# **Discoverability, Accessibility, and Availability of Full Clinical Trial Protocols: An Analysis of PubMed, ClinicalTrials.gov, and 6 Data Sharing Platforms**

# **Project Lead:**

Paije Wilson, *2018-2019 National Library of Medicine Associate Fellow*

# **Project Sponsor:**

Dr. Vojtech Huser, *Staff Scientist at the National Institutes of Health*

# **January 22, 2019**

**Table of Contents**

[**Discoverability, Accessibility, and Availability of Full Clinical Trial Protocols: An Analysis of PubMed, ClinicalTrials.gov, and 6 Data Sharing Platforms** 1](#_Toc535920188)

[**Project Lead:** 1](#_Toc535920189)

[**Project Sponsor:** 1](#_Toc535920190)

[**January 17, 2019** 1](#_Toc535920191)

[**Abstract** 3](#_Toc535920192)

[**Introduction** 4](#_Toc535920193)

[**Methods** 5](#_Toc535920194)

[Part 1: PubMed (Discoverability of Protocols) 5](#_Toc535920195)

[Part 2: ClinicalTrials.gov (Accessibility of Protocols) 6](#_Toc535920196)

[Part 3: Platform Analysis (Availability of Protocols) 6](#_Toc535920197)

[**Results** 7](#_Toc535920198)

[Part 1: PubMed (Discoverability of Protocols) 7](#_Toc535920199)

[Part 2: ClinicalTrials.gov (Accessibility of Protocols) 13](#_Toc535920200)

[Part 3: Platform Analysis (Availability of Protocols) 17](#_Toc535920201)

[**Discussion** 21](#_Toc535920202)

[Part 1: PubMed (Discoverability of Protocols) 21](#_Toc535920203)

[Part 2: ClinicalTrials.gov (Accessibility of Protocols) 21](#_Toc535920204)

[Part 3: Platform Analysis (Availability of Protocols) 22](#_Toc535920205)

[**Conclusion** 23](#_Toc535920206)

[**Works Cited** 24](#_Toc535920207)

[**Appendix 1: Narrative Version of PubMed Results** 25](#_Toc535920208)

[**Appendix 2: Narrative Version of ClinicalTrials.gov Results** 26](#_Toc535920209)

[**Appendix 3: Narrative Version of Platform Analysis Results** 27](#_Toc535920210)

# **Abstract**

**Background and Objective**

Data and design transparency have become essential in determining the legitimacy and impact of clinical trials. However, despite the increased sharing of data, many studies report that protocol discoverability and sharing have remained suboptimal. Though studies have been conducted on protocol sharing, few, if any studies have examined the discoverability of full protocols on PubMed, their accessibility on ClinicalTrials.gov, nor their general availability on data sharing platforms (i.e. websites). To fill these research gaps, this study seeks to address the following questions: 1) **How discoverable are protocols on PubMed?** 2) **How accessible are linked protocols on ClinicalTrails.gov?** and 3) **To what extent are protocols available on data sharing platforms?**

**Methods**

This study was divided into three parts: The first examined if search limitations for discovering full protocols are applicable to PubMed; the second attempted to determine the accessibility of full HIV protocols on ClinicalTrails.gov through testing a subset of its protocol links; and the third examined the availability of protocols on 6 additional data sharing platforms (Clinical Study Data Request, Project Data Sphere, dbGap, Biolincc (aka NHLBI), Nida Data Share, and Vivli).

**Results**

PubMed Search Results: The results of the study indicate that the discoverability of full protocols on PubMed is suboptimal.

ClinicalTrails.gov Results: The results indicate that there is moderate accessibility to HIV protocols linked on ClinicalTrials.gov for the casual user (with “casual user” referring to users that wish to access protocols without having to overcome request walls).

Platform Analysis Results: The results indicate that, cumulatively, there is a moderate availability of protocols on the 6 data sharing platforms tested.

**Conclusion**

The results of these analyses indicate that improvements could be made to enhance protocol discoverability on PubMed, accessibility on ClinicalTrials.gov, and availability on data sharing platforms. Though advancements have certainly been and are being made in protocol sharing, there is much work to be done to foster a truly open science environment.

# **Introduction**

**Data Sharing in Clinical Trials**

Data and design transparency have become essential in determining the legitimacy and impact of clinical trials. The sharing of such data is believed to enhance credibility and accelerate discovery through, among other things, optimizing data reuse, improving reproducibility, and facilitating external examination of the scientific process.1,2,3 Many claim there is also an ethical obligation to publicly share deidentified clinical trial data. In June of 2017, the New England Journal of Medicine (ICMJE) published a statement claiming that scientists have “an ethical obligation to responsibly share data generated by interventional clinical trials because trial participants have put themselves at risk.” 4 Similarly, in a 2017 article Bertagnolli et al. claim that “It is important to honor and reward the altruism of patients who participate in clinical trials . . . [in part, by sharing] the data gathered with other researchers in a responsible and meaningful way.” 5

**The Importance of Protocols**

One vital component of data sharing and design transparency in clinical trials is access to the full protocols of a study. Protocols report in detail on the “background, relevance, methods, administration, and ethical considerations” of a study before it is conducted, and are essential in ensuring a study is “implemented in a manner that is consistent with the research objectives and the intentions of the steering group.”6 Some of the benefits of sharing full protocols include the deterring of selective reporting (as researchers are required to state statistical and analytical approaches prior to collecting data and justify any changes they make during the implementation of these approaches); preventing unnecessary study duplication; informing new methodologies; enhancing confidence in the results and conclusions made in studies; and promoting greater understanding of trial data among researchers and the public. 2,7 The sharing of full protocols also facilitates study reproducibility, as external researchers may access complete documentation of the trial’s methodologies and more easily analyze the internal validity of the study.6,8

However, despite their importance in clinical reproducibility and inquiry, the sharing and discoverability of full clinical trial protocols has been relatively limited. According to a 2017 study conducted by Sutton et al., though there have been improvements in reporting guidelines and trial registration in clinical trials, the “availability and retrieval [of protocols] remain suboptimal.”7 This trend is persistent even in prominent medical journals with strong data sharing policies. According to a 2018 study conducted by Naudet et al, a survey of randomized clinical trails published in BMJ and PLOS Medicine found that only 46% of studies met eligibility criteria for sharing their data “with sufficient information to enable reanalysis.” 2 The reluctance to share full protocols may stem from multiple factors, including competitive forces (both between academic researchers and pharmaceutical companies, which results in a reluctance to share their intellectual property before it is published); lack of financial or professional incentive to share full protocols; fear of participant “unblinding” during a trail (i.e. participants uncovering the study blind during the trial through accessing the published protocol); and the potential for “naming and shaming” investigators with unreproducible or faulted studies, which may be uncovered as a result of examining full protocols.2,5

**Research Gaps**

While studies such as that conducted by Sutton et al. provide a comprehensive look into the searching of protocols through various methodologies, few, if any, studies examine the findability of full protocols on PubMed, their accessibility on ClinicalTrials.gov, nor their general availability on data sharing platforms. These are substantial information gaps, as PubMed houses over 28 million citations for biomedical literature and ClinicalTrails.gov contains over 293,000 studies spanning 207 countries.9,10 To fill these research gaps, this study was divided into three parts: The first examined if search limitations for discovering full protocols are applicable to PubMed; the second attempted to determine the accessibility of full protocols on ClinicalTrails.gov through testing a subset of its protocol links; and the third examined the availability of protocols on 6 additional data sharing platforms (Clinical Study Data Request, Project Data Sphere, dbGap, Biolincc (NHLBI), Nida Data Share, and Vivli).

It should be noted that both PubMed and ClinicalTrials.gov have made motions to enhance protocol discoverability and sharing. PubMed has announced MeSH will be releasing new subject headings in 2019, including the addition of “Clinical Trial Protocol” as a Publication Type (PT). Though this new term has the potential to significantly enhance the findability of new protocols on PubMed, the addition of the term will not address the issue of titles indexed before 2019.11 ClinicalTrails.gov has also implemented the “Final Rule” from the Federal Register, wherein clinical trials are required to share full protocols upon their registration to ClinicalTrials.gov.12,13 Similar to PubMed, however, the rule will only apply to trials registered after January of 2017. The wealth of information that can be acquired from these past protocols, in addition to the potential to examine extent of compliance to the Final Rule, further justifies the necessity of the study.

# **Methods**

## Part 1: PubMed (Discoverability of Protocols)

3 PubMed searches were conducted, with the top 20 results from each search (totaling 60 results) being examined for false positives (i.e. whether a result was not a protocol) and page length (see **Table 1** and **Figures 1** through **6**). These data were acquired to estimate the rough precision of each search and determine the “fullness” of the resulting protocols (as full protocols should exceed 30 pages in length, as indicated by the NIH-FDA Phase 2 and 3 Clinical Trial Protocol Template, which is 65 pages in length, by itself)).14

The source journals of the top 10 longest protocols were then analyzed for navigability (i.e. relative ease of locating protocols on the site), average protocol length, and availability of different protocol formats (e.g. PDFs, XMLs, etc.) (see **Tables 2** and **4**). The latter two characteristics were taken from a subset of 10 of the most recent protocols from each journal website (totaling 60 tested protocols). For general comparison, a 3-star scoring system was implemented for each of the tested characteristics (see **Table 3** for a key to the 3-star measure). These data served to evaluate the discoverability of protocols in following source journal linkages on PubMed (i.e. finding additional protocols indirectly through PubMed results by looking up the results’ source journals).

All of the above data were recorded into an Excel file, and all visualizations were produced in Excel or RStudio. Tables were created in Microsoft Word (Both Microsoft Word and Excel versions came from the Office 365 ProPlus package). The data served to evaluate the discoverability of full protocols on PubMed.

## Part 2: ClinicalTrials.gov (Accessibility of Protocols)

A subset of 9,857 ClinicalTrials.gov HIV research protocol links (provided by Dr. Huser and his team) was analyzed. This data was used because of its large size and the fact that trail registrations numbers and links had already been extracted, but not tested, by Dr. Huser and his team. Due to these conveniences, the data was chosen to serve as a representative sample of ClinicalTrials.gov protocols. From the data, the top 7 platforms (being platforms that contained the greatest number of protocol links on the ClinicalTrials.gov data subset) were extracted (see **Table 5** and **Figure 7**). A sample of 10 links per platform (totaling 70 links) were then tested for positive links (i.e. whether the linked protocols were found); path (i.e. “direct” if the link resulted in a direct download of the protocol; “linked” if the link leads to an additional page with a link to the protocol; or “indirect” if additional searching was needed to locate the protocol); available formats for the linked protocols; and the average page length of the linked protocols (see **Table 6**).

For visual comparison, a scatterplot was constructed to better compare the general page ranges of protocols from each platform (see **Figure 8**), and a 3-star scoring system was implemented to score the positive links, paths, formats, and average page lengths of linked protocols from each platform, with 1 star being a low score and 3 stars being a perfect score (see **Table 7** for a key to the 3-star scoring and **Table 8** for the scoring results).

All of the above data were recorded into an Excel file, and all visualizations were produced in Excel or RStudio. Tables were created in Microsoft Word. The data served to roughly evaluate the accessibility of HIV protocols linked on ClinicalTrials.gov.

## Part 3: Platform Analysis (Availability of Protocols)

6 additional data sharing platforms were examined in-depth for size (i.e. number of studies on the platforms’ site) protocol sharing, data sharing, availability of data dictionaries, data format, protocol format, extent of request walls, policies, and general navigability. To do this, 10 of the most recent clinical trial studies for each platform were examined (see **Tables 9** through **11**, and **Figures 9** and **10**). The tested platforms included Clinical Study Data Request, Project Data Sphere, dbGap, Biolincc (NHLBI), Nida Data Share, and Vivli.

The availability of protocols, data, and data dictionaries were quantified to evaluate how protocol availability compares with data and data dictionary availability, providing insight into which platforms include protocols as a necessary component of data sharing. The data were based on evidence of presence and not access (i.e. if there is evidence of the documents being present in a study’s materials, even if they are inaccessible due to request walls, they are still counted as being present).

Formats were evaluated to determine the relative interoperability of the information (i.e. the greater the number of formats, the greater the chances that the materials are interoperable). The presence of links was measured to evaluate the discoverability of information connected to the studies on the platforms (i.e. studies with more links should increase accessibility to relevant study information).

Request walls were measured to evaluate the relative difficulty researchers may experience when attempting to access study information (the more intensive the information required, the greater the relative difficulty of accessing the information). The term “request walls” here refers to methods of restricting access to content after a request is created (and sometimes approved by the platform’s request review committee and/or data holder). Requests vary in complexity, ranging from simple project descriptions (via webform or stand-alone document) to full research project proposals that have local IRB approval. Complex requests may also include signatures of an institution’s signing official. Request walls with complex request requirements can significantly delay data retrieval.

Navigability was a qualitative measure. The relative facility of navigating a platform was based solely on the experience of the author of this study. This measure served to evaluate the discoverability of information on each of the platforms.

To facilitate cross-platform comparisons, the 3-star scoring system was implemented, with 1 star being a low score and 3 being a high score (see **Table 12** for a key to the 3-star scoring and **Table 13** for scoring results).

All of the above data were recorded into an Excel file, and all visualizations were produced in Excel or RStudio. Tables were created in Microsoft Word. The data collected from each of these platforms served as indicators for the availability of full protocols in databases that were created to enhance reproducibility and transparency in research.

# **Results**

## Part 1: PubMed (Discoverability of Protocols)

For a narrative version of the results, please see **Appendix 1**.None of the 3 PubMed searches yielded full protocol results; only article or summary versions of the protocols (i.e. protocol documents less than 30 pages in length) were found (see **Table 1**).

**Table 1:** General Results Summary for Tested Searches

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Search** | **MeSH** | **Number of Results** | **False Positives** | **Precision (P = positives / total results tested)** |
| **Search 1** | "Protocol"[TI] AND "Clinical Study"[Publication Type] AND "Research Design"[Mesh] | 4,340 | 1 result / 20 tested | 95% |
| **Search 2** | (("Clinical Protocols"[MAJR]) AND "Humans"[MeSH Terms]) AND "Research Design"[MeSH Terms] AND "Protocol"[TI]) | 327 | 3 results / 20 tested | 85% |
| **Search 3** | “Protocol” [ti] AND “clinicaltrials.gov”[si] | 4,043 | 8 results / 20 tested | 60% |

**Table 1** shows a general summary of the tested searches, including the MeSH terms that were used, the number of results each search retrieved, the number of false positives that were identified in a sample of 20 results (i.e. results that were not protocols, long or otherwise), and the relative precision of each search.

The positive protocol results for each search were then tested for page length (see **Figures 1-6**).

**Figure 1:** Bar chart of Search 1 results that were marked positive for being protocols.

**Figure 2:** Pie chart of Search 1 results that were marked positive for being protocols. Colors and percentages are indicative of the proportion of results within the specified page ranges.

**Figure 3:** Bar chart of Search 2 results that were marked positive for being protocols.

**Figure 4:** Pie chart of Search 2 results that were marked positive for being protocols. Colors and percentages are indicative of the proportion of results within the specified page ranges.

**Figure 5**: Bar chart of Search 3 results that were marked positive for being protocols.

**Figure 6:** Pie chart of Search 3 results that were marked positive for being protocols. Colors and percentages are indicative of the proportion of results within the specified page ranges.

Of these results, the journal sources of the top 10 longest protocol articles were extracted and analyzed. These journals included *BMC Psychiatry*, *BMJ Open*, *JMIR Research Protocols*, and *Trials*. (see **Table 2**).

**Table 2**: Top 10 Longest Protocol Articles and their Source Journals

|  |  |
| --- | --- |
| **Protocol Article Page Length** | **Source Journal** |
| **26** | **JMIR Research Protocols** |
| **22** | **JMIR Research Protocols** |
| **18** | **BMJ Open** |
| **18** | **Trials** |
| **15** | **BMC Psychiatry** |
| **15** | **BMJ Open** |
| **14** | **Trials** |
| **13** | **Trials** |
| **12** | **Trials** |
| **12** | **Trials** |

**Table 2** shows the top 10 longest article protocols’ page lengths from the 3 searches, accompanied by the journals from which they derive.

The resulting 4 journals were then analyzed for site navigability, average protocol length, and available protocol formats (see **Table 3** and **4**).

**Table 3:** Key to the 3-star measure for source journal analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **Score** | **Site Navigability** | **Average Protocol Length** | **Available Formats** |
|  | **Low navigability** | **< 15 pages** | **1** |
|  | **Moderate navigability** | **> 15 pages but < 30 pages** | **2** |
|  | **Excellent navigability** | **> 30 pages** | **> 3** |

**Table 3** shows a key for the 3-star measure that was implemented for the source journal analysis. A perfect score in navigability, average protocol length, and available formats would equate to 9 stars.

**Table 4**: Source journal analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Journal** | **Site Navigability** | **Average Protocol Length** | **Available Formats** | **Total** |
| **JMIR Research Protocols** |  |  |  | **7/9** |
| **BMJ Open** |  |  |  | **7/9** |
| **Trials** |  |  |  | **5/9** |
| **BMC Psychiatry** |  |  |  | **5/9** |

**Table 4** shows the analysis of the source journals of the top 10 longest protocols. The journals were scored for site navigability, average protocol length, and available protocol formats. Average protocol length and available formats were based on samples of the 10 most recent protocols from each journal site.

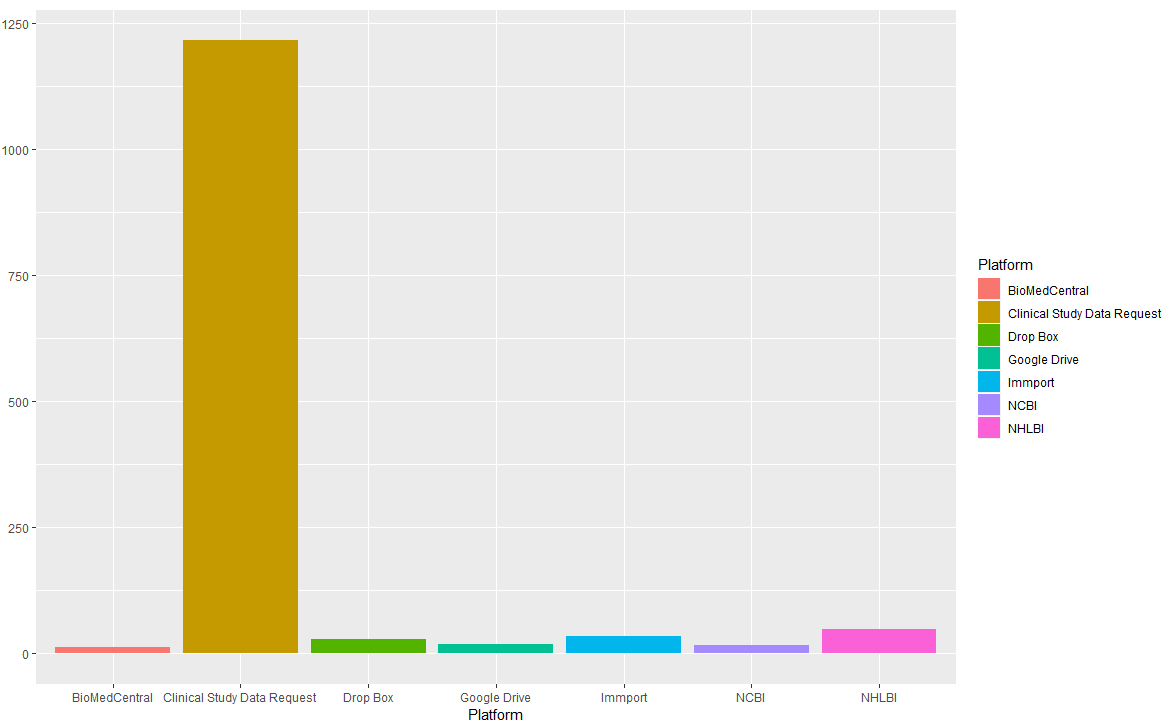
## Part 2: ClinicalTrials.gov (Accessibility of Protocols)

The top 7 linked source platforms from the ClinicalTrials.gov data subset included Clinical Study Data Request (1,215 links), NHLBI (Biolincc) (48 links), Immport (35 links), Drop Box (28 links), Google Drive (19 links), NCBI (17 links), and BioMedCentral (12 links) (see **Table 5** and **Figure 7**). For a narrative version of the results, please see **Appendix 2**.

**Table 5**: Top Source Platforms from ClinicalTrials.gov

|  |  |
| --- | --- |
| Platform | Links Found |
| BioMedCentral | 12 |
| Clinical Study Data Request | 1215 |
| Drop Box | 28 |
| Google Drive | 19 |
| NCBI | 17 |
| NHLBI | 48 |
| Immport | 35 |

**Table 5** shows platforms that had the greatest number of links in the ClinicalTrials.gov data subset.



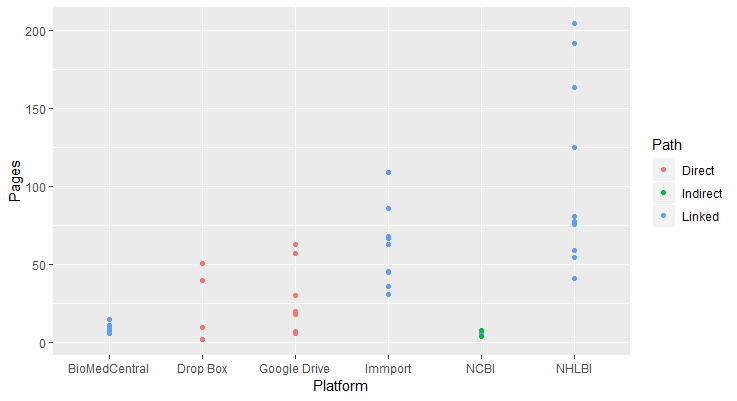
**Figure 7:** Bar graph of **Table 5** results. Shows the number of links per platform in the ClinicalTrials.gov data subset.

From each of these platforms, 10 random ClinicalTrials.gov protocol links (totaling 70 links) were selected and tested for positive links, path, available formats, and page length (see **Table 6**, **Figure 8**, and **Tables 7** and **8**).

**Table 6**: Linked platform analysis data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Platform | Positive Links | Path | Available Formats | Average Protocol Length (pages) |
| BioMed Central (BMC) | 10 | Linked | PDF | 9.3 |
| Clinical Study Data Request | 0 (permission barrier) | NA | NA | NA |
| Drop Box | 4 | Direct | PDF | 25.75 |
| Google Drive | 7 | Direct | PDF | 28.71 |
| Immport | 10 | Linked | PDF | 61.9 |
| NCBI | 4 | 3 Indirect and 1 Linked | PDF | 6.5 |
| NHLBI (Biolincc) | 10 | Linked | PDF | 107.6 |

**Table 6** shows the analysis of the top platforms linked in the ClinicalTrials.gov data subset. The data were acquired from 10 random ClinicalTrials.gov protocol links that were attributed to each platform.



**Figure 8**: Scatterplot of linked platform analysis data.

**Table 7:** Key to the 3-Star measure for linked platform analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Score** | **Positive Links** | **Path** | **Available Formats** | **Average Protocol Length** |
|  | **< 5** | **Indirect** | **1** | **< 15 pages** |
|  | **> 5 but < 10** | **Linked** | **2** | **> 15 pages but < 30 pages** |
|  | **10** | **Direct** | **> 3** | **> 30 pages** |

**Table 7** shows a key for the 3-star measure that was implemented for the linked platform analysis. 0 stars would indicate null values where the protocols were not found. A perfect score would equate to 12 stars.

**Table 8**: Linked platforms analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Platform** | **Positive Links** | **Path** | **Available Formats** | **Average Protocol Length** | **Total** |
| **BioMedCentral** |  |  |  |  | **7/12** |
| **Clinical Study Data Request** |  |  |  |  | **0/12** |
| **Drop Box** |  |  |  |  | **7/12** |
| **Google Drive** |  |  |  |  | **8/12** |
| **Immport** |  |  |  |  | **9/12** |
| **NCBI** |  |  |  |  | **4/12** |
| **NHLBI** |  |  |  |  | **9/12** |

**Table 8** shows the analysis of the top platforms linked in the ClinicalTrials.gov data subset. The data were acquired from the 10 random ClinicalTrials.gov protocol links that corresponded with each platform (see **Table 6** for a non-visual version of the data).

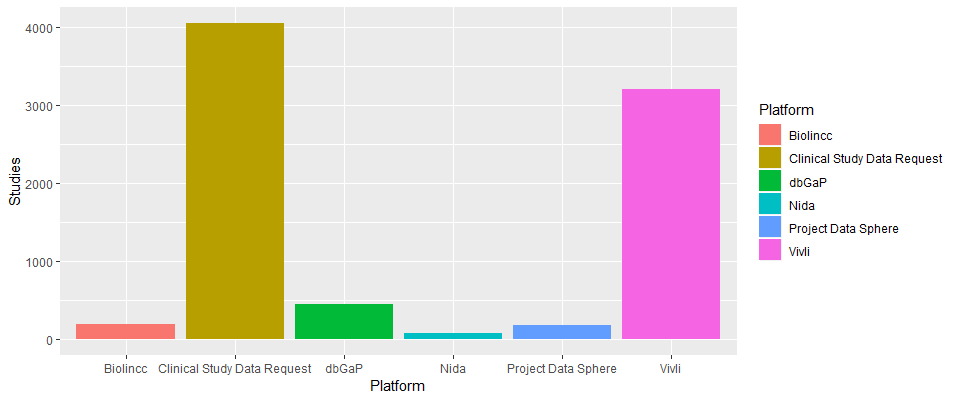
## Part 3: Platform Analysis (Availability of Protocols)

Clinical Study Data Request was the largest of the platforms, with 4,051 estimated studies; Vivli had 3,200 studies; dbGaP 446 studies; Biolincc 200 studies; Project Data Sphere 183 studies; and Nida 79 studies (see **Table 9** and **Figure 9**). For a narrative version of the results, please see **Appendix 3**.

**Table 9:** Platform Sizes

|  |  |
| --- | --- |
| Platform | Size (Estimated Number of Studies) |
| Clinical Study Data Request | 4051 |
| Vivli | 3200 |
| dbGaP | 446 (searching “e”) |
| Biolincc (NHLBI) | 200 |
| Project Data Sphere | 183 |
| Nida | 79 |

**Table 9** shows the estimated size of each of the platforms (by the number of studies they have present in their databases). If the study number was not present in their statistics, an estimated study count was acquired through searching “e” in the platform’s search bar.

**Figure 9**: Visualization of platform sizes (by number of studies within each platform).

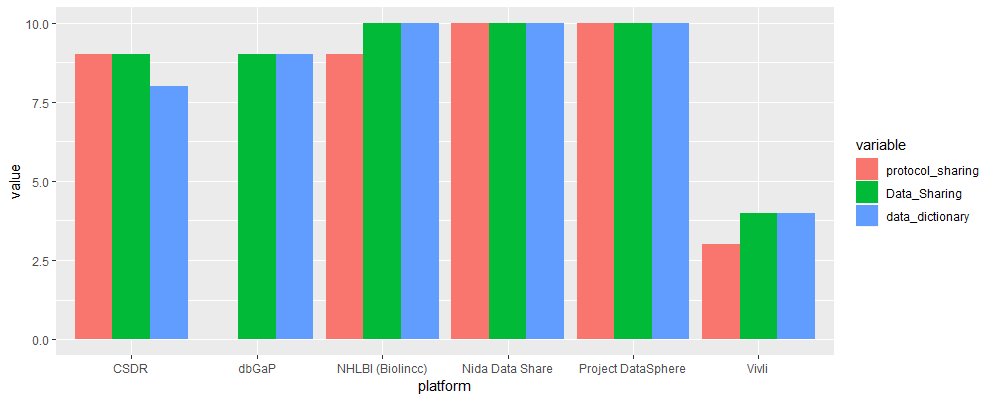
Each of the platforms were then analyzed for protocol availability, data availability, data dictionary availability, data formats, protocol formats, and available links (i.e. to trial registries). Policies found on each of the platform sites were also noted. A “0” or “NA” value indicates that evidence of the material was not found; this does not mean that the material does not exist, merely that a link to the material (even if the link does not grant access) could not be located (see **Table 10** and **Figures 10**). Request walls and navigability were also analyzed for each of the platforms, and a 3-star scoring system was implemented for visual comparison (see **Tables 11** through **13**).

**Table 10**: General Platform Analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Platform** | **Protocol Availability** | **Data Availability** | **Data Dictionary Availability** | **Data Formats** | **Protocol Formats** | **Links** | **Policies** |
| **Clinical Study Data Request** | 9/10 | 9/10 | 8/10 | SAS | NA | ClinicalTrials  EudraCT  Sponsor registry links | DSA  LAP |
| **dbGaP** | 0/10 | 9/10 | 9/10 | Image | NA | ClinicalTrials BioProject  BioSample  PMC  PubMed  Not all studies have links | CC  DUA  SBPP |
| **Nida data Share** | 10/10 | 10/10 | 10/10 | CRF  SAS  CSV | PDF | ClinicalTrials | DUA |
| **Project Data Sphere** | 10/10 | 10/10 | 10/10 | CSV  Zip | PDF | ClinicalTrials | DUA |
| **Vivli** | 3/10 | 4/10 | 4/10 | Image | PDF | ClinicalTrials | DUA |
| **NHLBI (Biolincc)** | 9/10 | 10/10 | 10/10 | NA | PDF | ClinicalTrials | DSA |

**Table 10** shows the protocol availability, data availability, data dictionary availability, available data formats, available protocol formats, links (i.e. to registries), and policies found for each of the platforms.

Key for Policies: DSA (data sharing agreement); LAP (limited access period); CC (code of conduct); DUA (data use agreement), SBPP (security best practices policy)



**Figure 10**: Comparative visualization of platforms’ sharing of protocols, data, and data dictionaries

**Table 11:** Platform Request Walls

|  |  |
| --- | --- |
| Platform | Request Walls |
| Clinical Study Data Request | -Account (basic)  -Project Proposal (extensive; numerous text fields) |
| dbGaP | -Institutional account (eRA)  -PI account (basic)  -Project proposal (extensive) |
| Nida | -No account is required  -Data use agreement (basic) for data  -**Other documentation accessible without barriers**. |
| Project Data Sphere | -Data request (extensive)  -Account given to approved requests |
| Vivli | -Account (basic)  -Project proposal (extensive)  -Some documentation available on clinicaltrials.gov. |
| Biolincc (NHLBI) | -Account (extensive)  -Project proposal (extensive) for data  -**Other documentation accessible without barriers** |

**Table 11** shows the request walls present on each of the platforms, and the extent of information needed to bypass these walls (“basic” referring to minimal and “extensive” to detailed information requirements).

**Table 12**: Key for 3-star measure for general platform analysis

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Score** | **Protocol Availability** | **Data Availability** | **Data Dictionary Availability** | **Data**  **Formats** | **Protocol**  **Formats** | **Links** | **Request Walls** | **Navigability** |
|  | **< 5 results** | **< 5 results** | **< 5 results** | **1** | **1** | **1** | **Extensive** | **High Difficulty** |
|  | **> 5 but < 10 results** | **> 5 but < 10 results** | **> 5 but < 10 results** | **2** | **2** | **2** | **Moderate** | **Moderate Difficulty** |
|  | **10 results** | **10 results** | **10 results** | **> 3** | **> 3** | **> 3** | **Low** | **Easy** |

**Table 12** shows a key for the 3-star measure that was implemented for the general platform analysis. 0 stars would indicate null values where the documents or formats were not found. A perfect score would equate to 27 stars.

**Table 13**: 3-star scoring for general platform analysis

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Platform** | **Protocol Availability** | **Data Availability** | **Data Dictionary Availability** | **Data**  **Formats** | **Protocol**  **Formats** | **Links** | **Request Walls** | **Navigability** | **Total** |
| **Clinical Study Data Request** |  |  |  |  |  |  |  |  | **14/27** |
| **dbGaP** |  |  |  |  |  |  |  |  | **10/27** |
| **Nida Data Share** |  |  |  |  |  |  |  |  | **20/27** |
| **Project Data Sphere** |  |  |  |  |  |  |  |  | **18/27** |
| **Vivli** |  |  |  |  |  |  |  |  | **10/27** |
| **NHLBI (BioLincc)** |  |  |  |  |  |  |  |  | **15/27** |

**Table 13** shows the analysis of 6 data sharing platforms. The platforms were scored for protocol sharing, data sharing, data dictionary sharing, available data formats, available protocol formats, links, request walls, and navigability (see **Tables 10** and **11** for textual versions of the data).

# **Discussion**

## Part 1: PubMed (Discoverability of Protocols)

**Research Question: How discoverable are protocols on PubMed?**

The results of the PubMed searches largely reflect the statement in the 2017 study of Sutton et al., in that full protocol retrieval and availability remains suboptimal.7 Though the searches succeeded in retrieving articles or summaries of protocols, full length protocols (i.e. that exceeded 30 pages) were not found in any of the PubMed searches. Analyzing the source journals of the lengthiest protocol articles/summaries found on PubMed yielded similar results, with most journals having an average protocol length of less than 15 pages. Only BMJ Open had full protocols that were found on their site.

Beings that PubMed is a leading repository for biomedical literature, it is concerning that the discoverability of full protocols was suboptimal in the experiment, as protocols play a significant role in data reusability and reproducibility in clinical trials research. Though the introduction of the “protocol” publication type will improve the discoverability of new protocols, the findings of this experiment emphasize the fact that full protocols published before 2019 will have poor discoverability unless retrospectively indexed with the new term, resulting in a substantial loss in clinical trials literature.

## Part 2: ClinicalTrials.gov (Accessibility of Protocols)

**Research Question: How accessible are linked protocols on ClinicalTrails.gov?**

The results indicate that there is moderate accessibility to HIV protocols linked on ClinicalTrials.gov for the casual user.

A majority of the protocol links on ClinicalTrails.gov led to 7 platforms, and there was roughly a 64% success rate of accessing a protocol in the subset of 70 links that were tested. However, the situation becomes more alarming when one generalizes the data. Most links on ClinicalTrials.gov (1,215) lead to Clinical Study Data Request, which, if the representative sample for this study can be generalized, are likely inaccessible to the casual user due to permission barriers. If this reasoning is accurate, a sizable chuck of ClinicalTrails.gov protocols are inaccessible to the casual user.

The paths received a moderate scoring. Most of the linked platforms used linked or direct paths, which greatly enhanced discoverability of the protocols. The available formats and average page lengths of the linked protocols, however, were suboptimal, with the only format available being PDF (decreasing interoperability) and only 2 of the linked platforms (Immport and NHLBI) having protocols with an average page length that exceeded 30 pages. The latter problem is similar to that shown in the PubMed searches, in that summaries or articles of protocols are more prevalent than full-length protocols. This is of concern, as summaries/articles of protocols are less likely to provide the breadth and depth in methodologies needed to support reproducibility and data reusability.

Though adequate, the accessibility of linked protocols on ClinicalTrials.gov could be improved. Standards in permission barriers, protocol page lengths, and, perhaps, an increase in available formats could greatly contribute to protocol access, thoroughness, and interoperability.

## Part 3: Platform Analysis (Availability of Protocols)

**Research Question:** **To what extent are protocols available on data sharing platforms?**

The results indicate that there is a moderate availability of protocols on the cumulative data sharing platforms tested. It also appears that protocol availability closely matches the extent of data and data dictionary availability on the platforms that were examined (with dbGaP being the exception). This finding is promising, as it indicates that data sharing platforms view protocol availability as a necessary component of open data. However, like ClinicalTrails.gov, many improvements can be made to enhance protocol availability, especially in terms of the availability of protocols in larger platforms; number of formats; request walls; and, in the case of dbGaP, navigability.

Of the platforms that were examined, Nida Data Share and Project Data Sphere had the highest scoring for the platform analysis, and consistently shared protocols, datasets, and data dictionaries with minimal (in the case of Nida Data Share) or moderate (in the case of Project Data Sphere) information requirements from the investigator. Of concern is the fact that Nida Data Share and Project Data Sphere are also the smallest of the platforms that were tested. If the smaller platforms’ quality of data sharing could be somehow applied to the larger platforms a great wealth of data could be made accessible for reanalysis and discovery.

Though the number of formats varied for datasets, available formats for protocols remained low (with the only verifiable format being PDF). This reduces the interoperability of protocols. Creating more format options could resolve this issue.

dbGaP and Clinical Study Data Request scored highest for the number of external links provided in their studies (i.e. that linked externally to additional information for the study, such as trial registries or articles). Other platforms only provided links to ClinicalTrials.gov registries. While additional linkages would be beneficial in enhancing connectivity of study information, the linkages each platform provided may be considered sufficient due to the wealth of summative information provided in the ClinicalTrails.gov registries.

Interestingly, it appears that request walls have a positive relationship with platform size (i.e. the greater the size of the platform, the more extensive permission barriers they have in place). Though some platforms allowed access to protocols and data dictionaries without request walls (i.e. Nida Data Share and Biolincc), the largest platforms (i.e. Clinical Study Data Request and Vivli) required extensive project proposals in order to access background documentation (though, for Vivli, this could sometimes be bypassed by searching for the study protocols on ClinicalTrials.gov). This effectively bars access for casual users and restricts accessibility to a great wealth of data. Though a larger analysis would need to be conducted to confirm this observation, the relationship makes sense, as the greater the amount of data a platform has, likely the greater the need to protect the identities of participants in the studies. However, this raises the question of whether such restrictions should be applied to background documentation like protocols. Is there enough personal identifiable information in a protocol to warrant request walls? And does the public (i.e. casual user) have the right to access these protocols?

All of the platforms exempting dbGaP received high scores for navigability, which enhanced discoverability of available materials. On the flip side, dbGaP was considerably difficult to navigate for an inexperienced user, making for poor discoverability and a rather frustrating search experience. Applying more interactive graphic user interfaces and standardizing the labeling of downloadable documents may help with this problem, and greatly enhance the discoverability of available materials.

# **Conclusion**

The results of these analyses indicate that improvements could be made to enhance protocol discoverability on PubMed, accessibility on ClinicalTrials.gov, and availability on data sharing platforms. Though advancements have certainly been and are being made in protocol sharing, there is much work to be done to foster a truly open science environment.

# **Works Cited**

1. Munafò, Marcus R., et al. "A manifesto for reproducible science." *Nature Human Behaviour* 1.1 (2017): 0021.
2. Naudet, Florian, et al. "Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine." *bmj* 360 (2018): k400.
3. Huser, Vojtech, and Dikla Shmueli-Blumberg. "Data sharing platforms for de-identified data from human clinical trials." *Clinical Trials* (2018): 1740774518769655.
4. Taichman, Darren B., et al. "Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors." *Annals of internal medicine* 167.1 (2017): 63-65.
5. Bertagnolli, Monica M., et al. "Advantages of a truly open-access data-sharing model." (2017): 1178-1181.
6. Chan, An-Wen, and Asbjørn Hróbjartsson. "Promoting public access to clinical trial protocols: challenges and recommendations." *Trials* 19.1 (2018).
7. Sutton, Anthea, et al. "Searching for trial protocols: A comparison of methods." *Research synthesis methods* (2017).
8. Nichols, Thomas E., et al. "Best practices in data analysis and sharing in neuroimaging using MRI." *Nature Neuroscience* 20.3 (2017).
9. “PubMed Help*.” PubMed.* U.S. National Library of Medicine, https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.FAQs. Accessed 1 Jan. 2019.
10. *ClinicalTrails.gov*. U.S. National Library of Medicine, https://clinicaltrials.gov/. Accessed 1 Jan. 2019.
11. “What’s New in MeSH.” *Medical Subject Headings.* U.S. National Library of Medicine (2019), https://www.nlm.nih.gov/mesh/whatsnew.html.
12. “FDAAA 801 and the Final Rule.” *ClinicalTrials.gov*. U.S. National Library of Medicine (2018), https://clinicaltrials.gov/ct2/manage-recs/fdaaa. Accessed 18 Dec. 2018.
13. “Clinical Trials Registration and Results Information Submission.” *Federal Register*. U.S. National Archives and Records Administration (2016), https://www.federalregister.gov/documents/2016/09/21/2016-22129/clinical-trials-registration-and-results-information-submission. Accessed 18 Dec. 2018.
14. “NIH-FDA phase 2 and 3 ind/ide clinical trial protocol template.” U.S. National Institutes of Health and Food and Drug Administration. 1.07 (2017).

# **Appendix 1: Narrative Version of PubMed Results**

Search 1 retrieved 4,340 results, with 1 false positive being identified out of the 20 results that were tested. The relative precision of the search was 95%. Search 2 retrieved 327 results, with 3 false positives being identified out of the 20 results that were tested. The relative precision of the search was 85%. Search 3 retrieved 4,043 results, with 8 false positives being identified out of the 20 results that were tested. The relative precision of the search was 60% (see **Table 1**)

The positive protocol results for each search were then tested for page length. Of the positive protocol results tested in Search 1, 2 (11%) were 1 to 4 pages in length, 8 (42%) were 5 to 10 pages, 8 (42%) were 11 to 15 pages, and 1 (5%) was 16 to 20 pages in length. No protocols were over 21 pages in length (see **Figures 1** and **2**). Of the positive protocol results tested in Search 2, none (0%) were 1 to 4 pages in length, 11 (65%) were 5 to 10 pages in length, 5 (29%) were 11 to 15 pages in length, and 1 (6%) was 16 to 20 pages in length. No protocols were over 21 pages in length (see **Figures 3** and **4**). Of the positive protocol results tested in Search 3, none (0%) were 1 to 4 pages in length, 7 (58%) were 5 to 10 pages in length, none (0%) were 16 to 20 pages in length, and 2 (17%) were over 21 pages in length (see **Figures 5** and **6**). As mentioned before, all of the protocols samples from the searches were articles of protocols, and not full protocols, themselves.

Of these results, the journal sources of the top 10 longest protocol articles were extracted and analyzed. These journals included *BMC Psychiatry,* *BMJ Open, JMIR Research Protocols,* and *Trials. Trials* had the greatest number of long protocol articles, totaling at 5 with an average length of 13.8 pages. *JMIR Research Protocols* had 2 with an average length of 24 pages. *BMJ Open* also had 2 with an average length of 16.5 pages. *BMC Psychiatry* had 1 which had a length of 15 pages (see **Table 2**).

The resulting 4 journals were then analyzed for site navigability, average protocol length, and available protocol formats (see **Table 3** for scoring key and **Table 4** for results). *JMIR Research Protocols*, being a sister journal of *JMIR Publications*, was easily navigable, had protocols with an average page length of 10.7 pages, and had protocols in HTML, PDF, and XML formats. Consequently, the journal received 3 stars for site navigability, 1 star for average protocol length, and 3 stars for available formats. *BMJ Open*, being a part of *BMJ Journals*, was also easily navigable, had protocols with an average page length of 87.5 pages , and had protocols in only PDF format. Consequently, this journal received 3 stars for site navigability, 3 stars for average protocol length, and 1 star for available formats. This source journal was also the only one that had full protocols in its representative sample. *Trials*, being a part of *BMC*, was easily navigable, had protocols with an average page length of 9.3 pages, and had protocols in only PDF format. Consequently, this journal received 3 stars for site navigability, 1 star for average protocol length, and 1 star for available formats. *BMC Psychiatry*, being a part of *BMC/Springer Nature*, was easily navigable, had protocols with an average page length of 8.8 pages, and had protocols in only PDF format. Consequently, this journal received 3 stars for site navigability, 1 star for average protocol length, and 1 star for available formats (see **Table 4**).

# **Appendix 2: Narrative Version of ClinicalTrials.gov Results**

The top 7 linked source platforms from the ClinicalTrials.gov data subset included Clinical Study Data Request (1,215 links), NHLBI (Biolincc) (48 links), Immport (35 links), Drop Box (28 links), Google Drive (19 links), NCBI (17 links), and BioMedCentral (12 links) (see **Table 5** and **Figure 7**).

From each of these platforms, 10 random ClinicalTrials.gov protocol links (totaling 70 links) were selected and tested for positive links, path, available formats, and page length(see **Table 6**, **Figure 8**, and **Tables 7** and **8**).

BioMed Central (BMC) had 10 positive links, linked paths, only had protocols in PDF format, and had protocols with an average length of 9.3 pages. Consequently, the platform received 3 stars for positive links, 2 stars for path, 1 star for available formats, and 1 star for average protocol length.

Clinical Study Data Request had 0 positive links, as the protocols could not be located due to permission barriers for accessing study information. Consequently, Clinical Study Data Request received 0 stars for all 4 categories.

Drop Box had 4 positive links, direct paths, only a PDF format, and had protocols with an average length of 25.75 pages. Consequently, the platform received 1 star for positive links, 3 stars for path, 1 star for available formats, and 2 stars for average protocol length.

Google Drive had 7 positive links, direct paths, only a PDF format, and an average protocol length of 28.71 pages. Consequently, the platform received 2 stars for positive links, 3 stars for path, 1 star for available formats, and 2 stars for average protocol length.

Immport had 10 positive links, linked paths, only a PDF format, and an average protocol length of 61.9 pages. Consequently, the platform received 3 stars for positive links, 2 stars for path, 1 star for available formats, and 3 stars for average protocol length.

NCBI had 4 positive links, mostly indirect paths (1 protocol was linked), only a PDF format, and an average protocol length of 6.5 pages. Consequently, the platform received 1 star for positive links, 1 star for path, 1 star for available formats, and 1 star for average protocol length.

NHLBI (Biolincc) had 10 positive links, linked paths, only a PDF format, and an average protocol length of 107.6 pages. Consequently, the platform received 3 stars for positive links, 2 stars for path, 1 star for available formats, and 3 stars for average protocol length.

# **Appendix 3: Narrative Version of Platform Analysis Results**

Each of the platforms were analyzed for protocol availability, data availability, data dictionary availability, data formats, protocol formats, and available links (i.e. to trial registries). Policies found on each of the platform sites were also noted. A “0” or “NA” value indicates that evidence of the material was not found; this does not mean that the material does not exist, merely that a link to the material (even if the link does not grant access) could not be located. Request walls and navigability were also analyzed for each of the platforms. All of the above results were acquired through examining 10 of the most recent studies from each platform (totaling 60 studies).

Clinical Study Data Request: 9 protocols; 9 datasets; 8 data dictionaries; 1 data format (SAS); and no protocol formats were located (though evidence of protocols were indicated with a checkmark on the site, the formats of them were not listed). Available links included ClinicalTrials.gov, EudraCT, and sponsor registry links. Policies on the site included a data sharing agreement and a limited access period to requested data (see **Table 10** for full results and **Figure 10** for a visual comparison of protocol, data, and data dictionary sharing between platforms). Request walls included the need for an account (which required basic user information such as first/last name, country, email, and institution), and an extensive project proposal. Protocols, data, and data dictionaries could not be accessed without satisfying these barriers (though their presence was indicated with checkmarks in the study profiles) (see **Table 11**). Consequently, Clinical Study Data Request received 14 out of 27 stars in the 3-star scoring for general platform analysis (see **Tables 12** and **13**).

dbGaP: 0 protocols; 9 datasets; 9 data dictionaries; 1 data format (Image); and no protocol formats were located. Navigating the site was markedly difficult; in consequence, evidence of protocols was not verified. Available links included ClinicalTrails.gov, BioProject, BioSample, PubMed Central, and PubMed; however, some of the studies did not provide any links. Policies on the site included code of conduct, data use agreement, and security best practices policies (see **Table 10** for full results and **Figure 10** for a visual comparison of protocol, data, and data dictionary sharing between platforms). Request walls included the need for an account for the researchers’ institution; an account for the principle investigator (which required basic user information); and an extensive project proposal. Access to protocols, data, and data dictionaries could not be carried out without satisfying the above permission barriers (see **Table 11**). Consequently, dbGaP received 10 out of 27 stars in the 3-star scoring for general platform analysis (see **Tables 12** and **13**).

Nida Data Share: 10 protocols; 10 datasets; 10 data dictionaries; 3 data formats (CRF, SAS, and CSV); and 1 protocol format (PDF) were located. Available links included ClinicalTrials.gov. The only policy found on the site was a data use agreement (see **Table 10** for full results and **Figure 10** for a visual comparison of protocol, data, and data dictionary sharing between platforms). The only access barrier was the signature of the data use agreement (which required the entering of a name and email address). Protocols, data, and data dictionaries could be accessed by any who signed the data use agreement (see **Table 11**). Consequently, Nida Data Share received 20 out of 27 stars for the 3-star scoring for general platform analysis (see **Tables 12** and **13**).

Project Data Sphere: 10 protocols; 10 datasets; 10 data dictionaries; 2 data formats (CSV and Zip files); and 1 protocol format (PDF) were located. Available links included ClinicalTrials.gov. Policies found on the site included a data use agreement (see **Table 10** for full results and **Figure 10** for a visual comparison of protocol, data, and data dictionary sharing between platforms). Request walls included the need for an extensive data request; and an account (which is given to the researcher upon approval of a request). Protocols, data, and data dictionaries could only be accessed with the fulfillment of the above requirements (see **Table 11**). Consequently, Project Data Sphere received 18 out of 27 stars for the 3-star scoring for the general platform analysis (see **Tables 12** and **13**).

Vivli: 3 protocols; 4 datasets; 4 data dictionaries; 1 data format (image); and 1 protocol format (PDF) were located. Available links included ClinicalTrials.gov. Policies found on the site included a data use agreement (see **Table 10** for full results and **Figure 10** for a visual comparison of protocol, data, and data dictionary sharing between platforms). Request walls included the need for an account (which required basic identifying information); and an extensive project proposal. Protocols, data, and data dictionaries could only be accessed with the fulfillment of the above requirements, though, in a few cases, documentation could be accessed if it was available through the ClinicalTrials.gov link (see **Table 11**). Consequently, Vivli received 10 out of 27 stars for the 3-star scoring for the general platform analysis (see **Tables 12** and **13**).

NHLBI (aka BioLincc): 9 protocols; 10 datasets; 10 data dictionaries; 0 data formats (unverified due to limited access); and 1 protocol format (PDF) were located. Available links included ClinicalTrials.gov. Policies on the site included a data sharing agreement (see **Table 10** for full results and **Figure 10** for a visual comparison of protocol, data, and data dictionary sharing between platforms). Request walls included the need for an account on the site (requiring extensive personal information) and an extensive project proposal. Data could only be accessed with the fulfillment of the above requirements; however, data dictionaries and protocols could be accessed in spite of these barriers (see **Table 11**). Consequently, NHLBI received 15 out of 27 stars for the 3-star scoring for the general platform analysis (see **Tables 12** and **13**).