BMJ Open

Data sharing through an NIH central database repository: a cross-sectional survey of BioLINCC users

Joseph S Ross,[1,2,3,4] Jessica D Ritchie,[1] Emily Finn,[2] Nihar R Desai,[1,5] Richard L Lehman,[1,6] Harlan M Krumholz,[1,3,4,5] Cary P Gross[2,3,7]

To cite: Ross JS, Ritchie JD, Finn E, *et al*. Data sharing through an NIH central database repository: a cross-sectional survey of BioLINCC users. *BMJ Open* 2016;**6**: e012769. doi:10.1136/bmjopen-2016-012769

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-012769).

Note: Emily Finn was affiliated with the Center for Outcomes Research and Evaluation, Yale New Haven Hospital, during the time the work was conducted.

Received 22 May 2016
Revised 29 July 2016
Accepted 30 August 2016

For numbered affiliations see end of article.

Correspondence to
Dr Joseph S Ross;
joseph.ross@yale.edu

ABSTRACT

Objective: To characterise experiences using clinical research data shared through the National Institutes of Health (NIH)'s Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) clinical research data repository, along with data recipients' perceptions of the value, importance and challenges with using BioLINCC data.

Design and setting: Cross-sectional web-based survey.

Participants: All investigators who requested and received access to clinical research data from BioLINCC between 2007 and 2014.

Main outcome measures: Reasons for BioLINCC data request, research project plans, interactions with original study investigators, BioLINCC experience and other project details.

Results: There were 536 investigators who requested and received access to clinical research data from BioLINCC between 2007 and 2014. Of 441 potential respondents, 195 completed the survey (response rate=44%); 89% (n=174) requested data for an independent study, 17% (n=33) for pilot/preliminary analysis. Commonly cited reasons for requesting data through BioLINCC were feasibility of collecting data of similar size and scope (n=122) and insufficient financial resources for primary data collection (n=76). For 95% of respondents (n=186), a primary research objective was to complete new research, as opposed to replicate prior analyses. Prior to requesting data from BioLINCC, 18% (n=36) of respondents had contacted the original study investigators to obtain data, whereas 24% (n=47) had done so to request collaboration. Nearly all (n=176; 90%) respondents found the data to be suitable for their proposed project; among those who found the data unsuitable (n=19; 10%), cited reasons were data too complicated to use (n=5) and data poorly organised (n=5). Half (n=98) of respondents had completed their proposed projects, of which 67% (n=66) have been published.

Conclusions: Investigators were primarily using clinical research data from BioLINCC for independent research, making use of data that would otherwise have not been feasible to collect.

Strengths and limitations of this study

- Data sharing policies are increasingly promoted and being adopted by research funders to improve access to clinical trial data to inform evidence-based practice. The National Institutes of Health (NIH)'s Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) has been actively sharing data from its clinical research data repository for >10 years.
- In the first survey of the experiences of investigators who have requested and been approved to use data from BioLINCC, we found that users were primarily focused on conducting independent research studies, making use of data that would otherwise have not been feasible to collect, because of insufficient time and resources.
- We also found that shared data from BioLINCC could be used to successfully pursue clinical research; 90% of BioLINCC users found the data to be suitable, half had completed their research projects thus far, and two-thirds had published their findings.
- Our study of user experiences with BioLINCC offers important insights for newly initiated and ongoing clinical trial data sharing efforts and illustrates the potential and value of data sharing for the broader scientific field, as well as the challenges that remain to be overcome.
- Our study is limited by a low response rate and may have been affected by recall bias and social desirability bias, perhaps suggesting that our findings overestimate the perceived value of BioLINCC data and their usability for the broader scientific community.

Over the past 5 years, several major research funders, including the US National Institutes of Health (NIH), the US Patient-Centered Outcomes Research Institute, the UK Medical Research Council and the Bill and Melinda Gates Foundation, as well as private industry,[1] have adopted policies supporting or mandating clinical research data sharing. In January 2015, the Institute of Medicine of the US National Academies further supported these efforts with its report, 'Sharing Clinical Trial Data: Maximizing Benefits,

Minimizing Risks', recommending that stakeholders foster a culture in which data sharing is the expected norm and commit to responsible strategies aimed at maximising benefits, minimising risks and overcoming challenges of sharing clinical trial data.[2] In January 2016, the International Committee of Medical Journal Editors issued a proposal to require authors to share with others the de-identified individual patient data underlying the results presented in the article no later than 6 months after publication as a condition of consideration for publication of a clinical trial report in its member journals.[3]

In response to these new policies and proposals, funded investigators will increasingly be asked to prepare and make collected data available to other investigators with whom they are not collaborating so that the second can pursue independent research. To support these efforts and inform developing policies, a number of prior studies have examined the willingness of clinical trial investigators to share clinical research data, generally finding broad support, and characterised anticipated challenges to and concerns with data sharing.[4–11] However, few studies have focused on the investigators who have actually received de-identified individual patient data from a centralised data sharing platform, in order to understand their perspectives regarding challenges encountered with requesting and using the data, and disseminating findings.

While most of these data sharing efforts have been relatively newly established, the US National Heart, Lung, and Blood Institute (NHLBI) of the NIH established a formal data repository in 2000, now managed by the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC), to facilitate access to, maximise the scientific value of, and promote the availability and use of the biorepository, data repository and other NHLBI-funded population-based biospecimen and data resources by investigators worldwide.[12 13] The BioLINCC data repository includes individual-level data on >580 000 participants from over 110 institute supported clinical trials and observational studies, beginning as far back as the 1980s. Each data set is prepared independently by the NHLBI-funded investigator to comply with specific requirements and data standards, with oversight by BioLINCC, including provision of baseline, interim visit, ancillary study and outcome data for clinical trials and provision of all examination and ancillary study data, along with follow-up information, for epidemiology studies. As BioLINCC has been actively sharing data for more than a decade and currently receives over 100 requests for clinical trial and other prospective cohort clinical data per year (ref: personal communication, Sean Coady, NHLBI data repository manager), there is an opportunity to learn from data users' experiences to inform clinical data sharing efforts. Accordingly, we surveyed all investigators who requested and received access to clinical research data from BioLINCC between 2007 and 2014. We specifically sought to understand their experiences with clinical research data sharing and status of their research project, as well as perceptions of the value, importance and challenges of accessing data through BioLINCC.

METHODS

Study sample and design

We conducted a cross-sectional survey from May to August 2015 of all investigators who requested and received access to clinical research data from BioLINCC between 2007 and 2014. This time period was chosen to ensure a contemporaneous sample of investigators whose contact information was less likely to have changed over ensuing years. In accordance with NIH policy, BioLINCC provided our study team with a list of investigators who had requested and received access using a *public email address*; contact information was available for the lead investigator who was responsible for the BioLINCC request, not each member of the study team. For investigators who had requested and received access using a *private email address*, BioLINCC first sent an opt-in/opt-out email in May 2015, asking if they would be willing to participate in the survey (see online supplementary appendix). Non-respondents were sent two follow-up requests by email; those that did not respond by the end of the third week were considered to have opted out. BioLINCC subsequently provided our study team with a list of those investigators who opted in.

In addition to contact information, BioLINCC provided our study team with information on the following for all investigators who had requested and received access to clinical research data: lead investigator location, affiliation with an academic institution or for-profit organisation and total number of requests ever submitted to BioLINCC, as well as the request year, the number of data sets requested and self-reported availability of external funding to support the research project using the requested data.

In May 2015, the Yale team sent all potential survey respondents an initial email to describe the purpose of the study, request their participation and provide a link to the survey; three follow-up requests were sent by email over the course of June 2015. Non-respondents were contacted by telephone to solicit their participation up to twice per week, but no more than once per day, until one contact was made. In July 2015, internet searches to update contact information for non-respondents were conducted. For all non-respondents whose updated contact information was identified, the initial survey email was sent, followed by three follow-up requests.

Invitations to participate did not reference a specific hypothesis of the study, but stated that investigator participation would further the understanding of investigators' experience with BioLINCC and inform future clinical trial data sharing efforts (see online

supplementary appendix). Participation was voluntary and included an opportunity to win one of five $100 gift certificates for Amazon. All internet-based responses were collected using a web-based survey platform (Qualtrics Labs, Provo, Utah, USA).

Survey instrument development

The design of our 50-item survey instrument was informed by previously published surveys,[4 5] a review of the literature on clinical trial data sharing, and discussion with multiple experts and stakeholders, including representatives from NHLBI and academic investigators. Experts recommended survey topics that they considered to be compelling for the field of data sharing and reuse of data. The survey was pretested with six medical students and staff at the Center for Outcomes Research and Evaluation, Yale New Haven Hospital (New Haven, Connecticut) and modified iteratively to improve clarity, face validity and content validity. Adaptive questioning was used to decrease response burden. Items were presented in multiple response, Likert scale and open-ended formats; many of the multiple response questions enabled respondents to select multiple answers. The complete instrument is provided within the online supplementary appendix.

Survey domains

Reasons for data request and planned research project

We used multiple response and yes/no questions to assess investigators' primary research purpose and reasons for requesting data from BioLINCC. Multiple response questions were also used to determine the primary research objective, funding used to support the project and other details of the planned research project. Knowing what these clinical research data are being used for will help tailor future data sharing efforts to the needs of investigators.

Interactions with original study investigators

We used yes/no questions to determine whether original study investigators were contacted prior to or after requesting data through BioLINCC to obtain the data or to collaborate. These were followed by multiple response questions to determine why collaborations were sought, whether the requests for data or collaboration were approved and reasons for not approving. Answers to these questions could potentially demonstrate the value of a data resource such as BioLINCC.

BioLINCC experience

Multiple response, yes/no and Likert-type questions were used to obtain information regarding investigators' experience using BioLINCC, including whether the data were suitable and useful for their project. Knowledge gained from these questions can help to improve BioLINCC and other data sharing efforts.

Project details

We used multiple response and yes/no questions to characterise the completion stage of investigators' projects. For those that did not complete their project, multiple response and yes/no questions were used to ascertain reasons why the project was incomplete. For those with completed projects, we used multiple response and yes/no questions to determine whether the final project differed from the prespecified project as well as to obtain publication information. Multiple choice and multiple response questions were used to identify any funding sources, and whether using the data from BioLINCC aided in any future grant applications. It is important to demonstrate that these data are being requested, and they are also being used to potentially generate new knowledge to advance science and public health.

Requestor demographics

Respondents were asked to characterise their primary employer and career status using multiple choice questions, including whether they had ever been closely involved (as principal or coinvestigator) in the conduct of a randomised controlled trial and/or ever deposited clinical trial data in the BioLINCC repository. Respondent sociodemographic characteristics, including age, gender and ethnicity, were also collected. While these characteristics were collected for descriptive purposes only, age, along with the professional characteristics collected, are of importance to demonstrate the value of the availability of BioLINCC data to investigators who are in certain stages of their career.

Patient involvement

Patients were not involved in the design or conduct of this study. Results will be directly disseminated via email to all individuals invited to participate in the survey on publication.

Statistical analysis

To compare characteristics of survey respondents and non-respondents, we used two-sided χ^2 tests and Fisher's exact tests when appropriate with a type 1 error level of 0.05. Next, we conducted descriptive analyses of the reasons for requesting data from BioLINCC, prior interactions with original trial investigators, experience using BioLINCC and project details, as well as respondent demographic characteristics. Data were analysed using JMP Pro V.11.2.0 (SAS Institute, Cary, North Carolina, USA).

RESULTS

There were 536 investigators who requested and received access to clinical research data from BioLINCC between 2007 and 2014 (figure 1). Investigators for which a public email address was not available were sent an opt-in/opt-out letter (n=74); 23 opted in, 3 opted out, 7 could

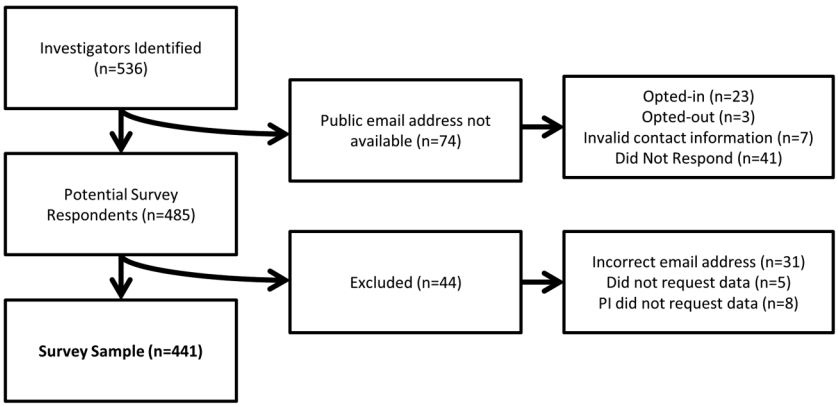

Figure 1 Inclusion flow chart used to identify potential survey respondents: investigators who had requested and received access to clinical research data from BioLINCC between 2007 and 2014. PI, principal investigator.

not be reached and 41 were not responsive. Survey participation requests were thus sent to 485 eligible respondents, 44 of whom were subsequently excluded due to the following reasons: invalid contact information (n=31), the investigator had no recollection of requesting the data (n=5) or the data had been requested by someone other than the investigator (n=8). Of the remaining 441 respondents, 195 completed the survey, yielding a survey response rate of 44.2%. However, of the 536 total investigators who requested and received access to clinical research data from BioLINCC, 195 completed the survey (response rate of 36.3%).

Survey respondents did not differ from non-respondents with respect to investigator location, affiliation with an academic institution or for-profit organisation and total number of requests ever submitted to BioLINCC, as well as the number of data sets requested (p≥0.10; table 1). However, respondents were more likely than non-respondents to have requested data more recently (p=0.004) and to have self-reported external funding to support the research project (p=0.009).

Half of survey respondents were between 35 and 49 years of age (n=97; 50%), while 59% were male (n=116), 68% were white (n=133) and 90% identified as not Hispanic/Latino (n=175; table 2). The vast majority of respondents were primarily employed by an academic institution (n=165; 85%) and 78% (n=152) have been engaged in clinical research for at least 3 years. While 42% (n=82) had been closely involved in the conduct of a randomised controlled trial, only 3% (n=5) had ever deposited data in the BioLINCC repository.

Table 1 Characteristics of survey respondents and non-respondents

	Respondents, no. (%) (n=195)	Non-respondents, no. (%) (n=246)	p Value
Investigator based in the USA?			
Yes	163 (84)	211 (86)	0.53
No	32 (16)	35 (14)	
Investigator based at academic institution?			
Yes	149 (76)	196 (80)	0.41
No	46 (24)	50 (20)	
Investigator based at for-profit institution?			
Yes	5 (3)	4 (2)	0.49
No	190 (97)	242 (98)	
Investigator's total submitted requests to BioLINCC (ever), no.			
1	120 (62)	169 (69)	0.12
>1 (includes renewals)	75 (38)	77 (31)	
Data sets requested, no.			
1	152 (78)	171 (70)	0.10
2–4	31 (16)	58 (24)	
5–9	7 (4)	14 (6)	
10+	5 (3)	3 (1)	
Request year			
2006	0 (0)	1 (<1)	0.004
2007	13 (7)	17 (7)	
2008	4 (2)	16 (7)	
2009	9 (5)	23 (9)	
2010	6 (3)	17 (7)	
2011	12 (6)	26 (11)	
2012	43 (22)	55 (22)	
2013	47 (24)	39 (16)	
2014	61 (31)	52 (21)	
External funding to support the research project?			
Yes	74 (38)	77 (31)	0.009
No	97 (50)	111 (45)	
Unknown	24 (12)	58 (24)	

Reasons for data request

Overall, respondents' motivations for requesting data from BioLINCC were largely focused on using the data to conduct and disseminate new research studies, as 89% (n=174) indicated that data were requested for an independent study, 17% (n=33) to use the data for pilot/preliminary analysis. For 63% (n=122) of respondents, the decision to request data through BioLINCC was influenced by the belief that collecting data of similar size and scope was not feasible, while insufficient financial resources for primary data collection (n=76; 39%), individual participant-level data being unavailable elsewhere (n=71; 36%), and insufficient time for primary data collection (n=64; 33%) were also commonly cited reasons for requesting data through BioLINCC (figure 2).

Table 2 Sociodemographic and professional characteristics of survey respondents (n=195)

Characteristic	No (%) of respondents
Age	
34 years or younger	29 (15)
35–49 years	97 (50)
50–64 years	47 (24)
65 years or older	14 (7)
Prefer not to answer	8 (4)
Gender	
Male	116 (59)
Female	74 (38)
Prefer not to answer	5 (3)
Race	
White	133 (68)
Asian	35 (18)
Black or African-American	10 (5)
Other	3 (2)
Prefer not to answer	14 (7)
Ethnicity	
Hispanic or Latino	8 (4)
Not Hispanic or Latino	175 (90)
Prefer not to answer	12 (6)
Primary employer	
Academic institution	165 (85)
Non-profit organisation	14 (7)
Government	8 (4)
Private industry	4 (2)
Other	4 (2)
Career stage	
In training (<3 years of active engagement in clinical research, still receiving formative training in research methods)	43 (22)
Early stage career (3–10 years of active engagement in clinical research)	83 (43)
Established in the field (>10 years of active engagement in clinical research)	69 (35)
Ever been closely involved (as PI or co-PI) in the conduct of a randomised controlled trial?	
Yes	82 (42)
Ever deposited clinical trial data in the BioLINCC repository?	
Yes	5 (3)

PI, principal investigator; co-PI, coinvestigator.

Planned research project

Respondents largely (n=149; 76%) planned research projects that used the requested BioLINCC data as a standalone data source for at least one project, while 43% (n=83) planned to combine the data with other data sources; of these, 27% (n=22) planned to conduct a meta-analysis. Nearly all respondents (n=186; 95%) indicated that at least one of their primary research objectives was to complete new research, whereas only 7 (4%) had a primary research objective solely to replicate prior analyses. Of those pursuing new research, 56% (n=104) planned to leverage the data for a research question unrelated to the original research design, while 40% (n=74) planned to examine subgroup populations and 32% (n=60) planned to examine secondary end points.

Only 13% (n=26) of respondents indicated that the focus of their research was a medical product or intervention; of these, 73% (n=19) planned analyses to examine product/intervention efficacy, 54% (n=14) safety. Finally, 52% (n=102) of respondents had funding to support the research project, most commonly from the NIH (n=44; 23%), whereas 43% (n=84) primarily self-funded the research project.

Interactions with original study investigators

Fewer than one in five (n=36; 18%) respondents indicated that they had contacted the original study investigators to obtain data prior to requesting the data from BioLINCC; among these, 44% (n=16) reported that the original study investigator approved their request and these investigators most commonly requested access to the data from BioLINCC anyway because the process to access data was more straightforward through BioLINCC (n=11). Among the 20 (56%) respondents who indicated that the original study investigator denied their request, the most common response given by the original investigator was to direct the respondent to BioLINCC (n=11; 55%).

Nearly one-quarter of respondents (n=47; 24%) indicated that they contacted the original study investigator to request collaboration, most commonly because of an interest in working with the original study investigators (n=23) and need for additional content expertise due to study design complexity (n=20). Of the respondents who requested collaboration, two-thirds (n=31; 66%) indicated that the request was accepted.

Data repository experience

Nearly all respondents indicated satisfaction with the data available through BioLINCC and that they were suitable for their originally proposed project (n=176; 90%). Among the 19 (10%) respondents who indicated that the data were not suitable, the two most commonly cited reasons were that the data were too complicated to use, preventing them from determining whether the data were suitable (n=5); and that the data were poorly organised, preventing adequate preparation for analysis (n=5).

Research project details

Half of all respondents (n=98; 50%) reported that their projects have been completed, of which 67% (n=66) have been published. Respondents who had requested data prior to 2012 were more likely to have completed their project when compared with those who had requested data in 2012 or afterwards (73% vs 44%; p=0.008). However, among those who completed their project, rates of publication did not differ among those who had requested data prior to 2012 and those who

had requested data in 2012 or afterwards (63% vs 69%; p=0.57). Of those who have completed their research, 48% (n=47) indicated that no substantive concerns were raised about the use of data from BioLINCC during the peer-review process, while 8% indicated that concerns were raised about research methodology and analysis (n=8), 7% about the original study design that the investigator could not address (n=7), and 6% about their research project design that they could not address without additional data (n=6).

Of the 97 respondents (50% of total) who have not yet completed their proposed projects, 84% (n=81) explained that they planned to complete their project; 65% (n=63) indicated that their project is in analysis/manuscript draft phase, while 28% (n=27) explained that they have thus far been too busy with other

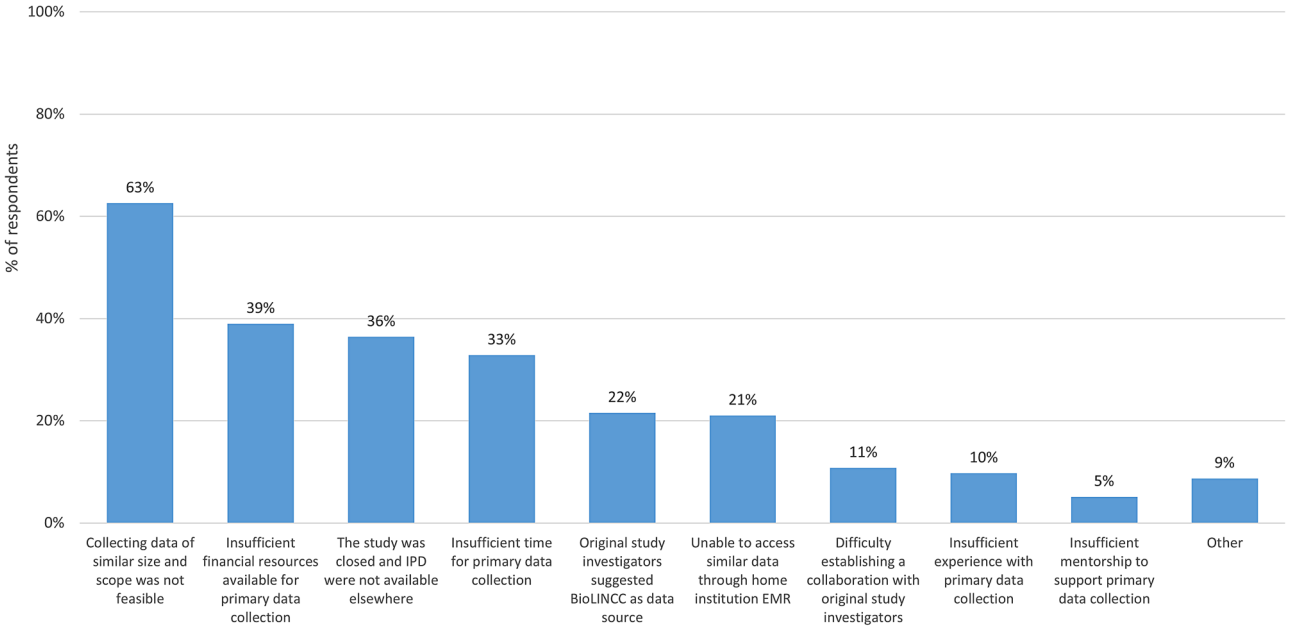

Figure 2 Factors influencing decision to request clinical research data through BioLINCC between 2007 and 2014 (n=195). **Note:** Respondents were able to select multiple answers in response to this question. BioLINCC, Biologic Specimen and Data Repository Information Coordinating Center; EMR, electronic medical record; IPD, individual participant data.

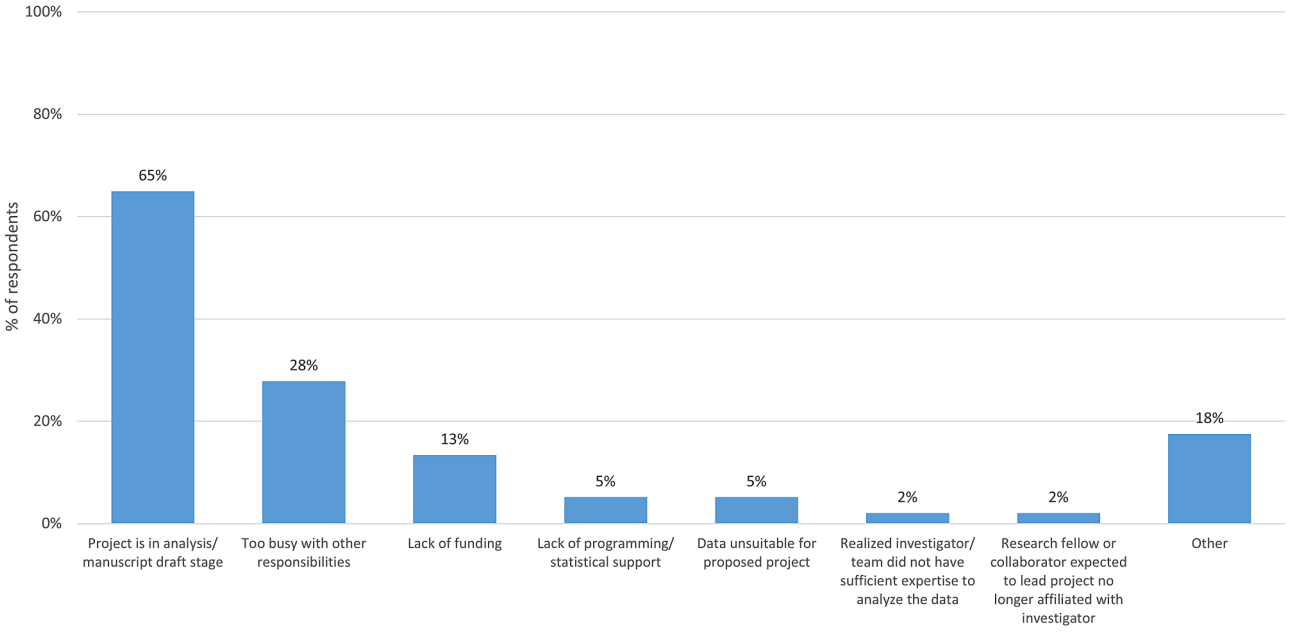

Figure 3 Flow chart showing completion rates of research projects using clinical research data requested from BioLINCC between 2007 and 2014. **Note:** Respondents were able to select multiple answers in response to this question.

responsibilities to complete the research project using the data from BioLINCC and 13% (n=13) reported that lack of funding to support the project was a problem (figure 3). A total of 16 investigators explained that they did not intend to complete their project, most often because the age of the data made the project now less relevant or because of data issues, such as missing values for the variable of interest.

Of the 179 respondents who already completed or planned to complete their proposed project, 54% (n=96) reported that there would be one research project resulting from their single request for data from BioLINCC, 23% (n=42) reported two and 23% (n=41) reported three or more. In addition, 15% (n=27) of respondents who have completed or planned to complete their project indicated that their completed/anticipated final project differed from their prespecified project; the most commonly modified aspects were the statistical analysis plan (n=18) and the selection of the main independent variables (n=12).

DISCUSSION

In this survey of investigators who had requested and received access to clinical research data from BioLINCC between 2007 and 2014, the vast majority had requested the data in order to conduct independent research projects, primarily because collecting data of similar size and scope was not feasible, due to insufficient time and resources. Half of the investigators had completed their research projects thus far, two-thirds of which published their findings, and among those investigators whose projects were incomplete, two-thirds were actively engaged in analysis or manuscript preparation. These findings offer important insights for newly initiated and ongoing clinical trial data sharing efforts and illustrate the potential and value of data sharing for the broader scientific field, as well as the challenges that remain to be overcome.

First, the BioLINCC experience suggests that when clinical research data are made available to investigators, there is likely to be interest in using the data for independent research projects. There are currently 654 publications associated with the data repository available through BioLINCC.[14] This large number of publications suggests that these data are being used by investigators, better maximising the NHLBI investment in and scientific value of clinical research data. Many investigators responding to our survey noted that collecting data of similar size and scope was not feasible, or that they had insufficient financial resources or time for primary data collection, justifying the need to request data from BioLINCC for their research.

Second, the BioLINCC experience suggests that clinical data can be collected by one set of investigators and made available to another set of investigators who, for the most part, can use it to successfully pursue an independent research project. While some surveyed investigators noted challenges in using the data made available through BioLINCC, 90% found the data to be suitable for their originally proposed project, even without input from the original research team. Few reported that the data were too complicated to use, preventing them from determining whether the data were suitable, or that the data were poorly organised.

Finally, the research enterprise is not optimally efficient, and the BioLINCC experience reflects this short-coming. In aggregate, >100 research projects were completed as a result of respondent investigators using data made available through BioLINCC. However, despite all investigators having received data from BioLINCC at no cost, only half of investigators who had received data had completed their research projects thus far. While many more continue to work on their projects and intend to complete their work, the investment by NHLBI to make these data available should be matched by the effort of investigators to ensure that the projects are completed. Moreover, even among completed projects, only two-thirds were published. While BioLINCC maintains an updated list of publications that have resulted from use of this shared data,[14] mechanisms should be established to ensure that results from research made possible through data sharing are publicly disseminated, either through publication or through a results reporting initiative similar to ClinicalTrials.gov.

For the potential and value of data sharing to be fully realised, more needs to be accomplished. Part of the success of BioLINCC may be attributed to the NHLBI policy that supported studies with direct costs equal to or ≥$500K in any 1 year and identified as being of high programmatic interest, along with cooperative agreements with 500 or more participants, are required to submit data as part of the grant award.[13] This policy establishes clear expectations for data sharing, so that data can be properly organised and de-identified and supportive documentation and materials prepared in anticipation of submitting data to BioLINCC. However, it is not clear whether this policy allows researchers to budget resources for this work. Currently, the NIH is seeking ways to broaden data sharing efforts across its institutes;[15] to enhance the likelihood of success of data sharing efforts, it should be clarified whether NIH-granted independent research funds can be used to prepare collected data for sharing through initiatives such as BioLINCC.

Similarly, financial support for investigators to use clinical research data that are being shared and made available would enhance efforts. In total, 43% of investigators using data from BioLINCC had self-funded their research efforts, while 23% were relying on funding from the NIH. However, among surveyed investigators who had not yet completed their proposed projects, lack of funding to support the project was a commonly cited problem. Without financial support, efforts to share data are likely to fail to achieve their

potential,[2] even despite the strong policies and proposals in favour of data sharing from other research funders, the Institute of Medicine and the International Committee of Medical Journal Editors.

There are important limitations of our study to consider. First, only 44% of potentially eligible respondents completed our survey, perhaps suggesting that our findings overestimate the perceived value of BioLINCC data and its usability for the broader scientific community. Individuals who chose not to respond to our survey may have found the data to be more problematic and less useful than those who responded. Furthermore, even among respondents, our findings may have been biased by recall bias, including an inability to remember using the data made available by BioLINCC, and social desirability,[16 17] as respondents may have been less likely to self-report experiences and project completion plans that may be negatively perceived by others. In addition, there were a few observed differences between survey respondents and non-respondents. As we would expect that investigators who made more recent requests and who had secured external funding to support the research project would be more likely to remain enthusiastic about the project and to complete it, our findings may be biased towards higher project completion rates. However, our response rate compares favourably with other surveys of physicians and investigators,[4 18–20] perhaps reflecting that we used several mechanisms to prospectively improve response rates, including a web-based survey platform for ease of completion, we employed several reminder contacts, including three emails and at least one telephone contact and we offered financial incentives for participation.

Second, our study was limited to investigators who had received data from BioLINCC and our findings may not be applicable to the experience of investigators obtaining data from other repositories. There is currently great interest and scrutiny of existing clinical trial data sharing efforts,[21–24] many of which require submission of a research proposal, as does BioLINCC, and some of which only make data available via a virtual, secure data sharing environment, as opposed to BioLINCC which provides de-identified data directly to approved researchers. One recent study evaluated how many clinical trials were publicly available to the research community through three open access data sharing platforms: ClinicalStudyDataRequest.com, the Yale University Open Data Access (YODA) Project and the Supporting Open Access for Researchers (SOAR) Initiative, finding that while >3000 trials were available, only 15.5% had been requested by a limited number of investigators.[25] The authors concluded that data sharing efforts are being underused, implicitly questioning the value of continued resource investment. However, the results of our survey of BioLINCC users suggests this conclusion may be premature, as use of data from these open access platforms can be expected to grow with time, although more remains to ensure the use of these data, and the successful completion and publication of the resulting research, to justify the investments being made in data sharing.

A third limitation of our study is that some information of interest was not asked in order to reduce survey response burden, including questions asking about the time and effort invested to manage and analyse the data from BioLINCC and the impact of the publications resulting from the research project. Finally, our study made no attempt to judge the impact of the research that was able to be completed because of the clinical research data made available through BioLINCC. Other efforts should consider whether the investment being made by NIH and NHLBI in data sharing is justified by the information and knowledge being generated for medical science and society.

In conclusion, we found that the vast majority of investigators who had requested and received access to clinical research data from BioLINCC between 2007 and 2014 had either succeeded in completing their research project or reported being actively involved in data analysis or manuscript preparation. In aggregate, >100 research projects were completed as a result of respondent investigators using data made available through BioLINCC. Experience with BioLINCC illustrates the potential of data sharing for the broader scientific field and the importance of funding these efforts, particularly when collecting data of similar size and scope is not feasible for many investigators.

Author affiliations
[1]Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, Connecticut, USA
[2]Department of Internal Medicine, Section of General Internal Medicine, Yale School of Medicine, New Haven, Connecticut, USA
[3]Department of Internal Medicine, Robert Wood Johnson Foundation Clinical Scholars Program, Yale School of Medicine, New Haven, Connecticut, USA
[4]Department of Health Policy and Management, Yale School of Public Health, New Haven, Connecticut, USA
[5]Department of Internal Medicine, Section of Cardiovascular Medicine, Yale School of Medicine, New Haven, Connecticut, USA
[6]UK Cochrane Center, Oxford, UK
[7]Cancer Outcomes, Public Policy, and Effectiveness Research Center, Yale Cancer Center, New Haven, Connecticut, USA

Acknowledgements The authors would like to acknowledge Mr. Sean Coady of the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health (NIH), who provided feedback during survey development and assistance with contact information for potential survey respondents, as well as Ms. Tiffany Chang, MPH, Ms. Julia Eichenfield and Mr. Daniel Shaw, who provided background research and assistance with survey distribution during the course of student employment/voluntary summer research experience at the Center for Outcomes Research and Evaluation (CORE), Yale New Haven Hospital.

Contributors JSR, HMK and CPG conceived the concept and design of study. JDR and EF were responsible for acquisition of data. All authors were responsible for analysis and interpretation of data and for critical revision of manuscript. JSR, JDR and EF drafted the manuscript. JSR and EF carried out statistical analysis. JSR supervised the study.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The authors assume full responsibility for the accuracy and completeness of the ideas presented.

Competing interests All authors have completed the International Committee of Medical Journal Editors (ICMJE) uniform disclosure form at http://www.icmje.org/coi_disclosure.pdf and all authors declare (currently or formerly) receiving support through Yale University from Medtronic and Johnson and Johnson to develop methods of clinical trial data sharing and from the Blue Cross Blue Shield Association (BCBSA) to better understand medical technology evidence generation. HMK and JSR receive support through Yale University from the Centers of Medicare and Medicaid Services (CMS) to develop and maintain performance measures that are used for public reporting and from the Food and Drug Administration (FDA) to develop methods for postmarket surveillance of medical devices. HMK chairs a cardiac scientific advisory board for UnitedHealth.

Ethics approval Ethics approval from the Yale University School of Medicine Human Research Protection Program was obtained prior to study conduct and consent was considered to be implied when participants completed the online survey.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Extra data can be accessed via the Dryad data repository at http://datadryad.org/ with the doi:10.5061/dryad.j38b7.

Transparency The lead author (JSR) affirms that this manuscript is an honest, accurate and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

REFERENCES

1. Pharmaceutical Research and Manufacturers of America (PhRMA) and European Federation of Pharmaceutical Industries and Associations (EFPIA). Principles for Responsible Clinical Trial Data Sharing. Our Committment to Patients and Researchers. July 2013. http://phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf (accessed 14 Jul 2016).
2. Institute of Medicine of the National Academies of Science. *Sharing clinical trial data: maximizing benefits, minimizing risks*. Washington DC: National Academies Press, 2015.
3. Taichman DB, Backus J, Baethge C, *et al*. Sharing Clinical Trial Data: A Proposal From the International Committee of Medical Journal Editors. *Ann Intern Med* 2016;164:505–6.
4. Rathi V, Dzara K, Gross CP, *et al*. Sharing of clinical trial data among trialists: a cross sectional survey. *BMJ* 2012;345:e7570.
5. Rathi VK, Strait KM, Gross CP, *et al*. Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey. *Trials* 2014;15:384.
6. Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE* 2009;4:e7078.
7. Tenopir C, Allard S, Douglass K, *et al*. Data sharing by scientists: practices and perceptions. *PLoS ONE* 2011;6:e21101.
8. Tenopir C, Dalton ED, Allard S, *et al*. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE* 2015;10:e0134826.
9. Oushy MH, Palacios R, Holden AE, *et al*. To Share or Not to Share? A Survey of Biomedical Researchers in the U.S. Southwest, an Ethnically Diverse Region. *PLoS ONE* 2015;10:e0138239.
10. Tudur Smith C, 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:e97886.
11. Hopkins C, Sydes M, Murray G, *et al*. UK publicly funded Clinical Trials Units supported a controlled access approach to share individual participant data but highlighted concerns. *J Clin Epidemiol* 2016;70:17–25.
12. Giffen CA, Carroll LE, Adams JT, *et al*. Providing Contemporary Access to Historical Biospecimen Collections: Development of the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). *Biopreserv Biobank* 2015;13:271–9.
13. Coady SA, Wagner E. Sharing individual level data from observational studies and clinical trials: a perspective from NHLBI. *Trials* 2013;14:201.
14. U.S. National Institutes of Health, National Heart L, and Blood Institute, Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). Publications associated with available BioLINCC resources. https://biolincc.nhlbi.nih.gov/publications/ (accessed 8 Jan 2016).
15. Hudson KL, Collins FS. Sharing and reporting the results of clinical trials. *JAMA* 2015;313:355–6.
16. Fowler FJ. *Survey research methods*. Newbury Park, CA: Sage, 1993.
17. Sudman S, Bradburn NM. *Asking questions: a practical guide to questionnaire design*. San Francisco, CA: Jossey-Bass, 1982.
18. Asch DA, Jedrziewski MK, Christakis NA. Response rates to mail surveys published in medical journals. *J Clin Epidemiol* 1997;50:1129–36.
19. Keyhani S, Federman A. Doctors on coverage—physicians' views on a new public insurance option and Medicare expansion. *N Engl J Med* 2009;361:e24.
20. Shanafelt TD, Boone S, Tan L, *et al*. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. *Arch Intern Med* 2012;172:1377–85.
21. Longo DL, Drazen JM. Data Sharing. *N Engl J Med* 2016;374:276–7.
22. Sydes MR, Johnson AL, Meredith SK, *et al*. Sharing data from clinical trials: the rationale for a controlled access approach. *Trials* 2015;16:104.
23. Tudur Smith C, Hopkins C, Sydes MR, *et al*. How should individual participant data (IPD) from publicly funded clinical trials be shared? *BMC Med* 2015;13:298.
24. Strom BL, Buyse M, Hughes J, *et al*. Data sharing, year 1—access to data from industry-sponsored clinical trials. *N Engl J Med* 2014;371:2052–4.
25. Navar AM, Pencina MJ, Rymer JA, *et al*. Use of Open Access Platforms for Clinical Trial Data. *JAMA* 2016;315:1283–4.