OPINION ARTICLE

The ideal repository for hosting data from clinical trials: blueprint using business process management [version 1; peer review: 1 approved with reservations, 1 not approved]

Mirko Gabelica[1], Damir Sapunar[2], Matko Marušić[3], Livia Puljak[4]

[1]Department for Ear, Nose, and Throat Disorders with Head and Neck Surgery, University Hospital Split, Split, 21000, Croatia
[2]Laboratory for Pain Research, Medical School at Split, Split, 21000, Croatia
[3]Medical School at Split, Split, 21000, Croatia
[4]Center for Evidence-Based Medicine and Health Care, Catholic University of Croatia, Zagreb, 10000, Croatia

v1 **First published:** 14 Jan 2021, **10**:23
https://doi.org/10.12688/f1000research.28008.1
Latest published: 14 Jan 2021, **10**:23
https://doi.org/10.12688/f1000research.28008.1

Abstract

In this article, we suggest a blueprint for an ideal open-access repository for clinical trial data with a description of a model of such a repository using a business process analysis approach. Firstly, we suggested which features an ideal repository should have. Secondly, we used business process management software to describe the whole process, from the decision to share clinical trial data to either publication of data in a repository or discarding data. The research community, legislators and society at large should be interested in a transparent open-access repository that will host clinical trial data. We hope this work can inspire relevant stakeholders to engage in discussion about the necessity of creating such repository, and that we will witness the creation of such a repository in the near future.

Keywords

repository, business process management, clinical trials, data sharing, raw data, individual patient data

Open Peer Review

Approval Status ? ✗

	1	2
version 1 14 Jan 2021	? view	✗ view

1. **Paul Grefen**, Eindhoven University of Technology, Eindhoven, The Netherlands
2. **Ida Sim**, University of California San Francisco, San Francisco, USA

   Rebecca Li, Harvard University, Cambridge, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Health Services gateway.

This article is included in the Research on Research, Policy & Culture gateway.

Corresponding author: Mirko Gabelica (gabelica@gmail.com)

Author roles: Gabelica M: Investigation, Methodology, Writing – Original Draft Preparation; **Sapunar D**: Conceptualization, Investigation, Project Administration, Supervision, Visualization, Writing – Review & Editing; **Marušić M**: Conceptualization, Supervision, Validation, Writing – Review & Editing; **Puljak L**: Conceptualization, Methodology, Project Administration, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Gabelica M, Sapunar D, Marušić M and Puljak L. **The ideal repository for hosting data from clinical trials: blueprint using business process management [version 1; peer review: 1 approved with reservations, 1 not approved]** F1000Research 2021, **10**:23 https://doi.org/10.12688/f1000research.28008.1

First published: 14 Jan 2021, **10**:23 https://doi.org/10.12688/f1000research.28008.1

Introduction

Considerable interest has been shown recently in increasing transparency of clinical trials. National Institutes of Health (NIH) defined clinical trials as research studies that explore whether a medical strategy, treatment or device is safe and effective for humans, and, if conducted well, they produce the highest level of evidence available for healthcare decision making among primary studies[1]. However, very often raw data from clinical trials are hidden from scientific community[2,3]. Sharing individual patient data (IPD) from clinical trials in a central openly available repository was suggested as a solution[4].

Although ideas about open data sharing come mostly from researchers, it has recently been shown that clinical trial participants support this idea too. A recent survey of individuals who participated in a diverse group of clinical trials showed that the overwhelming majority would support sharing of their data and that their willingness to share data would not be much different depending on the purpose of the data use[5].

In computer sciences, a repository is defined as a central location in which data are stored and managed[6]. Currently, there are no repositories that host exclusively open-access data from clinical trials. Our recent study (Gabelica *et al.*; unpublished data) indicated that there are 13 open-access repositories on the Internet, which host clinical trial data together with data from other types of studies. However, those repositories were highly heterogeneous, are not devoted to clinical trials exclusively and most allow data providers to restrict data access for the "shared" data. For this reason, there is a need for a universally adopted open-access repository devoted specifically and exclusively to data from clinical trials.

The US National Academy of Medicine (NAM), formerly called the Institute of Medicine (IoM), has indicated that it would be beneficial to have one single centralized data store, "to collect all clinical trial data worldwide into one central database". Such a model would benefit from economies of scale, and individuals or groups interested in data would need to search in one database only[7]. It is recognized that there are challenges associated with such an approach, but if we start addressing the challenges, and considering how such a repository would look like, we could make it a reality one day.

This manuscript aimed to propose how an ideal open-access repository for clinical trial data should look like and to develop a model of such repository using business process analysis approach.

Our approach

Firstly, we suggested which features an ideal repository should have. Some of the features were informed by the earlier study that searched for repositories hosting raw data from clinical trials and analyzing their characteristics. Insufficiencies of existing repositories were taken into account and several additional characteristics were suggested, as we continued envisioning an ideal repository[8].

Secondly, we used business process management BPA software, ARIS Express (Software AG, Darmstadt, Germany) to describe the flow of the entire process from decision to deposit data to either data being published or discarded. ARIS was selected as an adequate tool for this task, as it contains a visual representation of vital elements needed for the task. Documents, software, people at hand, etc. are placed in a 2D setting, connections are plain and unambiguous, the expected outcome is clear and perspective is not tainted with complex models. Business process management software allows precise problem identification, and reference model towards solving weak points, continuous quality control and monitoring[9].

Features of an ideal repository

Vision

The ideal clinical trial data repository should be the first place on the internet for searching clinical trial data from published and unpublished clinical trials. The ideal open-access repository for hosting raw data from clinical trials would be a public Internet-based resource.

General features:

1. Exclusivity. Repository accepts exclusively data from clinical trials, including raw data, analyzed data and meta-data. The user interface will be exactly tailored to fit the deposition of clinical trial data.

2. Mandatory use. In line with the requirements of the International Committee of Medical Journal Editors (ICMJE) for mandatory trial registration, relevant stakeholders such as employers, funders and journal editors could agree that clinical trial data need to be deposited in a clinical trial repository.

3. International governance. An international board of relevant stakeholders from academia, industry and funders is governing the repository. These stakeholders should be internationally renowned, non-profit organizations with stable funding, such as large universities, research funding agencies, European Medicines Agency (EMA), or similar.

4. The repository is self-sustainable. An ideal repository needs to develop a sustainable collaborative funding model that will ensure the maintenance and continuing development of the repository, providing new tools and storing new datasets, while ensuring that the repository is free to access and reliable[10]. Such a collaborative model could, for example, include financial support of governments, as part of their investment in research.

   Cost of data deposit is free or minimal. Data deposit is free for data depositors, or partially funded from grants, or institutions or government if such funds exist. Lack of funds should not prevent data deposition. Since funders now regularly cover the cost of publications in open-access journals, principal investigators applying for grants can also include the cost of data deposition in an open-access repository. Principal investigators without funds can apply for a fee waiver. The cost of data deposition, if there is any cost at all, should not be prohibitively high, and in line with the cost of manuscript publication.

Features related to user experience:

5. English is the main language of the repository framework, with the option to create sibling web sites in other languages. The uploaded files can be in any language, and language of files is indicated, with the preference for uploading files in the English language to achieve maximum visibility.

6. Simple user interface and searching. The user interface is a friendly and self-explanatory environment, which enables step by step upload. All the content in the repository is searchable.

7. Updates and corrections are archived. The repository enables subsequent updates and corrections to a deposited dataset, where each change is explained and recorded, and each version of the dataset is archived and accessible.

8. Mandatory inclusion of metadata with clinical trial data. Metadata include a comprehensive separate data set that should answer all potential answers about the clinical trial and data from a trial. Such metadata enables managing clinical data portfolio, enable assessment of the conduct and analysis of those trials, and reanalysis[11–13]

9. Instructions for preparing and depositing data and metadata. Extensive instructions on data preparation for deposition are available on-site, with clear statements on mandatory data and metadata deposition[13]. da Silva *et al*. concluded that most researchers store their data in various formats and the main reason for data loss is lack of appropriate annotation[14].

10. The maximum upload file size is 2 GB, while the maximum project upload size is unlimited so that researchers can upload all the files that are associated with a clinical trial. Image compression in an ideal repository should be lossless. Currently, there is a problem with deposition and archiving of medical images from MR, CT scans and similar devices that generate large file sizes. Mezrich and Siegel addressed the need for universal and technological appropriate guidelines regarding storing digital medical images, with the help of the information technology community. Image compression has been suggested as an approach but no guidelines have yet been made and adopted[15]. Although Koff's study found no difference in diagnosis based on low level compressed and uncompressed images, evaluating the standards for irreversible compression in digital diagnostic imaging has been proposed by the Canadian association of radiologists[16].

11. Persistent identifier is assigned to each dataset. A digital object identifier (DOI) is an important international standard for identification of online material. A DOI is therefore provided to each dataset (complete data and metadata for one study). It is vital for digital objects (articles, datasheets, images) to receive a DOI, as it helps to avoid several issues with citations, such as broken links (marked with a warning: error 404), copy-paste errors in citation text and copyright violation. Also, a DOI enhances verifiability, because it always leads to the correct web source[17]. According to Klump and Huber, a DOI is used in 75% of repositories as the most common persistent identifier, making it most successful persistent identification system currently in use[18]. Price of DOI is 1$ in the most expensive scenario for an article and 0.06$ for data set; the price varies according to DOI issuing agency[19].

12. Mandatory manual curatorship of datasets. After deposition, data are verified by at least two experts independently, such as a biocurator and a statistician[20,21]. The UK's Digital Curation Centre suggests outlines of their approach to digital curation procedure, with several steps, of which the following are relevant for ideal repository: i) conceptualization: considering which digital material will be stored, which data capture methods will be used and available storage options; ii) creation: production of relevant metadata because it enhances accessibility; iii) access and use: determining whether data are publicly accessible, whereas for ideal clinical trial open data repository limited accessibility is an option to consider, as well as embargo options before the publication of results; iv) appraisal and selection: determining what digital data is relevant, in respect to legal guidelines if they exist; v) disposal: discarding irrelevant data; vi) ingesting: placing digital objects to predetermined storage locations; vii) preservation: taking actions that will ensure long-term data protection and retention of the nature of digital material; viii) reappraisal: reevaluate material to ensure that is still relevant and is true to its original form; ix) storage: keeping the data secured; and x) access and reuse: routinely check that material is still accessible[22,23].

13. A limited embargo period is allowed. Investigators can deposit the data after obtaining results, before the manuscript is submitted, but with an embargo that will be in effect until manuscript publication. The maximum embargo period that investigators can request is 1 year[24].

14. Data and metadata are reusable. Curators need to confirm that data is reusable and analyzable by performing

minimal reanalysis according to the internally-agreed uniformed data reanalysis protocol, to confirm at least one result from the published manuscript.

15. Access to data is open after the registration. Users have open access to data after registration on the site, accredited via affiliation.

16. Data can be reused. Datasets are published under a Creative Commons Attribution 4.0 International License, which allows maximum dissemination and data reuse. Users will be free to share and remix the dataset, under the condition that they attribute the source of the dataset to the original author[25,26].

17. Enabled interconnectivity. Clinical trial data repository should be connected with protocol registries, such as ClinicalTrials.gov, ISRCTN and EudraCT, and with published journal article by using a DOI. Links to protocol registrations and journal articles are to be found on the repository website.

18. Researcher identification should be managed with an ORCID ID, a nonproprietary alphanumeric code whose purpose is to provide a unique persistent identifier to academic authors. An ORCID ID is thus similar to DOI. Even though the ORCID organization warns that they are not an identity verification system, many universities and publishers, along with commercial companies, promote and use ORCID[27].

19. Management of IPD is crucial and the most demanding process related to the design of the ideal repository. It has to be defined according to the EU General Protection Regulation (GDPR) which is designed to harmonize data privacy laws across Europe and to protect all EU citizen data privacy and redefine the way organizations across the EU approach data privacy. The enforcement date for GDPR was May 25, 2018, and since that date organizations in non-compliance will face heavy fines. Deposited data should be prepared and deposited in such a way that the data subject is no longer identifiable. For example, IPD should contain a code instead of participant real name, address, email, photo, phone number or social security number should not be in the table. It is not likely that someone could be identified via sex, age and arterial pressure, blood glucose and TNM stage. All other questions regarding GDPR compliance should be managed the by the clinical trial principal investigator or dedicated officer[28].

20. The repository should qualify for CoreTrustSeal, which guarantees that the repository has been created according to 16 guidelines for a sustainable and trustworthy repository[29].

21. Organization of metadata should be following the Dublin Core Metadata Initiative, specifically DCMI metadata terms[30]. DCMI maintains authoritative specification of metadata terms, and those terms are published as IETF RFC 5013 [RFC5013], ANSI/NISO Standard Z39.85-2007 [NISOZ3985], and ISO Standard 15836:2009 [ISO15836]. Description of every single metadata repository element is beyond the scope of this article

Benefits of the suggested ideal repository

For the uploader: The repository would be safe archival space where one can store an unlimited amount of data. These data may serve to others, but also as a backup for the uploader, in terms of data protection and preservation.

For other researchers: It would be a user-friendly interface and smart search engine that would provide easy access to clinical trial results, and dataset acquisition in just a few seconds. These data can then be reused and reanalyzed.

For stakeholders concerned with research integrity: It would help stakeholders check whether clinical research data have been fabricated or falsified. Preventing statistical analysis results unfavourable to the researcher (including the funding agency) from being concealed, and preventing fraud in a clinical trial results publishing[2,31].

For legislators: The repository would combine ethical science and reporting of results with good clinical practice. If problems emerge, all data from the beginning to the end of the study is available on site.

For science, at large: Implementation of the repository would lead to a reduction of waste in research. Currently, many RCTs do not report all data collected within the study. Other researchers or healthcare workers may benefit from knowing that certain data were collected and analyzed, and accessible in a repository[32]. Such a mandatory repository would mark an era of truly open science and data sharing in clinical trials.

Business process management approach for successful management and organization of an ideal repository

The focus of this project was a development of IPD and clinical trial data deposition and curation scheme, or to put it simply, what happens to data when a researcher deposits them to the repository. BPM was used to identify all necessary steps required for successful data management from deposition, checking whether data were adequately prepared via human curation and DOI assignment, and finally data publishing on the repository website. Figure 1 shows the entire process that we are proposing. The main goal was to ensure that there are no loose ends in the data management lifecycle; once the data is deposited in the repository, the project must end as published or discarded data.

The process in the middle of the model has left arm which describes the necessary documents and tools. The right arm defines the person responsible for performing a task that leads to another process until the process is finished as discarded data or published data. Every process begins with a decision whether or not one will enter the process. When researchers decide to share

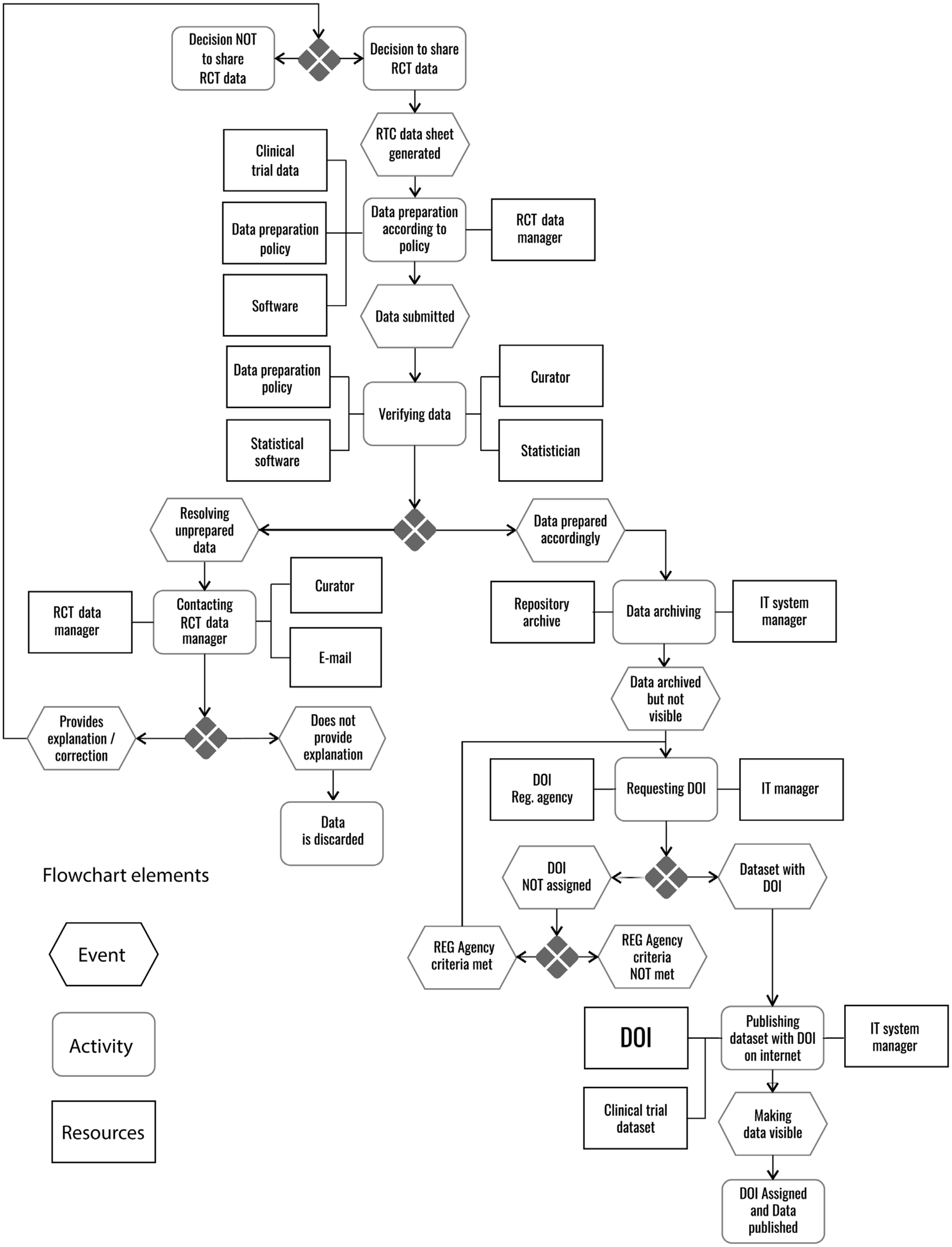

Figure 1. **Process management scheme for clinical trial raw data deposition in the repository.**

their data, clinical trial data manager puts an effort to generate RCT data sheet (block) they want to share. The next step is preparing data according to repository policy so data could be successfully ingested and manipulated until publication, e.g. images, sheets, videos, written data, should be submitted in acceptable formats. After preparation, data is submitted via an online form and the clinical trial data manager's job is now finished (Figure 1).

The repository should have curator(s) and statistician(s) proficient in the area of clinical trials. After data submission, curators will now verify whether deposited data has been prepared according to relevant policy; therefore, we see a branching process if data is not prepared accordingly (Figure 1). In that case, clinical trial data manager will be contacted to explain not following policy instructions. If an explanation is provided, the process begins again from the top; if an explanation is not provided, the process will terminate, and the data will be discarded (Figure 1).

If data are prepared properly, they are placed in repository archive but are not visible to others. Requesting a DOI for an archived dataset is the next process, which also branches in two arms, depending on registration agency criteria for DOI assignment. This means that the registration agency can decline DOI assignment. If that happens, the process is to be repeated until criteria are met to acquire DOI. Next step is publishing dataset with DOI on the repository website, making it fully visible online (Figure 1).

Other processes that precede data deposition, and processes after data publication are not addressed in this article. Ohman *et al.* advised 10 principles and 50 recommendations that should be taken into consideration for successful clinical trial data sharing in general. They described steps to be taken from data preparation to data sharing monitoring[33]. Ohman *et al.* focused on ideal data sharing prerequisites, while we focus on technical aspects regarding ideal repository features, data validation and publication of such data.

Discussion

This manuscript describes features of an ideal repository for hosting open-access data from clinical trials, and ideal schema for its creation, using a business management approach. Such a central, mandatory repository for clinical trials is necessary because we are witnessing numerous calls for action, statements, articles, open letters, organizations, projects, and initiatives regarding "open data" movement[34,35].

However, an ideal repository needs to be backed up by a legislative framework to be usable and sustainable. Without it, all efforts for creating sustainable clinical trials repository will be futile and dispersed[36]. The legislation behind ClinicalTrials.gov is a relevant example. The first milestone towards open clinical trial science was achieved in 2008, with legislation that backed up mandatory basic results reporting at ClinicalTrials.gov through Section 801 of the Food and Drug Administration Amendments Act (FDAAA 801), and again in January 2017 with the Final Rule for Clinical Trials Registration and Results Information Submission (42 CFR Part 11), the law generally includes interventional studies (with one or more arms) of FDA regulated drugs, biological products, or devices that meet one of the following conditions: the trial has one or more sites in the United States, or it is conducted under an FDA investigation and finally if the trial involves a drug, biologic, or device that is manufactured in the United States or its territories and is exported for research[37,38].

Now is the time for the next milestone in open clinical trial science, which will include the creation of one exclusive, domain-specific repository that will host raw data from clinical trials, with the strong support of the legislation, and publishing community.

Currently, the two most promising existing initiatives considering data sharing are ClinicalStudyDataRequest (CSDR)[39] and Vivli[40]. Both initiatives provide data on request, the request is processed, and a decision is made whether or not one is eligible to view the data. If the access to data is granted, one must sign a Data Use Agreement, which differs depending on the company holding the data; the agreement in some way restricts the use of data and publication of finding. Both initiatives are funded by pharmaceutical companies. The good thing regarding these initiatives is relative access to some individual participant data.

Strom *et al.* have described their experience with CSDR. As members of independent review panel deciding about requests for use of those data, Strom *et al.* reported on the first 2 years of applications for access to data from 3049 trials that were available through the website. Of the 177 research proposals that were submitted, the majority was granted, and of those only four reports were published by October 2016. Strom *et al.* suggest that this could indicate inefficiency of the approval process behind CSDR[39].

In our previous research (Gabelica *et al.*; unpublished data), we screened 1700 (re3data) repositories and found that individual participant data can be found in public repositories such as Dryad, Zenodo, OSF, B2share, Edinburgh data share, Easy/DANS, ICPSR, LSHTM Data Compass, SND, DRUM and University of Bath Research Data Archive. The only repository that was specifically created for hosting raw data from clinical trials was University Hospital Medical Information (UMIN)[31] Center's Individual Case Data Repository (ICDR) within the University of Tokyo Hospital; however, the UMIN repository was not open access; it is open only to researchers from Japan who previously registered their trials in it.

In 2016., Goldacre and Gray published a manuscript in which they described the creation of an open database for hosting all data and documents threaded together by an individual trial[41]. Their OpenTrials database has been created online and is available at the URL https://opentrials.net/. However, the way they envisaged their database has multiple serious issues that will not make it sustainable.

Firstly, there is an issue of funding. According to the information on the database web site, they secured funding for the first phase of the project, which allowed them to create a "practical data schema"[42]. Secondly, OpenTrials proposes web scraping as a method for populating the database. Web scraping, by definition, is the extraction of large amounts of data from websites. For the Open Trials database, web scraping will be done automatically with the help of software. However, certain web sites have barriers that will disable such work. Alternatively, manual web scraping is an option, but that is a very tedious work, which requires additional knowledge on what to scrape[41,43].

The third issue of the OpenTrials database is an expectation about crowdsourced curation. This sourcing model relies on a large network of internet users to participate in data curation, specifically clinical trial data. It is hard to believe that a significant number of people will be available for curating data from such a broad and complicated area of science[44].

It is also unclear how OpenTrials would match the data about the same trial from various sources. Goldacre and Grey wrote in their manuscript that they will "record linkage"[41], and on the OpenTrials web site, they write that this is "area of ongoing work and research"[45]. Thus, it is unclear how they planned to record linkage. Presumably, the idea is to connect deposited data with registered protocols and published manuscripts with results, but this is not clearly indicated, nor are methods to achieve so. Simple linking of records would have to be done manually, and it could not be automated, which would be a major disadvantage of such an approach.

Most importantly Goldacre and Grey propose against hosting IPD to protect patient privacy. However, the anonymity of data is a technical issue that is simple to solve. Planning of the ideal repository must consider full anonymization of the IPD's data under EU GDPR act[28]. Without the availability of IPD, there is no open science, and there will be a limited possibility for reuse and reanalysis of clinical trial data. Strom *et al*. state that meta-analysis of firewalled patient-level data from multiple sources is a grievous endeavour. Open access to data on a dedicated repository is obvious solution[46].

Without proper legislation, manual curation, automation of selected processes, regulation and sustainable funding, it is possible that OpenTrials may not fulfil expectations. Multiple such attempts were made before. One example is OneRepo, which aimed to solve the problem of institutional repositories fragmentation, and open access to scholarly articles[47]. OneRepo was described as a project whose aim is to "unify the world's green and gold open-access works" by providing a single access point for searching all of the world's institutional repositories[48]. However, it does not look like the research community is taking any notice. As of November 2017, there is no single scholarly article available in major indexing databases about OneRepo. In November 2017, personal communication with the OneRepo's founder Mike Taylor indicated that development of OneRepo has been halted due to funding issues.

An ideal central repository for hosting data from clinical trials should be modelled as a governmental database, such as DNA banks for convicted criminal offenders, fingerprint databases, bank account information, civil registration systems, land and property records, judicial records or vehicle information records[49]. In the case of an ideal repository, this would be a transgovernmental organization, such as INTERPOL[50]. The reason for this is that raw information from clinical trial data is too important to be insecurely funded or improperly curated.

We hope that relevant stakeholders will strive to create a repository which we idealistically propose. While such repository may not be perfect initially, it can be perfected over time, for the reasons that Strom *et al.* elaborated: "*Making trial data broadly available is ethically imperative and scientifically justified and has the potential to increase public understanding of and support for clinical research. But it seems critical to find ways to improve the use and output of data-sharing projects before the clinical research community invests the substantial effort and resources required to broaden the effort to include academic and other non-commercial investigators*"[46].

With this manuscript, we hope to foster further activities to reach a consensus of a wider research community about the suggested features of an ideal repository. We have suggested features that we consider important, but the wider community could likely suggest other features that would be important, and that some of the features we suggested could be trimmed. We are not suggesting a crude approach where everyone will simply have to deposit their full data, but a creation of an ideal repository, which will also address concerns regarding the possible re-identification of patients.

The novel part of our ideal repository is strictly defined technical aspect of deposited data validation, and data curation, while implementing all aspects of FAIR data principles, which include findability, accessibility, interoperability and reusability[51].

In conclusion, we described our idea of an ideal open-access repository for clinical trial data and developed a model of such a repository using a business process analysis approach. We hope this work can inspire relevant stakeholders to engage in discussion about the necessity of creating such repository, and that we will witness the creation of such repository in near future.

Data availability

No data are associated with this article.

References

1. National Heart Lung and Blood Institute: **What Are Clinical Trials?** National Institutes of Health; U.S. Department of Health and Human Services.
Reference Source

2. **Empty rhetoric over data sharing slows science.** *Nature.* 2017; **546**(7658): 327.
PubMed Abstract | Publisher Full Text

3. Schmucker C, Schell LK, Portalupi S, *et al.*: **Extent of non-publication in cohorts of studies approved by research ethics committees or included in trial registries.** *PLoS One.* 2014; **9**(12): e114023.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Smith CT, Dwan K, Altman DG, *et al.*: **Sharing individual participant data from clinical trials: an opinion survey regarding the establishment of a central repository.** *PLoS One.* 2014; **9**(5): e97886.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Mello MM, Lieou V, Goodman SN: **Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing.** *N Engl J Med.* 2018; **378**(23): 2202–2211.
PubMed Abstract | Publisher Full Text | Free Full Text

6. **Repository | Definition of repository in English by Oxford Dictionaries**. Oxford University Press. 2018.
Reference Source

7. Institute of Medicine: **Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk**. Washington, DC: The National Academies Press. 2015; 304.
Reference Source

8. Banzi R, Canham S, Kuchinke W, *et al.*: **Evaluation of repositories for sharing individual-participant data from clinical studies.** *Trials.* 2019; **20**(1): 169.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Sapunar D, Grković I, Lukšić D, *et al.*: **The business process management software for successful quality management and organization: A case study from the University of Split School of Medicine.** *Acta Med Acad.* 2016; **45**(1): 26–33.
PubMed Abstract | Publisher Full Text

10. Kitchin R, Collins S, Frost D: **Funding models for Open Access digital data repositories.** *Online Inform Rev.* 2015; **39**(5): 664–681.
Publisher Full Text

11. Canham S, Ohmann C: **A metadata schema for data objects in clinical research.** *Trials.* 2016; **17**(1): 557.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Raftery J, Young A, Stanton L, *et al.*: **Clinical trial metadata: defining and extracting metadata on the design, conduct, results and costs of 125 randomised clinical trials funded by the National Institute for Health Research Health Technology Assessment programme.** *Health Technol Assess.* 2015; **19**(11): 1–138.
PubMed Abstract | Publisher Full Text | Free Full Text

13. FORCE11: **Guiding principles for findable, accessible, interoperable and re-usable data publishing version B1.0.** 2017.
Reference Source

14. Silva JRD, Ribeiro C, Lopes JC, editors: **UPData: a data curation experiment at U. Porto using DSpace**. *Proceedings of the 8th International Conference on Preservation of Digital Objects.* 2011.
Reference Source

15. Mezrich JL, Siegel E: **Storing Medical Images in the Digital Age: The Need for Universal and Technologically Appropriate Guidelines.** *J Am Coll Radiol.* 2017; **14**(6): 752–754.
PubMed Abstract | Publisher Full Text

16. Koff D, Bak P, Matos A, *et al.*: **Evaluation of irreversible compression ratios for medical images thin slice CT and update of Canadian Association of Radiologists (CAR) guidelines.** *J Digit Imaging.* 2013; **26**(3): 440–6.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Wikipedia contributors: **Digital object identifier.** Wikipedia, The Free encyclopedia. 2004 [updated 1 August 2018 03:51 UTC.
Reference Source

18. Klump J, Huber R: **20 Years of Persistent Identifiers – Which Systems are Here to Stay?** *Data Sci J.* 2017; **16**: 9.
Publisher Full Text

19. Crossref: **Content registration fees.** 2017.
Reference Source

20. Buneman P, Cheney J, Tan WC, *et al.*: **Curated databases.** *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*; Vancouver, Canada. 1376918: ACM. 2008; 1–12.
Publisher Full Text

21. Bourne PE, McIntyre J: **Biocurators: Contributors to the World of Science.** *PLoS Comput Biol.* 2006; **2**(10): e142.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Wikipedia contributors: **Digital curation.** Wikipedia, The Free encyclopedia. 2008 [updated 9 July 2018 22:59 UTC.
Reference Source

23. JISC: **Digital Curation Centre.** 2005.
Reference Source

24. Suber P: **The evidence fails to justify publishers' demand for longer embargo periods on publicly-funded research**. The London School of economics and Political Science. 2014.
Reference Source

25. NATURE: **Scientific data.** 2017.

26. Forschungsdaten.info: **Laws governing database and repository usage.**

27. ORCID: **Does an ORCID iD assure my identity?** 2018.
Reference Source

28. **Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).** *Official Journal of the European Union.* 2016; **L119**: 1–88.
Reference Source

29. Data Seal of Approval, ICSU World Data System: **CoreTrustSeal Certification Launched.** 2017.
Reference Source

30. DCMI: **DCMI Metadata Terms.** 2018.
Reference Source

31. UMIN-ICDR: **Individual Case Data Repository**. UMIN Center, the University of Tokyo Hospital 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan: University of Tokyo Hospital, University hospital Medical Information Network (UMIN) Center. 2013.
Reference Source

32. Dwan K, Altman DG, Clarke M, *et al.*: **Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials.** *PLoS Med.* 2014; **11**(6): e1001666.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Ohmann C, Canham S, Banzi R, *et al.*: **Classification of processes involved in sharing individual participant data from clinical trials [version 2; peer review: 3 approved].** *F1000Res.* 2018; **7**: 138.
PubMed Abstract | Publisher Full Text | Free Full Text

34. European Commission Directorate General for Research and Innovation: **EOSC Declaration Action List**. 2017.
Reference Source

35. Morey RD, Chambers CD, Etchells PJ, *et al.*: **The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review.** *R Soc Open Sci.* 2016; **3**(1): 150547.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Guharoy V: **Clinicaltrials.gov: Is the Glass Half Full?** *Hosp Pharm.* 2014; **49**(10): 893–5.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Tse T, Williams RJ, Zarin DA: **Reporting "basic results" in ClinicalTrials.gov.** *Chest.* 2009; **136**(1): 295–303.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Department of Health and Human Services U: **Clinical Trials Registration and Results Submission.** Document Citation 81 FR 64981,CFR 42 CFR 11 Docket number: NIH-20110003. 2017; **0925-AA55**: 64981–5157.
Reference Source

39. Unknown: **Clinical study data request: ideaPoint. Inc.**
Reference Source

40. Center M: **Vivli: The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard**. 2018.
Reference Source

41. Goldacre B, Gray J: **OpenTrials: towards a collaborative open database of all available information on all clinical trials.** *Trials.* 2016; **17**: 164.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Goldacre B, Gray J: **OpenTrials.** 2016.
Reference Source

43. Wikipedia contributors: **Web scraping: Wikipedia, The Free Encyclopedia.** 2005.
Reference Source

44. Wikipedia contributors: **Crowdsourcing: Wikipedia, The Free Encyclopedia.** 2006.
Reference Source

45. Goldacre B: **Opentrials.net/FAQ/What is your record linkage strategy.** 2017.
Reference Source

46. Strom BL, Buyse ME, Hughes J, *et al.*: **Data Sharing - Is the Juice Worth the Squeeze?** *N Engl J Med.* 2016; **375**(17): 1608–1609.
PubMed Abstract | Publisher Full Text

47. Hammer S: **One repo project description: Indexdata.com**. 2015.
Reference Source

48. Taylor M: **OneRepo**. 2015.
Reference Source

49. Wikipedia contributors: **Government database: Wikipedia, The Free Encyclopedia**. 2007; updated 24 July 2018 22:12 UTC.
Reference Source

50. Runjic L: **Transgovernmental Organisations - a new kind of international organisms.** *Zbornik radova Veleucilista u Sibeniku.* 2014; 1-2/2014: 91-108.
Reference Source

51. contributors W: **FAIR data: Wikipedia**.
Reference Source

Open Peer Review

Current Peer Review Status: ? ✗

Version 1

Reviewer Report 30 June 2021

https://doi.org/10.5256/f1000research.30977.r86953

✗

Ida Sim
Division of General Internal Medicine, University of California San Francisco, San Francisco, California, 94143-0320, USA
Rebecca Li
Harvard University, Cambridge, MA, USA

This publication proposes an ideal repository for hosting data from clinical trials. The criteria for such a repository was derived and modeled using a business analysis approach. It is not clear why a business analysis approach was utilized for this endeavor which appeared to conduct a landscape scan of current repositories. Data sharing is conducted within legal and cultural contexts that require additional considerations beyond business process management.

Introduction - The researchers indicate that the ideal repository is an “open-access” concept and not a managed access or controlled access system. A thorough discussion of the many valid reasons for sharing clinical trial data under managed access rather than open access is needed to support the claim that open access is "ideal." The IOM report laid out the reasons why a fully open access approach is theoretically ideal but not realistic.

The authors refer to the existence of 13 such open access “heterogeneous” repositories that provide access to clinical trial data alongside other types of data. However, the authors do not consider these “ideal” repositories due to their heterogeneous nature. The authors do not provide adequate justification for their criteria nor are the 13 open access repositories named.

Other criteria also appear rather subjective and to conflict with the sustainability criteria if one were to put them into practice. For example, requiring the repository to be sustainable without charging fees to contributors nor requestors but rather placing the burden onto governments of countries who do not have a line-item budget for this type of effort (which countries should support this? and what is their incentive to do so?) is unrealistic. Additionally, many of the requirements are exceedingly expensive and would require highly skilled statistical staff such as requiring two data curators per dataset and that the curators perform re-analysis and reproduce one result from the manuscript. These requirements which on the surface appear to be reasonable QA would require extraordinary resources.

Other criteria appear arbitrary without justification such as the recommendation for the Dublin core metadata initiative (DCMI).

The paper also does not review Vivli's approach, even though the authors state that "Currently, the two most promising existing initiatives considering data sharing are ClinicalStudyDataRequest (CSDR)[39] and Vivli[40]. " Vivli demonstrates a working and sustainable model of data sharing that isn't a centralized repository. Readers would be enlightened if the authors could explain how a centralized model would better meet their own criteria.

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Are arguments sufficiently supported by evidence from the published literature?
No

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
No

Competing Interests: Ida Sim is a co-founder of Vivli and sits on the Board of Directors. Rebecca Li is the Executive Director of Vivli.

Reviewer Expertise: Data-sharing

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 12 March 2021

https://doi.org/10.5256/f1000research.30977.r80412

?

Paul Grefen
School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

This opinion article proposes the establishment of a global, true open access repository for clinical trial data, which the authors label as an 'ideal repository'. The proposed repository is further described by a set of requirements to this repository and its supporting software and by a business process for depositing data in the repository.

As a researcher, I am very sympathetic to the idea of the proposal. Open data repositories are important to the advancement of any science that includes empirical data analysis - and medical sciences are among the important of these. On the other hand, however, I find the proposal in the paper not very realistic, and from that point of view, not very strong. It is easy to propose an ideal repository and to wave away most practical problems by the remark that 'they should be solved'. Hence, the contribution of the paper in its current form can be questioned. Why not focus more on a structured strategy to address the problems?

The set of requirements to the repository is long, but appears rather arbitrary for two reasons. Firstly, there is no clear indication how this set of requirements was established. In other words, many of the listed requirements do not have a proper foundation and therefore appear no more than the (arbitrary) opinion of the authors. A typical example is Feature 10 about the maximum upload size (which, by the way, is not a feature but a constraint). I suggest to explicitly map the requirements/features to the challenges that are mentioned on Page 3.

Secondly, there is no proper structure in the requirements. They are rather arbitrarily split into 'general features' and 'features related to user experience'. This lack of structure makes it impossible to judge whether this is a complete requirements specification, or in the terminology of the paper, a complete feature set. I suggest that the authors adopt a proper requirements specification structure - there are many textbooks on requirements analysis available that cater for this in the information systems domain.

It is unclear to me why the process design for the data uploading process is included in this opinion paper. In my opinion, it is far too operational for an opinion paper. The process that the authors (try to) describe is rather trivial given the purpose of the repository, so does not add much to the 'opinion value' of the paper.

Apart from the (non-)usefulness of the process model in this paper, the process model is flawed for a number of reasons:

- The model does not have a proper starting point. Please add a proper entry point to the process.

- The model has one 'dangling' end point: it is not specified what happens with a submitted data set if the 'REG Agency criteria are NOT met'. Please complete this branch of the process.

- The concepts event and activity are used inconsistently and incorrectly at multiple points. For example, 'Resolving Unprepared Data' looks like an activity, not an event. Please consult the ARIS manual for proper use.

- Even though version management of data sets is (very rightfully) part of the listed requirements, even though not labeled as such (Feature 7), it is not included in the process model at all. Neither is the deletion of data sets on owner's request.

- It is not clear what the difference is between 'IT Manager' and 'IT System Manager'.

- In the model, it is not clear to which organizations the roles belong; the use of swim lanes can solve this problem.

- Why does a 'Dataset with DOI' require a DOI as a resource in the next step?

- Why is 'email' a resource but not 'data transfer mechanism'?

Apart from this, one might ask why the use of data sets is not included in the process model. Instead of all details of the uploading process, I would rather have liked to see a high-level life cycle model that includes all phases of the data life cycle, including the use of the data.

Finally, it is not clear why the use of ARIS is stressed, as there are many tools that can be used to create high-level process models.

On the textual side:

- Please clarify all acronyms, like IPD and RCT.

- Please check your use of articles (quite some are missing).

- Spell names consistently ('Gray' vs. 'Grey')

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Partly

Are arguments sufficiently supported by evidence from the published literature?

Partly

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Information Systems

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.