple web registration or no registration whatsoever, respectively. When WHO TRDS items are used, registries must be accessible online at no charge to the end user. From the 17 repositories reviewed in this study, only 2 have fees to access to data. Repositories must track and manage intellectual property rights and restrictions on use of data as disposed by its conditions. Intellectual property is disclosed on 7 data repositories.

6 Study Strengths and Limitations

The list of data repositories obtained in this study is limited by the keywords used in the search. Hence, some data repositories may not have been found. Nevertheless, the systematic review that was conducted in the first component of the project was very comprehensive, and it is unlikely that the search would have missed existing data repositories.

Availability of information on all standard components is limited in publicly available sources, which limited the gap analysis. Some components may have been considered as a gap, while in fact it was due to absence of information. Further information may be sought by questionnaires or interviews with custodians/curators in order to seek more details on specific components of governance.

7 Summary

In this review, we found six sets of standards for data repositories that address a broad range of components, ranging from technical aspects such as meta-data structure to governance and data access. We conducted an in-depth review of 17 existing data repositories, with a special emphasis on governance, and used the individual components of the standards as a functional checklist for evaluation. Based on our in-depth review, it appears that the majority of data repositories do not entirely follow these standards or some of the standards used do not address all elements relevant to governance. In other words, there are no harmonised standards that meet all needs. In order to develop these best practices, there must be a homologation of standards that includes the experiences and thoughts from all data repositories, taking particular attention to include big and small data repositories. Although our review was limited to the publicly available information sources, our gap analysis supports the need for more robust standards used in disease-specific or general population data repositories, and would support homogenization of the standards, mainly for data curation, security and longevity.

8 References

Clinical Data Interchange Standards Consortium (CDISC). (27 de June de 2016). Obtenido de https://www.cdisc.org/system/files/members/standard/foundational/sdtm/SDTM%20v1.5.pdf

Moride Y., et al. (2017). A Systematic Review of Interventions and Programs Targeting Appropriate Prescribing of Opioids. *Pharmacoepidemiology and Drug Safety, 26*(Suppl. 2), 635. doi:10.1002/pds

Society of American Archivists. (3 de August de 2012). *Open Archival Information System (OAIS)*. Obtenido de https://www2.archivists.org/groups/standards-committee/open-archival-information-system-oais

9 Appendix 1: Systematic Review of Online Repositories for Neglected Infectious Diseases

See PDF document attached.

10 Appendix 2: Repositories Contact Information

Repository	Contact
Population Data BC	General enquiries: Call: 604.822.8616 or Email: info@popdata.bc.ca
Platform for Aggregation of Clinical TB Studies (TB-PACTS)	Debra Hanna, Executive Director, Critical Path Institute dhanna@c-path.org +1 520-547-3440 info@c-path.org +1 520 547-3440
Reactome	help@reactome.org http://reactome.org/pages/about/reactome-scientific-advisory-board/
Worldwide Protein Data Bank (wwPDB)	info@wwpdb.org RCSB PDB: Team e-mail: info@rcsb.org PDBe: Team e-mail: http://www.ebi.ac.uk/pdbe/?tab=aboutus&subtab=contactus PDBj: Team e-mail: http://www.pdbj.org/contact BMRB: Team e-mail: bmrbhelp@bmrb.wisc.edu
Infectious Diseases Data Observatory (IDDO)	Email: info@iddo.org OR J.O'Callaghan@wellcome.ac.uk
Mendeley Data	https://service.elsevier.com/app/contact/supporthub/mendeley/ OR support@mendeley.com OR mendeley-community@mendeley.com.
International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC)	kajsa-stina.longuere@ndm.ox.ac.uk Phone (office): +44 (0)1865 612965
Harvard Dataverse	support@dataverse.org
Dryad	Meredith Morovati, Executive Director director@datadryad.org
Project Data Sphere	https://www.projectdatasphere.org/projectdatasphere/html/contactUs

	Christopher A. Viehbacher, President Dr. James Goodnight, Vice President Robert A. Ingram, 2nd Vice President Martin J. Murphy, DMedSc, PhD, FASCO, Chief Executive Officer David Handelsman, Vice President of Development
National Center for Biotechnology Information (NCBI)	https://www.ncbi.nlm.nih.gov/home/about/contact/
National Institute of Mental Health Data Archive (NDA)	Email: NDAHelp@mail.nih.gov Phone: 301-443-3265
ClinicalStudyDataRequest.com (CSDR)	support@clinicalstudydatarequest.com
The YODA Project	yodap@yale.edu http://yoda.yale.edu/leadership-contact-information
Médecins sans frontières (MSF)	Telephone: +41 22 849 8484
Figshare	Telephone: +44 (0) 20 7418 5573 Email: info@figshare.com
Vivli	Address: 14 Story Street, 4th Floor, Cambridge, MA 02138 USA Telephone: 617-496-9376 Email: contact@vivli.org