Access to Clinical Trial Data

Assessing the research potential of access to clinical trial data

March 2015

wellcome trust

technopolis |group|

January 2015

Assessing the research potential of access to clinical trial data

Final report to the Wellcome Trust

www.technopolis-group.com

technopolis |group|

Assessing the research potential of access to clinical trial data

Final report to the Wellcome Trust

technopolis |group|, January 2015

Peter Varnai

Maike Rentel

Paul Simmonds

Tammy-Ann Sharp

Bastian Mostert

Thyra de Jongh

3 Pavilion Buildings
Brighton BN1 1EE
T: +44 (0)1273 204320
E: peter.varnai@technopolis-group.com
E: info@technopolis-group.com

Please cite this study as: Varnai P., Rentel MC., Simmonds P., Sharp, TA., Mostert, B., de Jongh, T. (2014): Assessing the research potential of access to clinical trial data. A report to the Wellcome Trust. Study led by Technopolis Group (UK).

Table of Contents

Figures

Tables

Boxes

Executive summary

Introduction

Clinical trials generate systematic and vital information about the efficacy, safety and quality of medical interventions. However, there has been much discussion about the degree to which trial results and the underpinning data are currently made available to researchers outside of the team that conducted the original trial. Recent legislation and various initiatives have started to contribute to increased availability of data from clinical trials, predominantly at summary level, but also at the level of the individual trial participant. This enhanced access not only allows for more transparency in clinical research, but can also drive generation of knowledge, allowing researchers to tackle new research questions, to reduce duplication and optimise the design of trials, or to increase the efficiency of the research process by linking data from multiple trials. However, substantial barriers to accessing individual participant data (IPD) continue to exist. Overcoming these barriers offers the potential for improved access to clinical trial data and for fully exploiting existing research data to the benefit of the scientific community and, ultimately, the patient.

This report presents the findings of a study commissioned by the Wellcome Trust in April 2014. The study draws on independent research carried out by Technopolis Group, and was supported by an independent expert review group comprised of Mike Clarke (Queen's University Belfast), Trudie Lang (University of Oxford), Fiona Reddington (Cancer Research UK), Matt Sydes (Medical Research Council Clinical Trials Unit at University College London), and Catrin Tudur Smith (University of Liverpool).

Objectives

The primary aims of this study were to develop an understanding of the types of novel clinical research that may be possible using IPD from clinical trials, and to develop case studies of existing and future examples of such research and its benefits. The study also aimed at assessing the potential level of demand for a broader access model for such clinical trial data, and whether there are appreciable differences between the academic and commercial research communities in terms of needs and use of clinical trial data. The study ultimately wants to contribute to discussions about key mechanisms and practicalities that would need to underlie a broader access model for clinical trial data.

While increased transparency is an important outcome of data sharing, the emphasis of this study was on investigating the use of clinical trial data to drive generation of new knowledge.

Methodology

The study examined the history and set-up of existing data sharing initiatives, their current research use, and impacts achieved. It also gathered views regarding the current barriers to research using IPD and the need for broader access to IPD, via a central access point or otherwise, by consulting researchers from commercial and non-commercial entities, staff of clinical trial data repositories, and individuals affiliated with clinical research such as representatives of funding organisations and patient groups. Further questions gauged the perceived level of importance of various characteristics of a future access model, in order to allow researchers from the academic, non-profit, and industrial sectors to contribute data and share research benefits, while protecting patient privacy and respecting the wishes of trial participants regarding re-use of their data.

The study utilised both quantitative and qualitative research methods including: desk research; a global online survey of clinical trial researchers and relevant stakeholders from sectors including universities and research institutes, hospitals and healthcare professionals, industry, research funders, patients groups, and regulators[1]; a stakeholder workshop; and targeted interviews with relevant individuals active in this area of research.

Existing data sharing initiatives

We analysed the key characteristics of 18 existing IPD sharing initiatives, and found they clustered into the following five broad categories (named to reflect their main properties):

- Collaborative groups of trialists/trial sponsors
- Disease-specific data repositories
- Public-funder mandated repositories
- Commercial trial repositories and data portals
- Open data sharing by individual research groups/units.

Collaborative groups of trialists/trial sponsors were created as research collaborations, rather than initiatives to enable broad data access, with the aim of addressing a specific disease area or task. These initiatives hold data from academic and commercial trials, generally from both control and treatment arms. Database staff harmonise the data on receipt. While access by researchers from outside the collaboration is possible, data providers retain control over their datasets and can veto requests for access. We found that these types of initiatives tended to yield substantial benefits for research and patients.

Examples of this type of data sharing initiative include the Early Breast Cancer Trialists' Collaborative Group (EBCTCG), the WorldWide Antimalarial Resistance Network (WWARN), and initiatives of consortia of the Critical Path (C-Path) Institute.

Disease-specific data repositories were created with the aim of accelerating development of treatments through enhanced data access for the wider research community, and tend to be funded by disease charities. These include disease-specific data from academic and commercial trials, but generally only from the control arm of the trial, or from "failed" trials in disease areas the company providing the data is no longer active in. Database staff harmonise data, and grant access following guidelines agreed with the data providers. Some databases in this group have seen data access levels of between 50 and 200 times per year. We found that the organisations coordinating these databases were spending, or planning to spend, a lot of effort on promotion, or on further incentivising their use as many researchers were not aware of these resources.

Examples of this type of data sharing initiative include the PRO-ACT database for Amyotrophic Lateral Sclerosis (ALS), the CODR database for Alzheimer's Disease, and the database of the Sylvia Lawry Centre for Multiple Sclerosis (MS) Research.

Funder-mandated repositories were created as a platform for depositing data from publicly funded research. Such databases have been implemented by several institutes of the US National Institutes of Health (NIH) for research funded through their grant mechanisms, and are often linked to other types of data (genetic data, observational studies) and/or biospecimens. Harmonisation of data occurs at different points – some databases require the data provider to standardise data to their requirements prior to submission, others leave this to the user of the repository. This may be reflected in the observed differences in usage levels, ranging from more than 100 requests per year for datasets harmonised by the depositor, to less than 20 requests per year for repositories with unharmonised data. Several repositories indicated issues with timely deposition of data by the original researcher, and the need for monitoring compliance.

[1] The survey population is a broad sample of self-selecting respondents and hence non-random, i.e. the results may not be representative of the clinical trial stakeholder community as a whole. For example, only 10% of respondents were from the private sector; and although it was a global survey, 57% of total respondents were based in the UK and 24% in the USA.

Examples of this type of data sharing initiative include the BioLINCC repository of the National Heart, Lung, and Blood Institute (NHLBI), the data repository of the National Institute for Diabetes, Digestive, and Kidney Diseases (NIDDK), and the National Database for Clinical Trials (NDCT) of the National Institute of Mental Health (NIMH).

Commercial trial repositories and data portals were created as a platform or portal to allow access to data, in the first instance from commercial clinical trials. They are fairly recent initiatives, providing (or starting to provide) access to data from industry-sponsored clinical trials. Most datasets are held on the trial sponsor's server, and access is granted by an independent review board following an agreed application process. In some cases, however, companies retain a right to deny access. Approved researchers can analyse the data within a secure environment; in exceptional cases, data transfer to the user's server may be considered. We found that researchers in general welcomed these new initiatives, but we also heard about cases when access to data via a "remote desktop" presented challenges to efficient analysis.

Examples of this type of data sharing initiative include the Clinical Study Data Request (CSDR) portal and the Yale University Open Data Access (YODA) Project.

Open access datasets have been made available for download by individual research groups or units to allow broad access to IPD without the need to contact the original researchers. While most of the data are available without any restrictions, part of the dataset may be withheld from the open access interface in order to prevent misinterpretation of the data (e.g. the randomisation code for the FREEBIRD database). This approach ensures complete accessibility to datasets. However, as datasets are currently held on many different data platforms, in distributed locations, researchers may need support to be able to find and combine these to maximise data use.

Examples of this type of data sharing initiative include the FREEBIRD and International Stroke Trial (IST) databases.

Current research practices

Over recent decades, the number of articles reporting IPD meta-analyses has risen considerably: while only 57 articles were published before 2000, an average of nearly 50 articles per year were published between 2005 and 2009.

Access to IPD provides a number of potential advantages over access to summary-level data. IPD allows researchers to analyse clinical trial data outside the original purpose of the trial, including "dividing up" datasets to identify specific subgroups of trial participants, and investigate time-sequence events, the effect of multiple factors in different combinations, and rare events.

We found examples of meta-analyses using IPD, in a number of disease areas, leading to an improved understanding of treatment benefits and risks, the development of prognostic models, the development of new analysis methods, and identification of inconsistencies in clinical trial data collection and assays. Enhanced access to IPD from clinical trials is expected to increase such research outcomes (e.g. in disease areas not yet investigated in this manner). In addition, secondary analyses of IPD could lead to novel insights drawing on the statistical power of the large volume of combined data, such as an understanding of causes and treatments for common conditions or symptoms, where there is significant heterogeneity across the patient population (e.g. pain and rheumatoid arthritis), or the occurrence of extremely rare events, such as adverse events in patients who were not considered at risk initially (e.g. stroke in young persons), drawing on the large scale of high-quality data available.

A survey of clinical trial researchers and relevant stakeholders, conducted as part of this study with a total of 386 respondents, indicated that respondents were predominantly involved in, or aware of, projects using IPD to address cancer (54%). This was followed by cardiovascular disease (36%), central nervous system or neuromuscular conditions (32%), mental health and behavioural conditions (23%), and digestive/endocrine, nutritional and metabolic diseases (23%). The principal research objectives of these projects were comparison of effects of different interventions (82%), and assessment of potential adverse effects of a drug or other interventions (61%). Projects made use of data on health outcomes (83%), demographics (78%), clinical laboratory test results (73%), medical history (71%), and adverse events (64%). A higher proportion of respondents from companies indicated use of adverse events data (84%) as compared to the overall survey population.

Survey respondents indicated that a variety of statistical methods and techniques had been used to analyse IPD. Most projects involved multivariate (75%) and univariate (47%) analysis, and logistic regression (51%). The use of less traditional techniques, such as data mining (22%), machine learning (9%), and genetic algorithms (6%), was also noted.

Two-thirds of survey respondents (66%) indicated that IPD analysed in projects they were involved in, or were aware of, were generated and held by the organisation where they worked. This figure was even higher for respondents from companies, rising to 80%. Only 21% had obtained the data through an established repository. Nearly half of the survey respondents indicated that they had not made any data requests in the previous year (43%). This figure was higher for respondents from industry (65%).

The majority of survey respondents, including respondents from companies, thought the ability to access IPD from clinical trials would enhance the quality of research (34%), or even influence the direction of research (36%). Only 7% thought that enhanced access would not change the research, and that all the IPD currently needed were accessible.

Current research barriers and preferred characteristics of a broader access model

The survey asked respondents to rate the impact of a range of current barriers to IPD research. Similarly, respondents rated the importance of a number of characteristics of a potential future IPD access model. Rankings of barriers and characteristics by perceived level of impact and importance are provided in Table 1 and Table 2, respectively, along with a summary of preferred characteristics of a future data sharing model in Box 1.

Survey respondents indicated that the most serious barriers to research projects involving IPD were current access to relevant existing datasets (with 66% indicating this had a “significant impact” on research projects or completely “blocked” those), and incomplete knowledge of what data exist (with 52% indicating a “significant impact” on or “blocking” research). This was followed by concerns over data not being mapped to a common standard, concerns about participant consent, and being restricted to data analysis on the data owner’s or repository server (respectively, with 42%, 41%, and 40% of survey respondents indicating a “significant impact” on, or “blocking” research).

Compared to the overall population of survey respondents, respondents from industry tended to be more concerned about providing competitive advantage to others, with 43% indicating “significant impact” on or “blocking” the research project, compared to 26% of all respondents.

Table 1 Current barriers to research using individual participant data

Barriers to current IPD research	Score
Access to relevant existing datasets	2.8
Incomplete knowledge of what data currently exist	2.4
Available data are not mapped to a common standard	2.3
Data can only be analysed on data owner’s/repository server	2.2
Concerns about participant's consent for data sharing	2.2
Concerns about sharing research proposals due to current proposal review practices	2.0
Ownership terms of research results are not favourable to researchers	2.0
Stringent credentials required for data requestors to access data	1.9
Concerns about identification of participants in the data	1.9
Concerns about providing competitive advantage to others	1.7

Survey question: “Based on your experience, please rate, on a scale from 0 to 4, the extent to which the following current barriers have an impact on researchers conducting projects involving individual participant data.” n range: 312 – 370.

Table 2 Preferred characteristics for access to individual participant data

Characteristics of future IPD access model	**Score**
Researchers are provided with technical information in relation to trials/data sets within the repository	3.2
Datasets include both commercial and academic trial data	3.0
Datasets can be downloaded for analysis	2.8
Data are harmonised and presented in a single format	2.8
Datasets from all trials are accessible on a central repository	2.7
Datasets include trial data from all regions of the world	2.7
Datasets include historical data	2.5
Researchers can use any analysis software on a central data access server	2.5

Survey question: "Please rate, on a scale from 0 to 4, the importance of the following statements relating to the characteristics of a future data repository for the type of research you/your colleagues may want to conduct." n range: 320 – 331.

Box 1 Preferred characteristics of a future sharing model for individual participant data from clinical trials, based on survey responses

- One central repository
- Repository includes data from academic/non-commercial and commercial trials
- Data are held by a trusted third party
- Datasets are curated to a high standard
- Access is reviewed by an independent review board
- Data can be downloaded to user's server
- Datasets are harmonised
- Historical data are included
- Data from all regions are incorporated

Referring to a potential future IPD access model, respondents saw benefits to all characteristics investigated in the survey[2]. The majority felt that it was most important to provide researchers with technical information in relation to accessed data sets (with 77% indicating this was of "significant importance" or "essential"). Respondents also considered it "significantly important" or "essential" that a future sharing initiative include both commercial and academic trial data (70%), that datasets could be downloaded for analysis (68%), and that data were harmonised and presented in a single format (65%). Industry respondents assigned less importance to all characteristics listed in the survey, with the largest difference in the importance attributed to the ability to download data for analysis (with only 33% of industry respondents indicating this was "significantly important" or "essential", compared to 68% of the total survey population).

Survey respondents' main concern about enhanced access to IPD was "losing control" over the data (40%), which included potential issues around patient privacy, misinterpretation or deliberate misuse of data, potential lack of appropriate patient consent for secondary analysis, or fear of criticism of the original analysis. The fear that data would be exploited without any benefit for the original researcher or study sponsor was also seen as impeding researchers' willingness to deposit data (34%). A smaller number of respondents listed concerns about the cost and effort involved in preparing and uploading datasets (11%). Views on what would stop researchers from requesting access covered a range of issues. The largest

[2] All average scores above 2 ("moderately important"), on a scale from 0 ("not at all important") to 4 ("essential")

number of respondents cited concerns over the quality of deposited data (34%), and a burdensome administrative approval process (20%).

Compared to the current situation, many more survey respondents were expecting to make requests for data should enhanced access became available. While 43% had not requested any data over the last year, only 14% thought they would not request any data from a database with a suitable access mechanism. Similarly, respondents from industry signalled a shift in the number of requests, with the proportion of those who requested data one or more times increasing from 35% last year to 77%.

Key considerations for a broader data access model

The views of experts consulted as part of this study were largely positive regarding enhanced access to IPD via a central access point, and the research opportunities afforded through such an initiative. However, it was evident that there were substantial concerns about the practicalities and potential risks. The benefits highlighted and concerns expressed are summarised below.

The **benefits of a central access model for IPD** were that it would:

- Increase transparency
- Save time and effort required for new analyses, by providing a single/a small number of access points to data, with legal aspects of data sharing already taken care of
- Enhance data quality and value, and uncover potential issues in data collection and interpretation
- Increase data discoverability
- Avoid duplication of research
- Draw in new research communities, by lowering the effort required for researchers external to the core clinical trial community to access data.

The **drawbacks of a central access model for IPD** were that it would:

- Disconnect the original researcher from the dataset, and hence increase the potential risk of incorrect analysis
- Represent a significant cost to data providers and repositories, with the possibility that many datasets will never be re-used
- Put researchers in resource-limited countries at a disadvantage, by placing data at the hands of experts in highly-funded research institutions without research benefit for those who collected the data.

In addition, survey respondents and interviewees highlighted the misalignment between the benefits of data sharing and rewards for the original researchers/trial sponsors. This ranges from the cost and effort of preparing datasets for sharing, the lack of recognition of the data contribution made, and a loss of control over the dataset leading to potentially increased risks, such as misuse of data, giving competitive advantage to other researchers or companies, and loss of intellectual property.

Regarding the **scope of a central IPD access model** suitable to maximise research benefits, experts consulted broadly agreed that:

- Data from academic and non-commercial trials should be provided alongside commercial trial data, as these often addressed complementary research questions. Respondents did not foresee any real barriers to combining the data.
- Access to trials from all geographic regions was desirable but not practically achievable. Data from disease areas that would especially benefit from access to global data should be prioritised, rather than trying to gather all data from the outset.
- Access to historical data was desirable, especially in research areas where long-term follow up data exist, but not practically achievable across all disease areas given the cost implications. Data from disease areas that would especially benefit from historical data

should be prioritised. Researchers conducting secondary analysis needed to be made aware of potential pitfalls when analysing these data, such as differences in data collection due to changes in medical technology.

- Other types of data should be combined with, or at least linked to the numerical data from clinical trials. This includes data from observational studies, which provide important long-term datasets complementary to the shorter-term clinical trial data, and images, which are essential in some disease areas.

Regarding **access mechanisms for a future IPD access model**, to enable the broadest possible use of the data while keeping risks at an acceptable level, most survey respondents (61%) and interviewees felt that reviewed access to datasets held by a trusted custodian was most suitable. However, while half of survey respondents (49%) considered the open access model least suitable, a substantial proportion (25%) chose this as the most suitable model, indicating that the scientific community does not currently have a broad agreement on this point.

Concerns about data continuing to be held by the original research or trial sponsor included potential data censorship, increased difficulty in aggregating data if datasets were stored in multiple locations, and the often restrictive nature of commercial trial sponsors' data environments.

Potential **risks of enhanced access to IPD, and suggested mitigation measures** included:

- The potential for breach of patient privacy. This could be limited by removing additional data parameters from the trial dataset, and/or by limiting access to *bona fide* researchers vetted via a robust review process.
- Providing competitive advantage for others. For academic research groups, this could be limited by allowing sufficient time for the original researcher to exploit the data before external access is granted, or requiring the original researcher to be informed of, or potentially involved in, any subsequent projects. This risk is difficult to address in a commercial setting.
- Rogue analysis, either through lack of knowledge or malicious intent. Suggestions for how this risk could be limited included:
  - extensive data curation of deposited data, and availability of detailed technical information alongside the dataset(s)
  - limiting access to research teams with the right skills and credentials
  - requiring submission of a clearly outlined research proposal along with the request for access
  - requiring the original researcher to be informed of, and potentially involved in, any subsequent projects.

In addition, it was evident that the lack of clarity on patient consent forms concerning secondary use of data needs to be addressed, with some interviewees calling for the development of a standard question addressing this issue, to be included on all forms going forward.

Regarding **data format and the analysis environment**, survey respondents and interviewees broadly agreed that data needed to be curated to a high standard to make those valuable. Respondents also thought it important that researchers could download data to their servers, or at least use any analysis software they wanted on the remote desktop provided by the repository or data portal. Harmonisation of data across datasets held in a central database was desirable, but not realistic on a global scale. A number of views were put forward as to when data should be harmonised (at the point of data deposition or when requested) and by whom (data provider or data user), to optimise capturing the full value of the data while keeping this effort to a reasonable level. Existing databases use a range of models, which may account for different levels of data requests from the research community.

Key findings

Key findings regarding the types of novel clinical research that may be possible using IPD from clinical trials:

1. Over the last decade, the number of publications of secondary analysis using existing IPD from clinical trials has significantly increased in the scientific literature.
2. A survey carried out as part of the study showed that respondents were predominantly involved in, or aware of, research using IPD in the areas of cancer and cardiovascular disease, with the principal objectives of comparing the effects of different treatments, assessing the occurrence of adverse events by subgroup analysis, identifying new biomarkers, and aiding the design of new clinical trials.
3. Outcomes achieved include the development of disease-progression models, qualification of new biomarkers and endpoints for use in clinical trials, and dose optimisation in patient subgroups.
4. Enhanced access to IPD was expected to broaden these outcomes further across other disease areas and enable novel research to improve our understanding of the causes of, and treatments for common conditions with significant heterogeneity across the patient population, as well as the causes of rare events.
5. The majority of survey respondents were involved in, or aware of, research using IPD held by their own organisations, or shared within the academic community. Although a range of data sharing initiatives is available, study informants indicated that these were used to a lesser degree.
6. Eighteen data sharing initiatives were examined in more detail, and found to group into the following five categories: collaborations of trialists/trial sponsors, disease-specific repositories, funder-mandated access repositories, commercial trial data portals, and open-access initiatives. Individual data holdings exhibit varying degrees of "openness", scale, and focus.

Key findings about the potential level of demand for a broader access model for IPD from academic and non-commercial trials:

7. The main barriers to research employing IPD were identified as issues related to "not knowing what data exist", i.e., discoverability, and access to data. The majority of survey respondents thought the ability to access IPD through a central data access point would enhance the quality, and even influence the direction, of their research.
8. Broader availability of data was expected to increase the number of requests for sharing of datasets, especially from industry survey respondents.
9. Most survey respondents considered it "significantly important" or "essential" that a future data access model includes both commercial and academic trial data. This view was mirrored by interviewees who felt that these data complement each other, and that it was hence important to be able combine both types for research.

Key findings about the mechanisms and practicalities that would need to underlie a broader access model for clinical trial data:

10. Survey respondents deemed reviewed access, rather than open access, the most suitable data access mechanism, and indicated that data should ideally be held by an independent data custodian, accessible via a central point, and curated to a high standard.
11. The majority of survey respondents felt that it was "significantly important" that data were harmonised and could be downloaded to the user's server. Respondents from industry assigned less importance to these factors. In interviews, the points were raised that harmonisation of the entire body of data within a large repository was not practically achievable, and it was advisable to harmonise data in research priority areas initially.

Recommendations

Based on the evidence gathered, the following set of recommendations[3] was developed:

1. **Link current data sharing initiatives and prevent further fragmentation of data landscape**
   - Promote the establishment of larger data holdings, with the clear aim of incorporating IPD from both commercial and non-commercial clinical trials.
   - Initiate enhanced information exchange between existing data sharing initiatives and support linking of existing repositories and data portals.
2. **Confirm demand for IPD**
   - Establish a central information website, or consider adapting current clinical trial registries, with profiles and links to existing repositories and data portals.
   - Ensure that funding streams for sharing and/or secondary analysis of existing clinical trial data are available to facilitate generation of new knowledge.
   - Monitor actual demand and research outcomes following promotion of available repositories and data portals.
3. **Address current barriers to IPD research in a joined-up approach**
   - Establish a central repository or data portal to facilitate access to IPD from clinical trial data. Such an effort may need to take the form of a small number of regional repositories on compatible data platforms.
   - Establish a global discussion forum of potential funders of IPD sharing initiatives to develop global support and a joined-up approach leading to the implementation of globally "linkable" IPD repositories and data portals.
4. **Develop a suitable repository platform**
   - Evaluate current data sharing platforms against desired characteristics, and for suitability for expansion, to develop and implement a data sharing platform drawing on best practice from existing repositories.
   - In case different data sharing requirements prevent some data providers from joining the "new" repository or data portal from the outset, continue dialogue to allow data linkage at a future point.
5. **Scale the repository**
   - Global reach: implement a pilot repository in one or a small number of regions to develop a robust, cost-efficient solution that could function as a model for future efforts in other regions.
   - Historical data:
     - Adopt a case-by-case approach to incorporate historical data, i.e., only in research priority areas or as mandated by individual funders.
     - Establish clear processes for deposition of historical data in priority research areas.

[3] During the publication process of the present study, the US Institute of Medicine (IoM) published their independent report "Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk" in January 2015, which sets out guiding principles and a practical framework for the responsible sharing of clinical trial data. The recommendations formulated in this study and those in the IoM report are broadly in line and complementary to each other. The authors of the present study hope that information presented in these studies will contribute to further the thinking of international stakeholders around the issues at hand.

- Other types of data:
  - Support information exchange with existing IPD sharing initiatives from other disciplines (e.g. public health).
  - Identify options for future linkage across databases.

6. Enable research while ensuring appropriate use of data

- Access modality:
  - Develop a repository model with reviewed access and data held by a trusted third party.
  - Carry out a detailed comparison of review parameters of existing data sharing initiatives to identify best practice and challenges, and develop an effective, stream-lined process.
  - Incorporate open access options to allow data providers to make suitably de-identified data available without review, should they wish to do so, and monitor demand, actualised risks, and research outcomes to inform further efforts.
- Data format:
  - Adopt a case-by-case approach to data harmonisation, rather than aiming to harmonise all data from the outset.
  - Establish processes for harmonisation of IPD across trials in priority research areas that offer individual funders the option of carrying out these activities.
  - Adopt or develop, and test, data handling tools to facilitate data deposition
  - Investigate staffing needs and "data manager" roles to provide support at the repository or academic institutions to assure high data quality.
- Data analysis environment:
  - Implement an IPD repository model that allows the user to download data when permitted by the data provider.
  - Investigate the need for a secure data environment for analysis, as determined by the proportion of data providers who would not be able to deposit data if it were downloaded by repository users.

Glossary

Adverse event: an unfavourable change in the health of a participant, including abnormal laboratory findings, that happens during a clinical study or within a certain time period after the study is over. This may or may not be caused by the intervention being studied. http://clinicaltrials.gov/ct2/about-studies/glossary

Arm: a group of participants in a clinical trial who receives specific interventions, or no intervention, according to the study protocol. This is decided before the trial begins. http://clinicaltrials.gov/ct2/about-studies/glossary. In this report, we refer to treatment arm for the group receiving the intervention that is the focus of the trial, and control arm, for the group of participants who do not receive this intervention.

Cleaned data: reviewed data from a clinical trial; dataset with internally consistent data entries.

Clinical Study Report (CSR)*:* a detailed analysis of study efficacy data and complete adverse event data. The CSR, and the cleaned dataset it is based on, are made available to regulatory agencies to support applications for market authorisation of a medicinal product. http://books.nap.edu/openbook.php?record_id=18610&page=58

Clinical trial: a research study that prospectively assigns human participants to health-related interventions in order to evaluate their effects on health outcomes (WHO definition, WHO International Clinical Trials Registry Platform (ICTRP) http://www.who.int/ictrp)

Cross-analysis: combining data from more than one clinical trial to conduct an analysis.

Data curation: the process of reviewing data from an individual clinical trial submitted to a database, to ensure that entries make sense and are internally consistent (i.e., cleaned). Unlike data harmonisation, data curation is not aimed at preparing data from more than one clinical trial for cross-analysis.

Data harmonisation / standardisation: the process of preparing data from two or more clinical trials to enable combining of datasets for analysis (the terms are used interchangeably).

Data portal: a central website allowing researchers to request data held by multiple individuals or organisations. A data portal does not store data. Datasets can be transferred from one or multiple providers to the user directly, or to a secure environment, for analysis.

Data provider: individual or entity responsible for enabling access to clinical trial data, e.g. through linking to a data portal or deposition in a repository.

Data repository: a database established to store data from multiple sources. Datasets may or may not be harmonised.

De-identification: The removal of parameters from data of trial participants in order to prevent identification of the individual.

Individual Participant Data (IPD): The definition of ‘participant-level data’ for this study is based on the Institute of Medicine’s Discussion Framework for Clinical Trial Data Sharing document. The data of interest are coded, transcribed, abstracted and cleaned from raw data sources and made accessible in a final cleaned and locked analysable database. http://www.iom.edu/Reports/2014/Discussion-Framework-for-Clinical-Trial-Data-Sharing.aspx

Market authorisation: product licence for a medicinal product, granted by regulatory agencies after review of efficacy and safety of the product, for the intended use.

Meta-analysis: a pooling and analysis of all appropriate data on a drug or other intervention. http://www.abpi.org.uk/our-work/library/industry/Documents/Clinical trial reporting - Definitions and guiding principles.pdf

Observational study: a study in which subjects are not randomly assigned to a treated group or a control group.

Original researcher: a researcher who was responsible for, or at a minimum directly involved in, conducting a clinical trial.

Randomised controlled trial: trials which typically compare an experimental intervention with usual care or a placebo, whereby patients are randomly assigned to the treatment or the control group.

Secondary analysis: an analysis of data that were collected for a different reason, in order to pursue a research interest which is distinct from that of the original work. http://sru.soc.surrey.ac.uk/SRU22.html

Summary-level data: the aggregated results of an analysis, without access to the individual data points from which these results were derived.

Acknowledgements

The project team would like to thank the many experts who contributed their knowledge and views to this study, by responding to our survey, participating in the workshop, giving interviews, and replying to our e-mailed questions. In particular, we are grateful to Will Greenacre and Jane Simmonds (Wellcome Trust) for their assistance throughout the study and the members of the Expert Review Group for their advice: Mike Clarke (Queen's University Belfast), Trudie Lang (University of Oxford), Fiona Reddington (Cancer Research UK), Matt Sydes (Medical Research Council Clinical Trials Unit at University College London), and Catrin Tudur Smith (University of Liverpool).

1. Introduction

This document is the final report of a study commissioned by the Wellcome Trust in April 2014. The study draws on independent research carried out by Technopolis Group under guidance from members of an Expert Review Group that was specifically set up to support this study.

1.1 Study objectives

The primary aims of this study were:

- To develop an understanding of the types of novel clinical research that are and would be possible using data from clinical trials, and to develop case studies of existing and future examples of such research and its benefits;
- To assess the potential level of demand for a broader access model for clinical trial data, including where demand lies, and whether there are appreciable differences between the academic and commercial research sectors that would dictate whether access to data from academic and other non-commercial trials should be provided alongside, or in combination with, data from commercial trials;
- To use the outcomes from steps 1 and 2 to begin to identify the key mechanisms and practicalities that would need to underlie a broader access model data, including: how data would be accessed; where it would be held (i.e. by the sponsor or creator of the data, or by an independent "gatekeeper"); what safeguards against inappropriate use or disclosure need to be in place; and whether data from multiple trials should be accessed through the same system and analysed together.

While increased transparency is an important outcome of data sharing, the emphasis of this study was to investigate the use of clinical trial data to *drive knowledge generation*, i.e. using previously generated data to tackle new research questions, to reduce duplication of trials, or to increase the efficiency of the research process by linking results from several trials.

1.2 Methodology

The study was conducted using several investigative techniques. These are briefly described in this section; additional detail is provided in Appendices C through G, as noted.

This study examined the history and set-up of existing data sharing initiatives, their current research use, and impacts achieved. It also gathered the views of members of the research community regarding the need for broader access to individual participant data (IPD), and current barriers to the use of IPD for secondary research. The study then investigated what characteristics future access models to IPD should have in order to allow researchers from the academic, non-profit, and industrial sectors to contribute data and share research benefits, while protecting patient privacy and respecting the wishes of trial participants regarding re-use of their data.

Our approach and methodology involved gathering of information from a variety of sources including publications, the web, researchers from commercial and non-commercial entities, staff of clinical trial data repositories, and individuals affiliated with clinical trials research such as representatives of funding organisations and patient groups.

The following points outline the methodological approach followed:

- Desk research: to compile data on existing IPD from clinical trial repositories and sharing initiatives, and their research impacts
- Stakeholder workshop, June 2014: to understand current uses and sharing mechanisms of IPD from clinical trials, and explore research opportunities given a large (hypothetical) central IPD repository (see Appendix G)
- Online survey: to gather views of researchers directly involved in running clinical trials or in the analysis of clinical trial data, and others with an interest in clinical trials, on the

need for enhanced access to IPD and the preferred characteristics for a future access model (see Appendix C). The survey was distributed by e-mail to relevant umbrella organisations, with the request for further dissemination via their mailing lists, or by directly e-mailing relevant individuals[4]. The identity and total number of individuals who received the survey is therefore unknown. As a result, the survey population represents a non-random sample of self-selecting respondents: it cannot be assumed that the views reported on in this study represent the views of the entire stakeholder community. A total of 446 survey responses were analysed in detail. Survey results are reported for the entire survey population, and for a sub-set of 45 survey responses from commercial entities[5].

- 40 Qualitative interviews with:
  - members of the research community, funders, and patient representatives: to discuss opportunities and challenges for enhanced access to IPD in more detail;
  - representatives from existing IPD repositories: to understand the rationale, characteristics, benefits and challenges of existing IPD sharing initiatives, and the level of demand experienced.

  (see Appendix E for list of interviewees and Appendix F for topic guides)
- Targeted e-mails: to address specific questions on database models or research use of IPD with relevant individuals (12 individuals, 9 of which were also interviewed).
- Development of case studies: to describe models of existing IPD sharing initiatives (11), and research impacts achieved (11), combining information gathered through desk research, survey responses, and interviews. All 22 case studies are presented in Appendix A.

1.3 Report structure

The remainder of the report is set out as follows:

- Section 2 – presents an **overview of existing data sharing initiatives**, and clusters these into broad "families" according to their intended purpose, the source and type of data they contain, and key characteristics. An example of a database from each cluster is provided.
- Section 3 – describes **how IPD are currently used and shared** in the academic, non-profit, and commercial research communities, and considers the advantages participant-level analysis has over summary-level data analysis. A range of case studies is presented to further illustrate **research opportunities opened up by access to and use of IPD**.
- Section 4 - provides an **overview of the results of a survey** conducted as part of this study, and presents survey respondents' views on current barriers to IPD research, important characteristics of a potential future IPD access model, the demand such data might see, and the main issues that could block researchers' willingness to make data available for, or use data from, such an initiative.

[4]This included industry associations, non-governmental funders such as charities, governmental funders, professional societies and other relevant associations, regulators, patient groups, research and clinical trials coordination networks, individual researchers, and staff of existing data sharing initiatives and repositories.

[5] Industry is a major provider of IPD from clinical trials, and represents a key stakeholder group. In contrast to academic institutions, companies may assign a single individual to represent the position of the *entire* organisation. We therefore report survey results from the "industry" group separately; taking the average across the overall population of respondents (386) would obscure the industry view (45 responses). Given the small size of the industry sample, the result for responses from all respondents versus those from "non-commercial" entities do not differ substantially and hence the latter is not reported separately.

- Section 5 – describes the **results of the survey in detail**, and integrates these with qualitative information gathered through **interviews**, from the **literature**, from responses to open survey questions, and from an **expert workshop** organised as part of this study. This section discusses views concerning the benefits and drawbacks of access to IPD via a central access model, considerations around the scope of such an initiative, incentives for data providers to share data, access mechanisms and potential risks, and a number of technical points such as data format and analysis environment.
- Section 6 – then presents the study **conclusions and recommendations**, drawing selectively on, summarising, and reflecting on the information and views presented in the previous sections of the report, in order to address the main study objectives regarding opportunities for research enabled by enhanced access to IPD from clinical trials, the potential level of demand for such data, and the key mechanisms and practicalities that would need to underlie a broader access model. This section also suggests **areas for further investigation** that were highlighted by study informants but were beyond the scope of this study.

Supporting material is set out in a series of **appendices**, for the purposes of reference and providing additional detail on the evidence presented in the main report:

- Appendix A presents a compendium of extended case studies of existing IPD sharing initiatives and research impacts achieved by combining and re-analysing IPD across trials
- Appendix B presents a more detailed comparison table of 18 existing data sharing initiatives investigated as part of this study
- Appendix C provides a more in-depth explanation of the survey and survey results
- Appendix D provides a copy of the survey questions
- Appendix E lists experts who have been interviewed, and contributed to the study
- Appendix F provides copies of the interview topic guides
- Appendix G outlines the programme and discussions of the expert workshop organised to inform this study in June 2014

1.4 Policy context

The ethical underpinnings of clinical trials, as research involving human subjects, require that the results be publicly available to inform medical practice as well as future research[6]. Sharing of information from clinical trials can occur at various levels:

1) Trials are registered before a trial is initiated, and updated with final enrolment numbers and date of completion, irrespective of the outcome of the trial,
2) a report at summary-level is made available, such as presented in a clinical study report (CSR), and/or
3) data at the level of the individual trial participant, collected as part of the trial, can be accessed by individuals not involved in the original study.

In the US, it is already mandatory[7] that all applicable clinical trials are registered in advance and results are made available subsequently on the national registry ClinicalTrials.gov. At the time of writing, the FDA and NIH were considering amendments to this requirement to

[6] World Medical Association. Declaration of Helsinki: ethical principles for medical research involving human subjects, as amended by the 48th World Medical Assembly, Somerset West, Republic of South Africa, October 1996. (Available at http://www.wma.net/en/30publications/10policies/b3/, accessed 19 Jan 2015)

[7] Food and Drug Administration. Amendments Act of 2007. Public Law No 110-85

include results of clinical trials addressing surgical techniques and of negative clinical trials. These would also require that more data be deposited in a more timely fashion.

In Europe, trial registration has been on-going since 2004 in the European Register (EudraCT), in line with the 2001 Clinical Trials Directive. A new Clinical Trials Regulation was adopted in 2014, to come into effect in 2016. At the time of writing, it was foreseen that a new publicly accessible clinical trial register would be developed to enable registration of all drug trials conducted in Europe before they start, and publication of summary results within one year of marketing authorisation, including negative results. This will be a significant step forward, as currently not all trials are registered and the findings of many trials are never made public. The European Medicines Agency (EMA), responsible for establishing the EU Portal and Database, recently announced that it had adopted a policy to publish clinical reports on all authorised medicines from 1 January 2015, and re-confirmed its plans to consult on plans to make IPD available in the future[8]. These changes will only apply to new trials.

On a global scale, the World Health Organisation (WHO) released a statement on public disclosure of clinical trial results in November 2014[9], calling for all trials to be registered prior to initiation, and updated with a summary of key findings after completion of the trial. The statement also proposes that the final results should be made available publicly through an open access mechanisms, or published in a peer-reviewed journal.

Traditionally, IPD collected in clinical trials could be difficult to access by researchers outside of the original research team (see Section 5.2.1). Recently, there have been renewed calls for responsible sharing of comprehensive participant-level data beyond the summary results reported in registries or the scientific literature[10] [11] [12] [13]. A driving force behind these initiatives has been the demand for greater transparency and better use of available data to scrutinise drug safety and efficacy, following a number of high-profile cases where companies stood accused of failing to release safety data (e.g. Merck's Vioxx[14], GSK's Paxil[15] and Takeda's Actos[16]). In addition, sharing of data, including at the participant level, is encouraged or expected by many funding organisations and by journals – however, often without explicit guidance on the exact mechanism and timing. Examples of funding bodies with a data sharing policy include the UK National Institute for Health Research's Health Technology Assessment Programme[17], the UK Medical Research Council[18], the Wellcome Trust[19], the Bill and Melinda Gates Foundation[20], and the US National Institutes of Health (NIH)[21].

[8] http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2014/10/news_detail_002181.jsp&mid=WC0b01ac058004d5c1. Announced on 15 October 2014, this policy applies to all products for which the application was submitted in or after 2014. Accessed 19 Jan 2015

[9] http://www.who.int/ictrp/results/Draft_WHO_Statement_results_reporting_clinical_trials.pdf?ua=1 (accessed 5 Nov 2014)

[10] Goodlee, F (2012) Clinical trial data for all drugs in current use. BMJ 345:e7304.

[11] Ross, JS et al (2012) The importance of clinical trial data sharing: Toward more open science. Circulation: Cardiovascular Quality and Outcomes 5(2):238-240.

[12] Loder, E (2013) Sharing data from clinical trials: where we are and what lies ahead. BMJ 47:f4794.

[13] Mello, MM et al (2014) Preparing for Responsible Sharing of Clinical Trial Data. NEJM 369: 1651.

[14] http://www.nature.com/news/2007/071113/full/450324b.html (accessed 5 Nov 2014)

[15] Doshi, P (2013) Putting GlaxoSmithKline to the test over paroxetine. BMJ 347:f6754

[16] http://www.bloomberg.com/news/2014-04-07/takeda-actos-jury-awards-6-billion-in-punitive-damages.html (accessed 5 Nov 2014)

[17] http://www.nets.nihr.ac.uk/about/adding-value-in-research (accessed 5 Nov 2014)

[18] http://www.mrc.ac.uk/research/research-policy-ethics/data-sharing/ (accessed 5 Nov 2014)

[19] http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/ (accessed 5 Nov 2014)

[20] http://www.impatientoptimists.org/Posts/2014/11/Knowledge-is-Power (accessed 11 Dec 2014)

[21] http://grants.nih.gov/grants/policy/data_sharing/ (accessed 16 Nov 2014)

The US Institute of Medicine (IoM) published their report "Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk" in January 2015[22], which sets out guiding principles and a practical framework for the responsible sharing of clinical trial data. The report concludes that sharing data is in the public interest, but that a multi-stakeholder effort was needed to develop a culture, infrastructure, and policies to foster responsible sharing. The study was sponsored by 23 public- and private-sector sponsors in the United States and abroad, demonstrating the scale of current interest in this area.

1.5 Summary

Recent legislation and initiatives have already increased the availability of clinical trial data, predominantly at summary level, but also at the level of the individual participant. Enhanced access not only allows for more transparency in clinical research, but can also drive knowledge generation, allowing researchers to tackle new research questions, to reduce duplication of trials, or to increase the efficiency of the research process by linking results from several trials.

Employing a range of investigative techniques, this study examined the history and set-up of existing repositories, their current research use, and impacts achieved. It also gathered the views of members of the research community regarding the need for broader access to IPD and the characteristics a potential future access model should have in order to allow researchers from the academic, non-profit, and industrial sectors to contribute data and share research benefits, while protecting patient privacy and respecting the wishes of trial participants regarding re-use of their data.

[22] Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk http://www.iom.edu/Reports/2015/Sharing-Clinical-Trial-Data.aspx (accessed 16 Jan 2015)

2. Existing data sharing initiatives

At present, most clinical trial datasets are held by the researcher, research units, and companies who conducted or sponsored the trials. External researchers who want to make use of existing datasets have to submit multiple requests and negotiate access with each individual data provider. Over the past 30 years, a number of initiatives have been launched in order to facilitate this process, and/or to address research questions in specific disease areas. This section provides an overview of a range of existing IPD sharing initiatives.

2.1 Landscape of data sharing

We investigated the history and characteristics of 18 existing IPD sharing efforts. Taking these databases' principal aims and characteristics into account, we found that they clustered into the following five "families"[23]:

1. Collaborative groups of trialists/trial sponsors
2. Disease-specific data repositories
3. Public-funder mandated repositories (NIH)
4. Commercial trial repositories and data portals
5. Open data sharing by individual research groups / units

While these clusters are not precisely delineated, with overlaps for some dimensions and intra-group variation for individual initiatives, they provide a useful description of the existing data sharing landscape. The following sections outline the key features of each cluster in turn, and provide one case study each to illustrate further. Additional (and extended) case studies of data sharing initiatives are available in Appendix A.

2.1.1 Collaborative groups of trialists/trial sponsors

Collaborative groups consist of researchers from academia and industry who contribute their data to a common database. They are created to address a specific question or progress a particular area of research. Collaborative groups may contain members from academia, industry, government, and non-profit organisations. In most cases, the coordinators of the collaboration identify potential group members through a review of published and unpublished trials, and invite them to submit their data and join the collaborative group. Data, generally from treatment and control arms of clinical trials, are harmonised by staff at the database, after which datasets can be combined for analysis. The original researchers who contributed data retain ownership and control over their datasets, and remain involved in the research. This often leads to co-authorship in any ensuing publications. As opposed to initiatives in the other four clusters presented in this section, collaborative groups are not set up with the intent to share data outside of the group; however, all groups interviewed confirmed that this was possible in principle. External researchers have to obtain permission directly from the investigator who contributed the data to the database. The database then enables access to the relevant (harmonised) datasets.

[23] We are aware that categories of data sharing activities were recently developed by other groups. While our analysis of repositories was focussed on the history and intent of existing initiatives, along with key access and data storage characteristics, other efforts have taken a more conceptual approach. Hence, our work resulted in an overlapping, but also different set of categories. For example, the IoM's Discussion Framework for Clinical Trial Data Sharing[22] defined four categories: "Open Access", "Controlled Access to Individual Company, Institution, or Researcher Data", "Controlled Access to Pooled or Multiple Data Sources", and "Closed Partnerships/Consortium". Categories "Open Access" and "Closed Partnership/Consortium" correspond to our clusters 5 ("Open data sharing by individual research groups /units") and 1 ("Collaborative groups of trialists/trial sponsors"), respectively. (We chose a different name for cluster 1, as the intent of databases in this cluster is to enable collaboration of data providers, rather than to form a closed partnership.) Our remaining clusters 2, 3 and 4 all fit into the IoM category "Controlled Access to Pooled or Multiple Data Sources". Mello et al[13] describe four possible models with current examples, focussed on who controls access, if the raw data or only analysis results are made available, and which criteria are used for releasing data. Again, while there were overlaps with the clusters we identified, the 18 existing databases we examined were not optimally described by these categories.

This study found that this type of data sharing initiative has yielded high impact research outcomes. They do however require continuous funding for staff to coordinate the collaboration, on top of costs for data infrastructure. They are also less accessible to the wider research community, and have rarely (or not at all) been used by external researchers; this may however reflect the fact that the data has already been analysed exhaustively by the collaboration group itself.

Examples of collaborative groups of trialists include the Early Breast Cancer Trialist Collaborative Group (EBCTCG) (see Box 2), the Worldwide Antimalarial Resistance Network (WWARN), and the Critical Path (C-Path) Institute's research consortia.

The key features of this group are:

- Created as a research collaboration, rather than an initiative to enable data access
- Disease or task-specific
- Includes data from academic and commercial trials
- Generally includes data from control and treatment arms
- Database staff harmonise data
- Data providers retain ownership and control of their datasets, and remain directly involved in research project
- External access possible, in principle, but not primary objective of the initiative.

2.1.2 Disease-specific data repositories

Disease-specific repositories are set up with the aim of progressing and accelerating the development of treatment options for defined patient groups, by opening up access to clinical trial data to the wider research community. The repositories are generally initiated by non-profit organisations that approach potential data providers from academia and industry to gather all relevant data into a single database. As opposed to databases in the other four clusters, disease-specific repositories generally contain only control arm data; where data from the treatment arm are included, it tends to be from "failed" trials (i.e., those that did not lead to market authorization, or for diseases that the company is no longer pursuing). All data are standardised by dedicated staff after which datasets can be combined for analysis. In order to gain access to the data, researchers have to register and submit a research proposal. Applications are reviewed by repository staff according to guidelines that were developed and agreed with the contributing data providers. Data providers are kept informed through update reports at overview level, but do not play a role in individual data access decisions.

This type of data sharing initiative requires a high level of initial input from database staff to gather and harmonise data, shifting effort away from the researcher, with the aim of encouraging broad research use. This study found that the coordinating organisations were spending, or planning to spend, a lot of effort on promoting the databases, or on further incentivising their use. Many researchers were not aware of these resources, and interest levels are also likely to depend on (and be limited by) the size of the research community investigating the disease the database is focussed on. Some databases in this group have seen access levels of between 50 and 200 times per year.

Box 2 Example of a collaborative group of trialists / trial sponsors

The Early Breast Cancer Trialists' Collaborative Group (EBCTCG)

The Early Breast Cancer Trialists' Collaborative Group (EBCTCG) overview is a major collaborative endeavour that investigates the treatment of women with early (or operable) breast cancer. The collaboration first came together in 1983 to discuss the combination of the results of randomised clinical trials of tamoxifen and chemotherapy. Currently, the collaboration involves around 300-400 research groups across the world - essentially all groups conducting randomised trials on treatments of women with early-stage breast cancer, where a main outcome is mortality.

The EBCTCG overview takes place in cycles lasting approximately 5 years, going through the stages of study identification, data collection, processing and analyses, presentation and discussion of the results by the collaborating researchers, and publication of these results. Following extensive searches for published and unpublished trials, investigators from academia and industry who conduct randomised trials on early breast cancer with survival as the major outcome are invited to join the group. Trial data that address one of EBCTCG's priority research questions are collected in the central database, located at the Clinical Trial Studies and Epidemiological Studies Unit (CTSU) at the University of Oxford, UK, the base for the EBCTCG Secretariat. The database staff select the variables relevant to the EBCTCG in discussion with the EBCTCG Steering Committee and converts the submitted data to a highly structured format (excluding data that might be submitted but are not required for the overview, such as data on quality of life measures or some toxicity effects, as these are outside the remit of the group).

While the data are held in Oxford, the contributing investigators retain ownership of their data. Other researchers can request access to datasets in the database to conduct their own analyses, but have to contact the data owner for approval before it can be transferred by the EBCTCG Secretariat[24].

The EBCTCG database currently holds data from around 700 clinical trials.

Examples of disease-specific data repositories are the PRO-ACT repository for amyotrophic lateral sclerosis (ALS) (see Box 3), the C-Path Online Repository for Alzheimer's Disease (AD), and the Sylvia Lawry Centre for Multiple Sclerosis (MS).

The key features of this group are:

- Created to accelerate development of treatments by enabling data access to the wider research community
- Disease or task-specific
- Includes data from academic and commercial trials
- Generally only data from control arms of trials
- Database staff standardises data
- Access to data is granted by repository staff, following guidelines agreed with data providers.

[24] This, however, happens rarely, as the data have already been exhaustively analysed through the EBCTCG overview.

Box 3 Example of a disease-specific data repository

The PRO-ACT database

The PRO-ACT[25] database is a project coordinated and implemented by the non-profit organisation Prize4Life, whose mission is to accelerate the discovery of treatments and a cure for ALS (amyotrophic lateral sclerosis, also called motor neuron disease), in partnership with the North Eastern ALS Consortium and the Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital.

PRO-ACT went live in December 2012, after 2 years of discussions with sponsoring companies, followed by a period during which the NCRI cleaned, harmonised, aggregated and anonymised the data. It currently houses around 8,500 ALS patient records from 17 completed Phase II/III ALS clinical trials (10 commercial and 7 academic trials). For most trials, both treatment and control arm data are included, where trials generally "failed" (i.e. results were clinically and statistically not significant), with only one 'modestly effective' treatment currently available on the market (extending patients' lives by an average of 2 months).

The database is open to anyone with an acceptable research proposal. Eligibility guidelines were agreed by a data access committee, which includes representatives of the companies contributing the data. Prize4Life staff review individual requests for fit with these guidelines, and keep the data access committee informed through update reports at overview level. If the request is approved, researchers can download all or some of the data types in the database, as Excel or text files, and can run their analyses as needed.

By July 2014, Prize4Life had received over 350 requests from researchers from industry and academia. This relatively high number of requests is likely to be a result of a promotion campaign for PRO-ACT, and the attention the database received through the prize challenge (see Section 3.4.2).

2.1.3 Public-funder mandated repositories

While many public funders of research require grantees to provide access to generated data on request[17,18], only some of the institutes of the US National Institutes of Health (NIH) appear to have set up specific databases to gather data in a central location. These "NIH databases" contain only data from academic investigators funded through their grant programmes, i.e., while they do not focus on a specific disease, the data mirror the broad areas of research supported by the individual NIH institute (such as "heart, blood, and lung diseases" or "treatments of drug abuse"). Frequently, the NIH databases are linked to other types of data (genetic data, observational studies) and/or biospecimens. Standardisation of data occurs at different points – some databases require the data provider to standardise data to their requirements before submission. Others curate the data upon receipt, but do not harmonise data and leave this to the user of the repository. In most cases, access to the database is subject to an administrative review by repository staff.

Individual NIH Institutes appear to have developed data repositories independently, and the existing IPD databases are not linked. The different models for harmonisation of data may reflect the observed differences in usage levels, ranging from more than 100 requests per year for datasets standardised by the data provider, to less than 20 requests per year for repositories where users need to harmonise across datasets. Several repositories indicated that timely compliance with the requirement to deposit data was an issue, despite the fact that grantees are encouraged to include a separate budget for the preparation of data for submission in their research proposals.

Examples of (NIH) public-funder mandated repositories include BioLINCC (National Heart Blood and Lung Institute, NHLBI) (see Box 3), the data repository of the National Institute

[25] The acronym PRO-ACT stands for Pooled Resource Open-Access Clinical Trials.

of Diabetes and Digestive and Kidney Diseases (NIDDK), the National Database for Clinical Trials in Mental Health (National Institute of Mental Health, NIMH), and the Immune Tolerance Network TrialShare database (National Institute of Allergy and Infectious Diseases, NIAID).

The key features of this group are:

- Created as a platform to enable access to data from publicly-funded research
- Not specific to a particular disease, but within the research area covered by the funding institute
- Includes data from academic trials only (NIH grant holders)
- Includes treatment and control arm data
- Responsibility for data curation and harmonisation differs between repositories
- Access to data is granted by repository staff, following institutional guidelines
- Often includes, or is linked to repositories containing, other types of data (genetic, observational studies) and/or biospecimens.

Box 4 Example of a public-funder mandated repository

The BioLINCC repository of the National Heart, Lung, and Blood Institute

BioLINCC was set up by the US NIH's National Heart, Lung, and Blood Institute (NHLBI) in 2000 to facilitate sharing of datasets and biospecimens from NHLBI-funded research. It contains treatment and control arm data from 82 clinical trials and 33 observational studies on heart, blood, and lung diseases (excluding cancer). The most "famous" dataset included is probably the (observational) Framingham Heart Study, which has been running since 1942. Where available, BioLINCC also provides access to biospecimen collections associated with these studies, which are stored in the BioLINCC biorepository.

Researchers wanting to request data submit information on the study protocol or proposed research plan and the data security measures to be used. They also have to provide ethical approval from their Institutional Review Board, or a waiver statement, for any level of access to the data. The request goes through an administrative review by the Repository Allocation Committee (NHLBI staff), confirming that the proposed use of the data is consistent with the data agreement. After approval, data are transferred to the researcher in the format that it was received in, with the NHLBI not offering custom data solutions. (Data harmonisation is under consideration, with its potential advantages being balanced against the high burden of cost.)

Since 2000, approximately 640 investigators have received data. Nearly 35% of the requested datasets include data from clinical trials, i.e. around 220 requests over 14 years. (It should be noted that the actual re-use frequency of the datasets may be masked by the fact that most studies supported by the NHLBI share data readily with outside investigators, and do not require the involvement of BioLINCC.)

2.1.4 Commercial trials repositories and data portals

Recent initiatives have enabled, or are in the process of enabling, access to data from industry-sponsored clinical trials, via a repository or coordinated through a data portal. These initiatives were at least partially developed in response to calls for greater transparency in commercial clinical trials, including recent policy developments at the EMA (see Section 1.4) and guidelines published by the European Federation of Pharmaceutical Industries and Associations (EFPIA) and the Pharmaceutical Research and Manufacturers of

America (PhRMA) in January 2014[26]. Commercial trial sharing initiatives do not specifically aim to accelerate research progress in defined disease areas. They mostly provide access to data held within a secure environment, but for some datasets, options for analysis outside the data platform can, in principle, be discussed. Most datasets are held on the trial sponsor's servers. Access is granted by an independent review board following an agreed application process. In some cases, however, companies retain a right to deny access.

There are currently two active examples of this type of data sharing initiative – a repository accessible via the ClinicalStudyDataRequest (CSDR) portal, and the Yale University Open Access (YODA) Data Project. While data sharing initiatives in this group contained only data from commercial trials at the time of writing, they are in principle not limited to industry data; for example, the YODA Data Project is open to incorporating data from academic trials in the future. Recently, CSDR has approved requests for access to datasets from different trial sponsors (see Box 5)[27].

While existing repositories / portals have received between 20 and 50 data requests in their first year, this may increase as more data are added and access mechanisms become more established. For example, researchers have reported issues working with the "remote desktop" interface of the CSDR portal; it remains to be seen how this develops as the initiative matures.

The key features of this group are:

- Recent initiatives, at least partially in response to calls for greater transparency
- At the time of writing, included data from commercial trials only
- Not specific to a particular disease, but within the research area of contributing companies
- Include treatment and control arm data
- Access to data is granted by an independent review board.

[26] Available at http://transparency.efpia.eu/responsible-data-sharing (accessed 19 Jan 2015)

[27] We did not investigate the *practicalities* of conducting analyses across datasets from different sponsors (e.g. issues with harmonisation of data, and combining datasets) as part of this study.

Box 5 Example of a commercial trial data portal

The Clinical Study Data Request portal

Clinical Study Data Request (CSDR) is an online portal enabling researchers to view available studies conducted by a number of clinical trial sponsors and to request access to the underlying anonymised individual participant data. The repository was initiated by GlaxoSmithKline (GSK), and launched in May 2013. At the time of writing (Dec 2014), 1599 distinct datasets were available via the CSDR portal, from eight pharmaceutical companies: Boehringer Ingelheim (190), GSK (1058), Lilly (81), Novartis (6), Roche (60), Takeda (145), UCB (21), and ViiV Healthcare (38). Another three companies (Astellas, Bayer, and Sanofi) will make their datasets available via the CSDR portal in the future. These data concern medicine that had received market authorisation or those from terminated research programmes. In addition, some sponsors may accept enquiries from researchers about the availability of data from other studies not currently listed on the website. Both raw and analysis-ready datasets are provided, along with supporting documentation.

Researchers may gain access to requested datasets by submitting a research proposal. The first step is a ”requirements check” by study sponsors to ensure the information is complete and meets the requirement for informed consent. Some sponsors may decline access to their data in exceptional circumstances, for example, where there is a potential conflict of interest or an actual or potential competitive risk. In a next step, the proposal is vetted by an Independent Review Panel for overall feasibility, scientific rationale, and relevance of the proposal's approach, and qualification of the team. Once the request has been approved and a data sharing agreement signed, the relevant sponsor(s) provide access to anonymised data in a password-protected workspace. Researchers can combine data from different studies, conduct research using statistical software provided (SAS and R), and finally download their analyses. Controls are in place to prevent researchers downloading the data to their computer.

All approved requests with signed data sharing agreements can be viewed on the CSDR website. Of the 23 projects with signed data sharing agreements by May 2014, 12 have gained access to data from one trial only, 9 gained access to data from 2 or 3 trials; the remaining two projects gained access to 8 and 11 datasets. Only one of the projects involves data from more than one trial sponsor. These projects cover a range of research objectives, including the identification of predictive markers or risk factors in individual patients, development of prognostic models, comparison of effectiveness and safety of drug combinations, dose optimisation, identification of improved clinical endpoints, and improvement of future trial design.

2.1.5 Open data sharing by individual research groups / units

Many funders and medical journals require researchers to provide access to their data if requested by external researchers as a condition of funding and publication. This occurs generally through direct contact between the requestor, who is interested in the published study, and the primary investigator, who then transfers the IPD (peer to peer transfer).

A number of individual research units have enabled external researchers to view and download IPD from clinical trials through a web-interface, to permit additional secondary analyses and facilitate the planning of future trials. The anonymised datasets can be freely downloaded. Examples of open datasets are the FREEBIRD database at the London School of Hygiene and Tropical Medicine, and the International Stroke Trial (IST) Database at the University of Edinburgh.

This approach ensures complete accessibility, but data are held on many different data platforms in distributed locations, likely limiting discoverability and use.

While most of the data are available without any restrictions, FREEBIRD withholds part of the data (the randomisation code) in order to prevent misinterpretation (see Box 6).

The key features of this group are:

- Created to allow broad access to IPD without having to contact the original researchers
- Data from within the research areas of the contributing research unit
- Includes treatment and control arm data
- Anonymised dataset can be downloaded without any restrictions
- For FREEBIRD only: Randomisation code is withheld but can be requested from the data provider.

Box 6 Example of open data sharing by an individual research group / unit

The FREEBIRD database

The FREEBIRD database was set up in 2011 by the Clinical Trials Unit at the London School of Hygiene and Tropical Medicine (LSHTM). It currently consists of two large clinical trials, CRASH and CRASH-2, which investigated the effect of treatments for adult trauma patients. Together, the studies involved more than 30,000 patients from across 49 countries. The database set-up was funded by the UK's National Institute for Health Research (NIHR), and running costs are absorbed by the Clinical Trials Unit budget. It is strongly supported by the consumer network and includes consumer testimony about the importance of data sharing.

FREEBIRD is available to any member of the public. After filling in a simple registration form, the anonymised dataset can be downloaded in CSV format, without an approval process. In addition, the randomisation code is withheld (i.e. the data do not show which treatment was allocated to which patient), in order to prevent users from drawing inappropriate conclusions about treatment effects, which the trial design would not support. Users can request the randomisation code, accompanied by a detailed proposal for the study team to review for suitability. To date, this has occurred twice; for one project, the protocol is in preparation and for the second, the requester did not respond to the study team's additional questions.

One of the underlying premises for making the CRASH and CRASH-2 data widely available is that the LSHTM investigators do not consider themselves "owners" of these data: it was generated in more than 300 hospitals around the world, by numerous researchers.

Table 3 summarises these key characteristics of each of the categories of data sharing initiatives. Additional information is available in Appendices A and B:

- Appendix A provides case studies, describing the history, objectives, implementation and research uses of a selection of data sharing initiatives.
- Appendix B provides a more detailed comparison table of individual data sharing initiatives.

Table 3 Individual participant data sharing initiatives

	Collaboration of trialists/trial sponsors	**Disease-specific data repository**	**Funder-mandated access**	**Commercial trial repository and data portal**	**Open data sharing by individual research units**
Disease-specific data	Yes	Yes	No	No	Yes[d]
Data source	Academic and commercial	Academic and commercial	Academic	Commercial	Academic
Trial arm included	Both	Control arm only[a]	Both	Both	Both
Data harmonised by	Database staff	Database staff	Data provider/ Database staff / User	Data user	n/a
Access approved by	Data provider	Database staff (scientific)[b]	Database staff (administrative)	Independent review board	None
Data held by	Data custodian	Data custodian	Data custodian	Trial sponsor	Original research unit
Funding source	Public funders, industry, foundations	Public funders, industry, foundations	Public funders (US NIH)	Industry	Public funders
Examples	EBCTCG, C-Path consortia, WWARN, IMPACT, EORTC	PRO-ACT, C-Path CODR AD, Sylvia Lawry Centre, Project Data Sphere	NIH: NIDDK, BioLINCC, NCDT, NIDA, ITN TrialShare[c]	CSDR, YODA, Bristol-Myers Squibb/Duke U	FREEBIRD, IST

[a] some treatment arm included, generally for academic trials or inconclusive commercial trials

[b] using guidelines agreed with each of the original researchers or data providers

[c] open access

[d] research area of individual research unit

2.2 Summary

We analysed the key characteristics of 18 existing data sharing initiatives, and found they clustered into five broad "families" (named to reflect their main properties):

- Collaborative groups of trialists/trial sponsors,
- Disease-specific data repositories,
- Public-funder mandated repositories,
- Commercial trial repositories and data portals, and
- Open data sharing by individual research groups / units.

Collaborative groups of trialists/trial sponsors were created as research collaborations, rather than initiatives to enable broad data access, with the aim of addressing a specific disease area or task. These initiatives include data from academic and commercial trials, generally from both control and treatment arms. Database staff harmonise the data on receipt. While access from researchers outside the collaboration is possible, data providers retain control over their datasets and can veto requests for access. We found that these types of initiatives have yielded substantial benefits for research, and patients.

Disease-specific data repositories were created with the aim of accelerating development of treatments through enhanced data access for the wider research community, and tend to be funded by disease charities. They include disease-specific data from academic and commercial trials, but generally only from the control arm of the trial, or from "failed" trials in disease areas the company is no longer active in. Database staff harmonise data on receipt. Access is granted by repository staff, following guidelines agreed with the data providers. We found that the organisations coordinating these databases were spending, or planning to spend, a lot of effort on promotion, or on further incentivising their use. Many researchers were not aware of these resources, and interest levels are likely to depend on the size of the research community investigating the disease the database is focussed on. Some databases in this group have seen data access levels of between 200 and 50 times per year.

Funder-mandated access repositories were created as a platform for depositing data from publicly funded research. This type of database has been implemented by several institutes of the US National Institutes of Health for research funded through their grant mechanisms. Databases are often linked to other types of data (genetic data, observational studies) and/or biospecimens. Individual NIH Institutes appear to have developed data repositories independently. As a result, harmonisation of data occurs at different points – some databases require the data provider to standardise data to their requirements before submission, others leave this to the user of the repository. This may be reflected in the observed differences in usage levels, ranging from more than 100 requests per year for datasets harmonised by the depositor, to less than 20 requests per year for repositories with unharmonised data. Several repositories indicated issues with timely deposition of data by the original researcher.

Commercial trial repositories and data portals were created as a platform or portal to allow access to data, in the first instance from commercial clinical trials. They are fairly recent initiatives, providing (or starting to provide) access to data from industry-sponsored clinical trials. Most datasets are held on the trial sponsor's server, and access is granted by an independent review board following an agreed application process. In some cases, however, companies retain a right to deny access. Approved researchers can analyse the data within a secure environment; in exceptional cases, data transfer to the user's server may be considered. We found that researchers in general welcomed these new initiatives, but also heard that in some cases, access to data via a "remote desktop" presented significant challenges to efficient analysis.

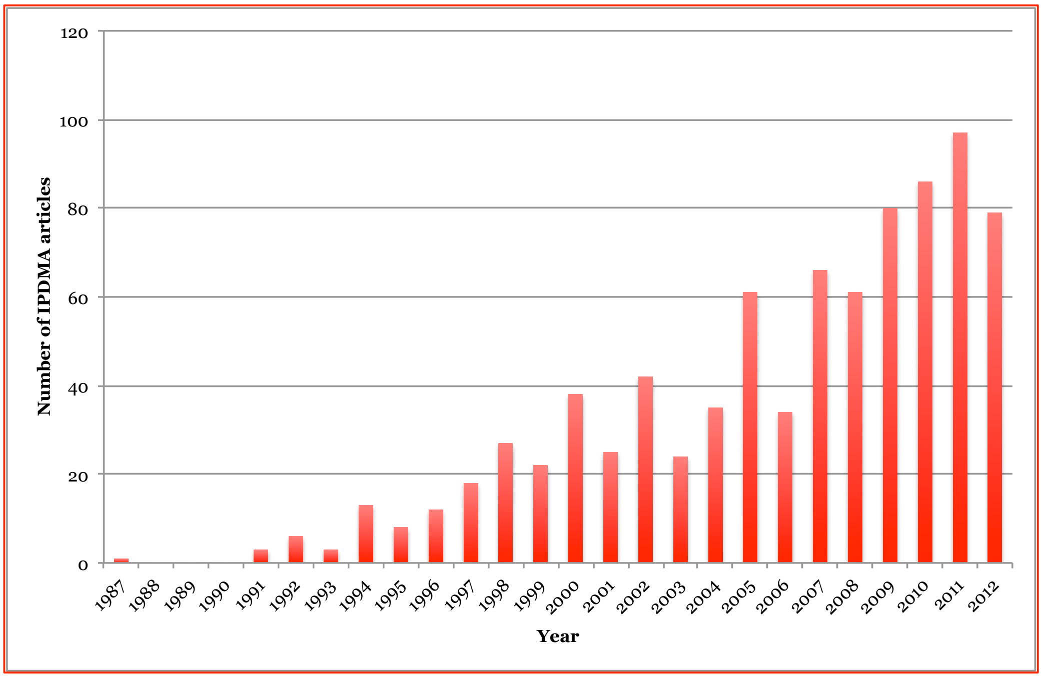
Open access datasets have been made available for download by individual research groups or units to allow broad access to IPD without having to contact the original researchers. While most of the data are available without any restrictions, one of these initiatives (FREEBIRD) withholds part of the data (the randomisation code) in order to prevent misinterpretation of the data. The code is available on request and after discussion with the original researchers. This approach ensures complete accessibility to datasets, but data are held on many different data platforms, in distributed locations. To maximise discoverability and use, researchers may need support to be able to find and combine these.

3. Current research practices

3.1 Current uses of individual participant data by the research community

Over recent decades, the number of articles reporting IPD meta-analyses has risen considerably: while only 57 articles were published before 2000, an average of 50 articles per year were published between 2005 and 2009[28]. The 383 articles published up to March 2009 focussed predominately on cancer, cardiovascular disease, and diabetes, and most studies assessed whether a treatment or intervention was effective, often in subgroups of patients. Nearly a quarter (22%) assessed risk factors for disease onset or prognostic factors for disease outcome. A recent review found that the number of IPD meta-analyses published annually continues to rise (see Figure 1), with each article including a median of 8 studies involving 2,563 patients[29].

Figure 1 Increase in individual participant data meta-analysis publications



Source: Reproduced from Huang et al[29]

The survey conducted as part of this study (see Appendix C) provided similar results in relation to the research areas that are most common for IPD meta-analyses. The majority of respondents indicated that projects using IPD they were involved in, or aware of, addressed cancer, followed by cardiovascular disease, central nervous system or neuromuscular conditions, mental health and behavioural conditions, and digestive/endocrine, nutritional and metabolic diseases (Figure 2a). These results may reflect disease areas with unmet clinical need that have both data and funding sources or large market opportunities available.

Most respondents indicated that the principal research objectives of these projects were comparison of effects of different interventions, and assessment of potential adverse effects

[28] Riley, RD et al (2010) Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 340: c221.

[29] Huang, Y et al (2014) Distribution and Epidemiological Characteristics of Published Individual Patient Data Meta-Analyses. PLoS ONE 9(6): e100151.

of a drug or other interventions (Figure 2b). A high proportion of research projects also aimed to assess the effects of interventions in specific sub-groups of the trial, e.g. children, ethnic minorities, or patients with different disease progression types, to aid the design and methodology of clinical trials, and to identify new biomarkers.

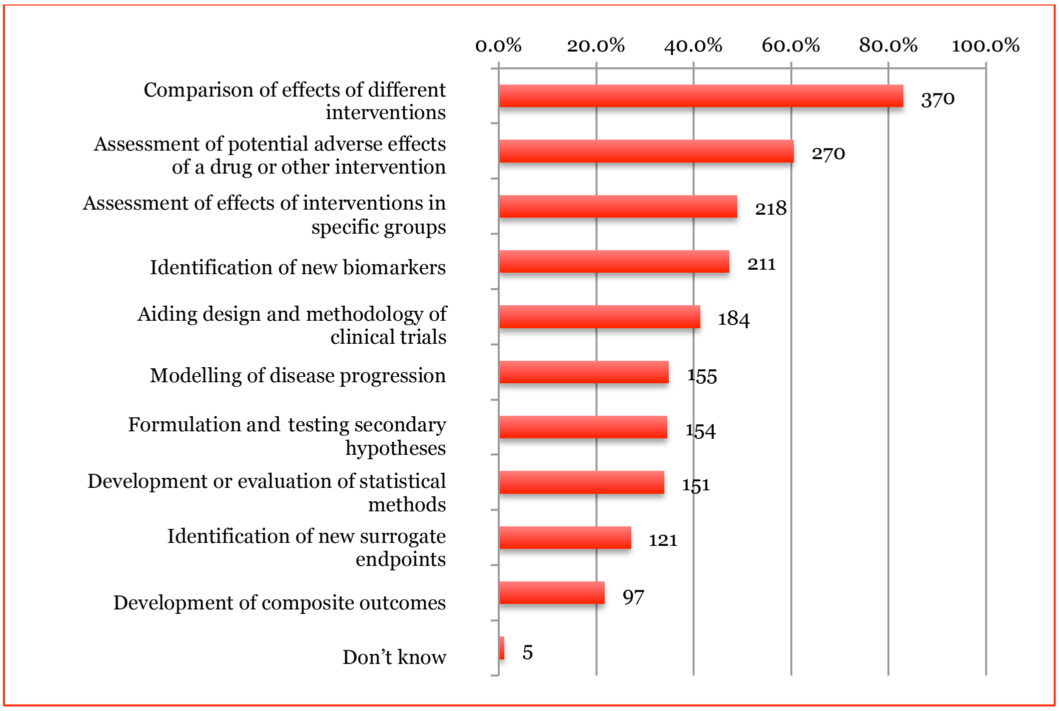
More than two thirds of projects made use of data on health outcomes, demographics, clinical laboratory test results, medical history, and adverse events; one third of projects used radiology reports and images (Figure 2c). A higher proportion of respondents from companies indicated that projects had involved data on adverse events (84%, compared to 64% of all respondents).

Survey respondents indicated that a variety of statistical methods and techniques had been used to analyse IPD (Figure 2d). Most projects involved multivariate and univariate analysis, and logistic regression. The use of less traditional techniques, such as data mining, machine learning and the use of genetic algorithms, was also noted. This may indicate the potential for the use and testing of a wide range of approaches, provided that clinical trial data, in the right format, are accessible to researchers.

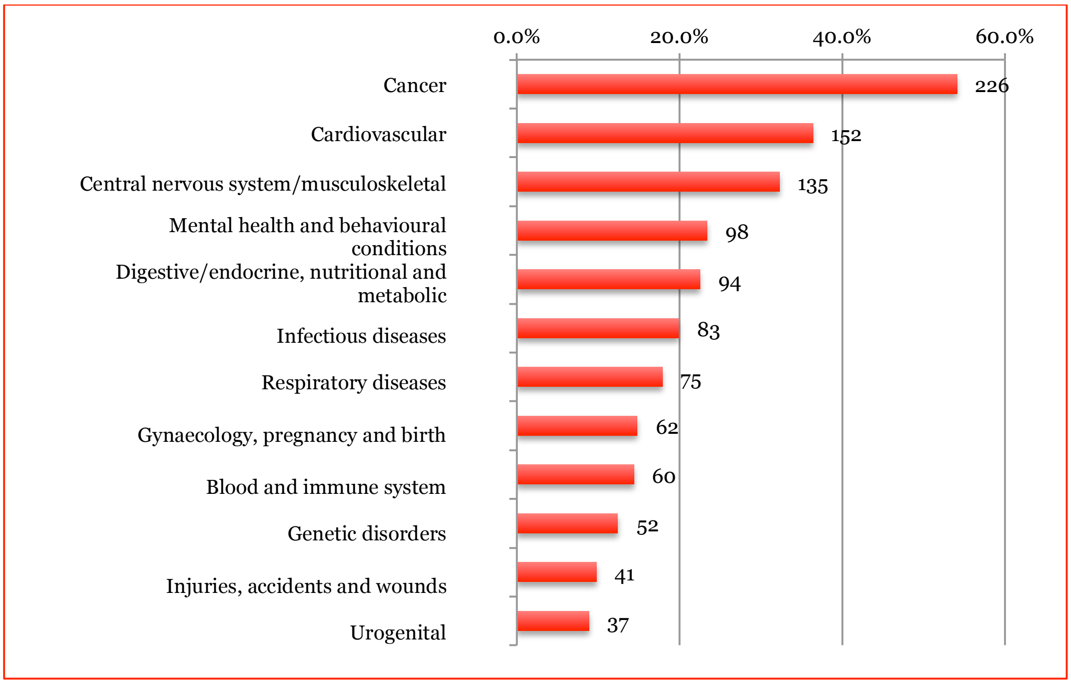
Figure 2 Current practices in research using individual participant data

(Note that multiple responses were allowed; answers do not add up to 100%. Data labels within the chart indicate the number of respondents.)

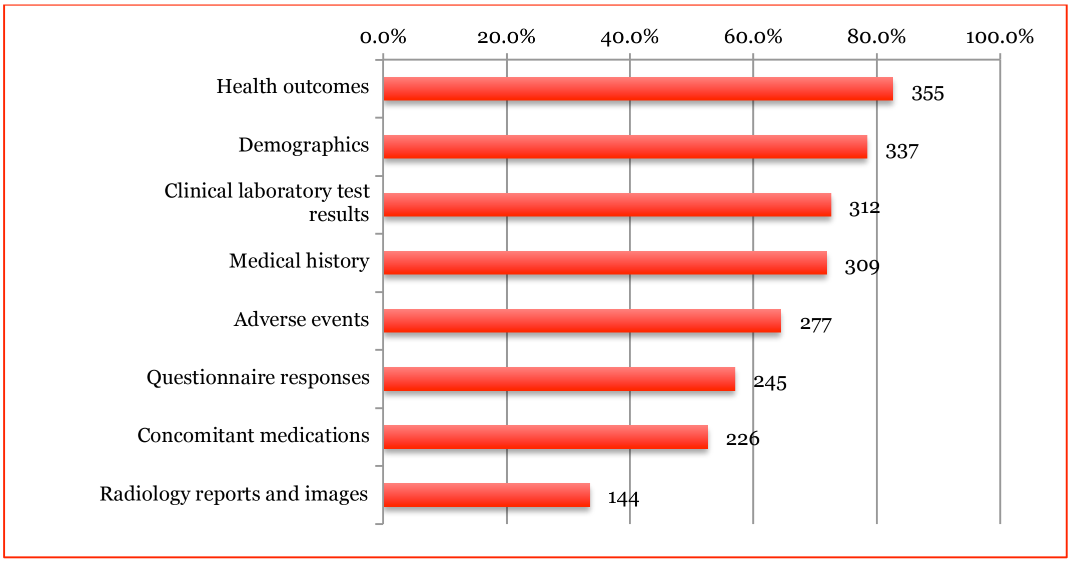
a) Please indicate the principal objectives of the research using IPD you were involved in / aware of. (n=446)



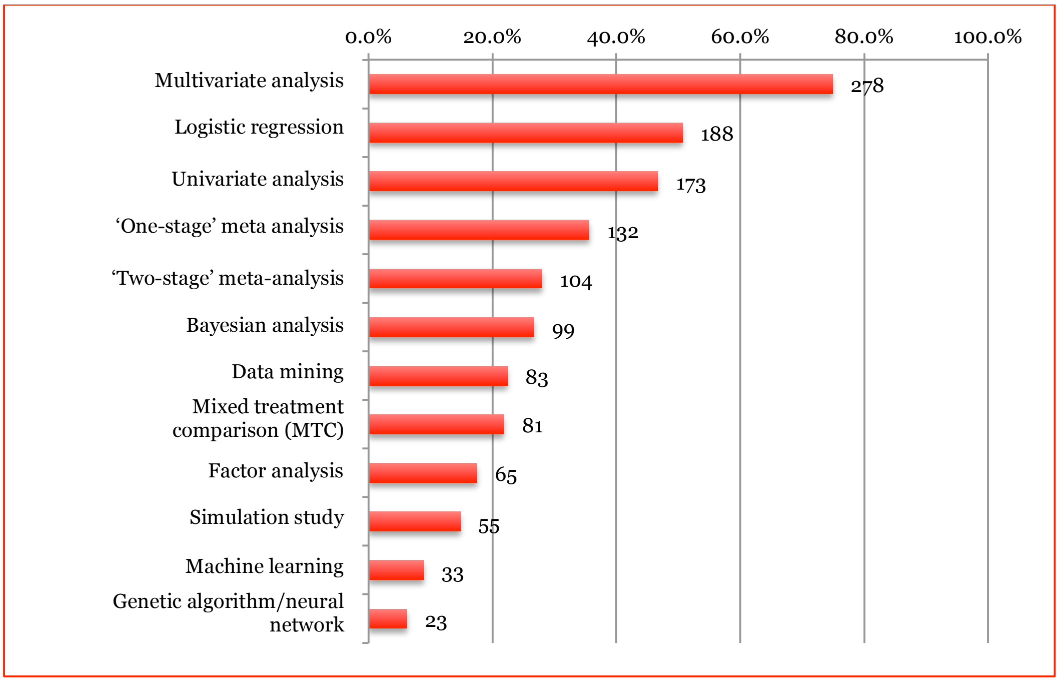
b) Disease area (n=418)



c) Type of individual participant data used (n=430)



d) Analysis method used (n=371)



3.2 Current data sharing practices

Two recent surveys, one of clinical trials authors[30] and one of members of the Cochrane Collaboration's IPD Meta-Analysis Methods Group[31], indicated that the academic research community was in principle willing to share their data through a data repository. In the first survey, 74% of the 317 respondents thought that sharing de-identified data through data repositories should be required, and 72% thought that investigators should be required to share de-identified data in response to individual requests. In all, 47% of respondents had received a request to share their clinical trial data. Of these, 77% had granted, and 38% had denied, at least one request. The second survey concluded that 83% of the 30 respondents agreed that a central IPD repository was a good idea, and 83% indicated that they would provide IPD for central storage.

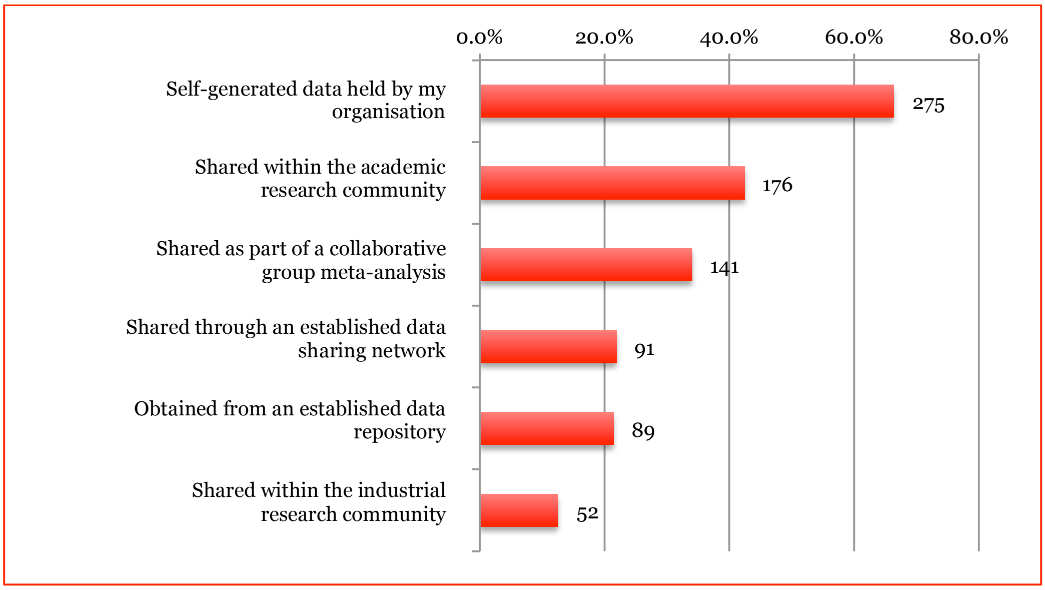
In our study, two-thirds of survey respondents indicated that IPD analysed in projects they were involved in, or were aware of, was generated and held by the organisation where they worked (Figure 3). This figure was even higher for respondents from companies, rising to 80%. For around one third of all projects, data were shared from within the academic community or as part of a collaborative group. Only 21% had obtained data through an established repository, and 13% from within the industrial research community.

[30] Rathi, VK et al (2012) Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey. BMJ 345: e7570.

[31] Tudur-Smith, C et al (2014) Sharing Individual Participant Data from Clinical Trials: An Opinion Survey Regarding the Establishment of a Central Repository. PLoS One 9: e97886.

Figure 3 Source of individual participant data

(Note that multiple responses were allowed, and hence answers do not add up to 100%; n = 415; data labels within the chart indicate the number of respondents.)



The survey asked respondents to indicate the number of data requests they made over the past year. Nearly half of the respondents indicated that they had not made any data requests, while 38% had requested data two or more times (see Table 7). The proportion of those who had not requested any data was even higher for respondents from industry (65%). This may be linked to the survey finding that the primary data source for companies is data generated within their organisation (see above); access to this internal data may not require a formal data request.

Existing data repositories reported a broad range of external requests (average over 12 months), from 200 requests to the PRO-ACT database, to around 50 requests to the C-Path Institute and the ClinicalStudyDataRequest portal, to less than 20 requests to the NIH BioLINCC and NIDDK repositories[32]. (Note that these repositories are at varying stages of establishment; hence, for some of these initiatives, the number of requests may change substantially over the coming years.) Demand may be high either because a database represents a value-added set to researchers, with a lot of the work of assembling and curating, and potentially harmonising, data already carried out, or because the data are unavailable through other channels, e.g. commercial data. Conversely, the number of requests to the repository may be lower if data can be obtained by contacting the original researcher directly, e.g. for NIH-funded academic trials.

In interviews, company representatives indicated different levels of use of in-house IPD. One company made use predominantly of summary result data, e.g., to inform the design of future trials, and did not routinely use IPD. The company was however starting to carry out some data mining projects and the interviewee expected that use of IPD would increase significantly in the future. Other interviewees from industry indicated that they made extensive use of their IPD, e.g. analysing characteristics of patients and subgroups, and planning future clinical trials. Interviewees indicated that their companies tended to share data with the academic community through direct requests for data, and within collaborative

[32] This figure includes only requests for clinical trial data, and excludes requests for data from observational studies, genetic data, and biospecimens.

groups. Some examples mentioned included the EU Innovative Medicines Initiative, TransCelerate BioPharma, and collaborative groups organised by Project Data Sphere and the C-Path Institute.

3.3 Advantages of using individual participant data

Access to IPD provides a number of potential advantages over access to summary-level data. The ability to look at individual data parameters allows researchers to analyse clinical trial data outside the original purpose of the trial. This includes "dividing up" datasets for analysis of:

- Specific subgroups of trial participants, for example those receiving a particular treatment, those with a particular genetic characteristic, or those with a particular biomarker level
- Time-sequence events, for example changes of biomarkers at different time points of the trial
- Multiple factors in different combinations, for example multiple biomarkers and genetic factors and their interaction
- Identification of rare events by pooling of data from separate clinical trials.

Other research outcomes that require the use of IPD include:

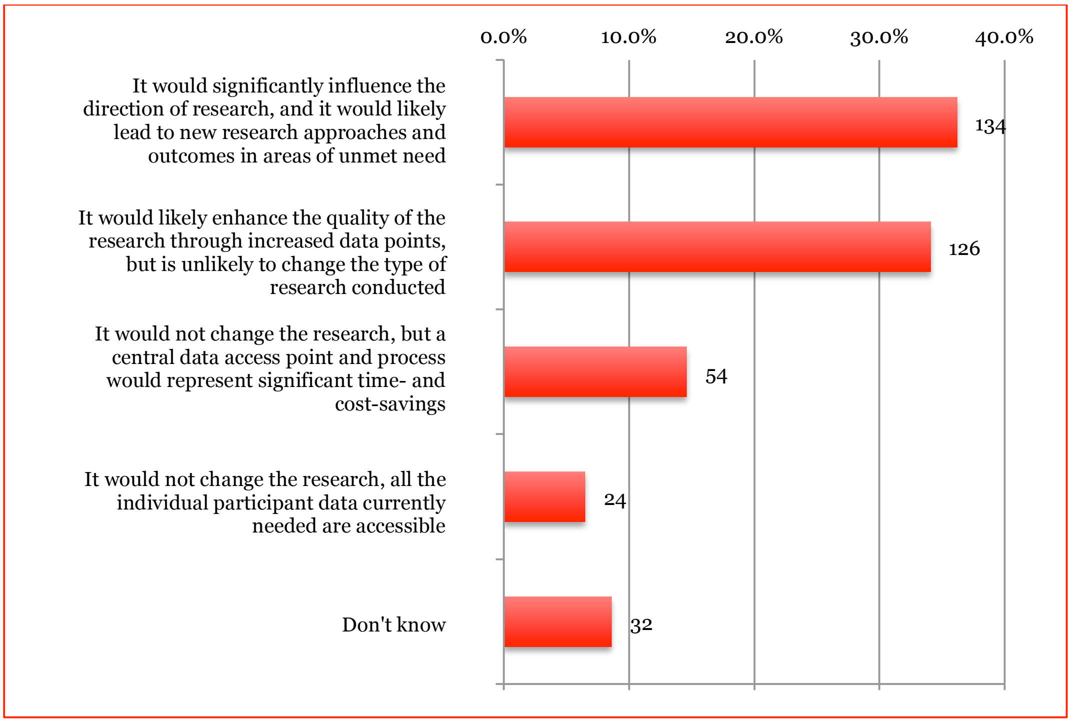
- Development of prognostic models using part of the dataset and subsequent validation of the model against the remaining data
- Enhanced understanding of treatment benefits and harms through addition of relevant data points collected after the conclusion of the clinical trial (follow-up data)
- Development and application of new analysis methods, such as those based on machine learning and neural networks
- Better identification of inconsistencies in clinical trial data collection or performance of assays in different trial centres.

A paper by Riley et al[28] presents several analyses where use of summary-level data and participant-level data yielded different results. One of the examples described relates to a study on the effect of gender on effectiveness of hypertension treatment[33]. Analysis of summary-level data from 10 trials indicated that the treatment effect was significantly lower in men than in women. However, a meta-analysis of the underlying IPD revealed that the within-trial difference in treatment effect was not clinically significant. In addition, the IPD analysis showed a non-linear effect of age on the treatment effect: up to the age of 55 years, the treatment effect increased for each year increase in age; but after 55 years there was no evidence of differential treatment effects according to age.

Our survey asked respondents to indicate how the ability to access IPD from industrial and academic trials via a central access point, such as a repository, might change their, or their organisation's, current research. Overall, the majority of respondents thought this would enhance the quality (34%), or even influence the direction of research (36%) (Figure 4). About 15% of respondents stated that a central data access point would represent significant time- and cost-savings. Respondents from companies held comparable views.

[33] Riley, RD et al (2008) Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Stat Med 27:1870-93.

Figure 4 Impact of enhanced access to individual participant data on research



Survey question: "How would the ability to access a clinical trial data repository, containing individual participant data from industrial and academic trials, change your / your organisation's current research?" n = 375; data labels within the chart indicate the number of respondents.

Survey respondents highlighted a range of research areas they expect to be supported by enhanced access to IPD. These are presented in Box 7.

Box 7 Research areas supported by enhanced access to individual participant data

- Better planning and simulation of new trials
- Consider individual patient factors to predict response to therapy
- Focus on finely phenotyped cohorts on an international scale, particularly important for the study of rare and complex disorders
- Development of models for outcome prediction to guide treatment (e.g., tumour-specific outcomes)
- Provide a robust base-line for expected outcome in rare and ultra-orphan indication
- Analyses of adverse events, disease progression and prediction
- Validation of surrogate endpoints and biomarkers
- Enhance the ability for methodological development, e.g., machine learning, data mining, causal inference techniques, missing data imputation
- More new cross-disciplinary research teams to look for new approaches

During the workshop, participants put forward more detailed examples of research questions that could be addressed by pooling large numbers of IPD datasets. Three examples are presented below.

- Understanding and dealing with patient heterogeneity

Availability of IPD at a large scale, from across many trials, is likely to progress the understanding of causes and treatments for common conditions, or symptoms, where there is significant heterogeneity across the patient population, or for those that commonly co-exist with other disease conditions. Conditions that could be addressed through access to large-scale IPD include pain, dementia, and inflammatory conditions such as rheumatoid arthritis.

For example, pain is a symptom that can be experienced as a result of many different underlying conditions, presents in many different ways, and across diverse patient populations (e.g. from children to the elderly). In addition to physiological causes, pain is complicated by psychological components. It is, therefore, difficult to address the complexities of pain in any single trial. However, given that data on pain are collected as part of many trials, bringing this vast amount of data together for secondary analysis could generate new hypotheses, such as through identification of patient subgroups, and differential effects of drug regimens or other treatments.

A second example discussed during the workshop was concerned with opportunities for progress in the area of inflammatory phenotypes, such as rheumatoid arthritis. These conditions co-present with many other diseases, are very heterogeneous across the patient population, and are complicated by the existence of acute as well as chronic phenotypes. The underlying causes are not well understood. Access to IPD in this area was expected to help to:

a) Understand basic disease etiology, such as year of onset, progression over time, changes in the condition and relapses. This could include identification of predictive markers; longitudinal data would be valuable to achieve this.

b) Define patient sub-groups, potentially by identifying genotypic markers.

These analyses can subsequently help to inform and define targeted clinical trials.

- Increased data for rare diseases

Areas that are likely to benefit from pooling of IPD are rare diseases, for which limited data are available due to the low number of trials conducted. Examples of current data sharing initiatives helping to understand characteristics of specific diseases are the PRO-ACT database for ALS, and the C-Path Institute consortium on polycystic kidney disease (see Appendix A).

- Investigating extremely rare events

The ability to analyse across a large number of datasets affords the opportunity to investigate the occurrence of extremely rare events, such as adverse events in patients who were not considered at risk initially. The primary questions addressed by the clinical trials underlying the data are not directly relevant to this secondary research question but the scale of data available from multiple clinical trials could make such an analysis possible (and could be augmented if combined with data from cohort studies and / or linkage to patient health records).

For example, for many adverse events such as stroke (or death), age is currently the only predictor. However, against all odds, strokes do occur in young people. A large IPD database could provide sufficient numbers of these rare events for further analysis. As trial datasets generally include detailed clinical and biochemical data that were collected ahead of the event (stroke), an analysis may enable identification of new biomarkers. This approach could include machine-learning techniques to predict some complex factors.

3.4 Examples of research successes using individual participant data

Re-use of IPD from clinical trials promises to enable researchers to address new questions and to increase the efficiency of the research process.

The following section presents three examples of research combining IPD from multiple sources, and reports on some of the impacts this work achieved. These, and other examples, are summarised in Table 4. All of the case studies are described in more detail in Appendix A.

Table 4 Examples of research using trial participant data from multiple sources

Research category	Research topic	Data gathered by:
Efficacy and safety of therapies	Tamoxifen in treatment of early breast cancer	EBCTCG
Modelling disease progression Identification of new biomarker candidates	Algorithms to predict ALS disease progression	PRO-ACT
Informing policy (driving standards)	Prognostic model for epileptic seizure recurrence	Individual research group
Aiding design and methodology of clinical trials	A clinical trial simulation tool for Alzheimer's Disease trials	C-Path Open Data Repository
Dose optimisation in a patient subgroup Assessment of parasite drug resistance levels	Antimalarial combination therapy in young children	WWARN
New surrogate outcome measures	Qualification of biomarker in polycystic kidney disease	C-Path consortium
Identification of an earlier clinical endpoint	Approved use of 12 week endpoint, rather than 24 week, in chronic Hepatitis C trials	FDA study
Early detection of emerging drug resistance	Molecular markers of malaria parasite resistance	WWARN (use of clinical and molecular data)
Prognostic models Common data standards Improved trial design	Dealing with heterogeneity in causes, pathophysiology, treatments and outcomes of traumatic brain injury	IMPACT, FREEBIRD
Comparison of efficacy and safety profile of different treatments Aiding design and methodology of clinical trial	Anti-epileptic drugs	Individual research group
Treatment efficacy in patient subgroups	Surgical interventions	Individual research group

3.4.1 Efficacy and safety of therapies: Tamoxifen for women with early breast cancer

The Early Breast Cancer Trialists' Collaborative Group (EBCTCG) overview is a large collaborative effort investigating the treatment of women with early (or operable) breast cancer (see Box 2). Combining data from multiple trials has allowed the group to reliably assess *moderate* treatment effects.

For example, a 1998 meta-analysis of clinical trials provided strong evidence that tamoxifen treatment substantially improved the 10-year survival of women with endocrine receptor positive (ER+) tumours, irrespective of other patient characteristics or co-treatments[34]. As data on long-term outcomes become available, the EBCTCG carries out updates of their meta-analyses. Following on from the 1998 paper, a study published in 2011[35] looked at the long-term outcomes of around 21,500 women with early-stage, ER+ breast cancer who had received more than 5 years of tamoxifen treatment (99% of all women known to have been randomly assigned into trials of about 5 years of adjuvant tamoxifen). The median follow-up for the group in this analysis was 13 years. The findings demonstrated that rather than simply delaying an inevitable event, 5 years of tamoxifen treatment prevented a high proportion of recurrences, even 10 or more years after the end of treatment, potentially curing many patients. These results allow clinicians and women to make well-informed decisions about treatment, with confidence about the likely effects of tamoxifen on breast-cancer events and overall survival.

Findings published by the EBCTCG have been embedded into clinical practice and guidelines for treatment of women with early breast cancer across the world, and have informed the design of planned clinical trials. The results have been incorporated into clinical decision and survival prediction tools, and fed into clinical treatment guidelines. In addition, the collaboration has given rise to an extremely well networked research community, facilitating information exchange between groups and avoiding potential duplication of efforts.

3.4.2 Modelling disease progression: Algorithms to predict disease progression

Amyotrophic lateral sclerosis (ALS), also referred to as Motor Neuron Disease in the UK, is a progressive neurodegenerative disease that leads to paralysis affecting one in 1000 individuals. Following the onset of symptoms, patients live for another 3-5 years on average; however, the disease progresses at markedly different rates – a long-surviving well-known patient is Professor Stephen Hawking who was diagnosed more than 50 years ago. The reasons for these differences in progression rates are currently unknown.

The PRO-ACT database is a project coordinated and implemented by the non-profit organisation Prize4Life, to accelerate the discovery of treatments and a cure for ALS. PRO-ACT houses around 8,500 ALS patient records from 17 completed Phase II/III ALS clinical trials (10 commercial and 7 academic trials).

In 2012, ahead of the launch of PRO-ACT, Prize4Life in collaboration with the DREAM Project, ran a prize competition in which participants used a subset of the PRO-ACT dataset to develop algorithms to predict the progress of ALS. The six best performing algorithms were able to identify several novel ALS predictive features, such as blood pressure, pulse, phosphorus, and creatinine levels, representing potential new lines of inquiry as ALS biomarkers. In addition, modelling suggests that use of this tool to predict disease progression could reduce the number of patients needed for a new clinical trial by 23%, representing a significant reduction in trial cost.

The algorithms are currently being tested by companies re-visiting their clinical trial data to determine if patient stratification can explain the (negative) study results, as well as by companies planning new clinical trials[36].

[34] Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. Lancet (1998) 351: 1451-67.

[35] Early Breast Cancer Trialists' Collaborative Group (EBCTCG) et al (2011) Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet 378: 771-784.

[36] The challenge is described here: Küffner, R. et al. (2014). Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression, Nature Biotechnology, doi: 10.1038/nbt.3051

3.4.3 Validation of a prognostic model for seizure recurrence following a first unprovoked seizure, and implications for driving

In the UK and other European Union countries, the majority of people who have had a first unprovoked seizure are allowed to return to driving a car following six months without a subsequent seizure. This driving guideline is in part informed by prognostic modelling of data from a randomised clinical trial, the Multicentre Study of Early Epilepsy and Single Seizures (MESS). The model included data from more than 600 participants, and estimated after 6 seizure-free months, the risk of a subsequent seizure within the next 12 months had dropped below 20%. In addition, data from MESS was used to develop a more detailed prognostic model allowing stratification of patient groups.

Before a predictive or prognostic model can be introduced into routine practice, it should be externally validated, i.e. tested for satisfactory performance in datasets that are fully independent of the development data. A subsequent study[37] to MESS used three external datasets of IPD to validate the prognostic model for seizure recurrence: two observational studies from the US and UK and a clinical trial from Italy, with a total of more than 1400 individuals. The analysis demonstrated that the prognostic model generalised relatively well, confirming its validity for predicting risk of seizure recurrence following a first seizure in people with various combinations of risk factors.

Following this external validation, the model was fitted to a pooled population comprising all three validation datasets and the development dataset. Again, the model fit well, providing support for a single, worldwide overall prognostic model for risk of second seizure following a first, which will enable driving regulations worldwide to be harmonised.

3.5 Summary

Over recent decades, the number of articles reporting individual participant data meta-analyses has risen considerably. A comparison between analyses conducted using either summary-level data or participant-level data showed that these two approaches can yield different results, and made a strong case for analysis using IPD.

The ability to look at individual data parameters allows researchers to analyse clinical trial data outside of the original purpose of the trial. This includes "dividing up" datasets for analysis of specific subgroups of trial participants (e.g., those receiving a particular treatment, those with a particular genetic characteristic, or those with a particular biomarker level), time-sequence events (e.g. changes of biomarkers at different time points of the trial), and multiple factors in different combinations (e.g., multiple biomarkers and genetic factors and their interaction). All of these objectives were found to have been addressed by past and current IPD research projects, leading to the development of prognostic models, an enhanced understanding of treatment benefits and harms, the development of new analysis methods, and identification of inconsistencies in clinical trial data collection and assays.

Enhanced access to IPD from clinical trials is expected to increase research outcomes such as those described above (e.g. in disease areas not currently addressed in this manner), and could in addition lead to novel insights that cannot be gained from summary-level data or small IPD holdings, such as an understanding of causes and treatments for common conditions or symptoms, where there is significant heterogeneity across the patient population (e.g. pain and rheumatoid arthritis), or the occurrence of extremely rare events, such as adverse events in patients who were not considered at risk initially, drawing on the large scale of high-quality data available.

The survey conducted as part of this study indicated that respondents were predominantly involved in, or aware of, projects using IPD addressing cancer. This was followed by cardiovascular disease, central nervous system or neuromuscular conditions, mental health and behavioural conditions, and digestive/endocrine, nutritional and metabolic diseases.

[37] Bonnett, LJ et al (2014) External Validation of a Prognostic Model for Seizure Recurrence Following a First Unprovoked Seizure and Implications for Driving. PLoS One 9:e99063.

The principal research objectives of these projects were comparison of effects of different interventions, and assessment of potential adverse effects of a drug or other interventions. More than two thirds of projects made use of data on health outcomes, demographics, clinical laboratory test results, medical history, and adverse events, with a higher proportion of respondents from companies indicating use of adverse events data.

Survey respondents indicated that a variety of statistical methods and techniques had been used to analyse IPD. Most projects involved multivariate and univariate analysis, and logistic regression. The use of less traditional techniques, such as data mining, machine learning and the use of genetic algorithms, was also noted.

Two-thirds of survey respondents indicated that IPD analysed in projects they were involved in, or were aware of, was generated and held by the organisation where they worked. This figure was even higher for respondents from companies, rising to 80%. Only 21% had obtained data through a repository. Nearly half of the survey respondents indicated that they had not made any data requests. This figure was even higher for respondents from industry (65%). However, the majority of survey respondents, including respondents from companies, thought the ability to access IPD from clinical trials would enhance the quality (34%), or even influence the direction of research (36%).

Existing data repositories reported a broad range of external requests ranging from 200 to less than 20 requests per year (average).

In interviews, company representatives indicated different levels of use of in-house IPD for secondary analysis, ranging from use of summary result data (rather than IPD) to extensive use of IPD, e.g. analysing characteristics of patients and subgroups, and planning future clinical trials. Companies tended to share data with the academic community through direct requests for data, and within collaborative groups.

4. Current research barriers and preferred characteristics of a broader data access model

This section provides an overview of the quantitative results of survey questions asking about current barriers to IPD research, and important characteristics of a potential future access model. A more detailed discussion incorporating qualitative information from interviews, the survey, and the literature is presented in Section 5.

4.1 Current barriers to individual participant data research

The survey asked respondents to indicate the extent to which a range of potential barriers impacted on researchers conducting projects involving IPD (Figure 5). Answers were converted to numerical values and barriers ranked by average "impact score", as detailed in the legend of Table 5, to provide an indication of differences.

On average, respondents were most concerned about access to relevant existing datasets, and incomplete knowledge of what data currently exists. This was followed by concerns over a lack of common data standards, being restricted to data analysis on the data owner's or repository server, and concerns about participant consent. Respondents were least concerned about providing competitive advantage to others. These barriers ranked in the same order when the percentages of respondents indicating 'significant impact' and 'blocks the project' were added. On average, respondents from companies were more concerned about providing competitive advantage and about identification of trial participants (0.5 and 0.3 difference in impact score, respectively). Industry respondents tended to be less concerned about sharing research proposals due to current proposal review practices (0.4 difference in impact score), limitation of analysis to data owner's / repository server, and the stringency of credential required for access to data (both 0.3 difference).

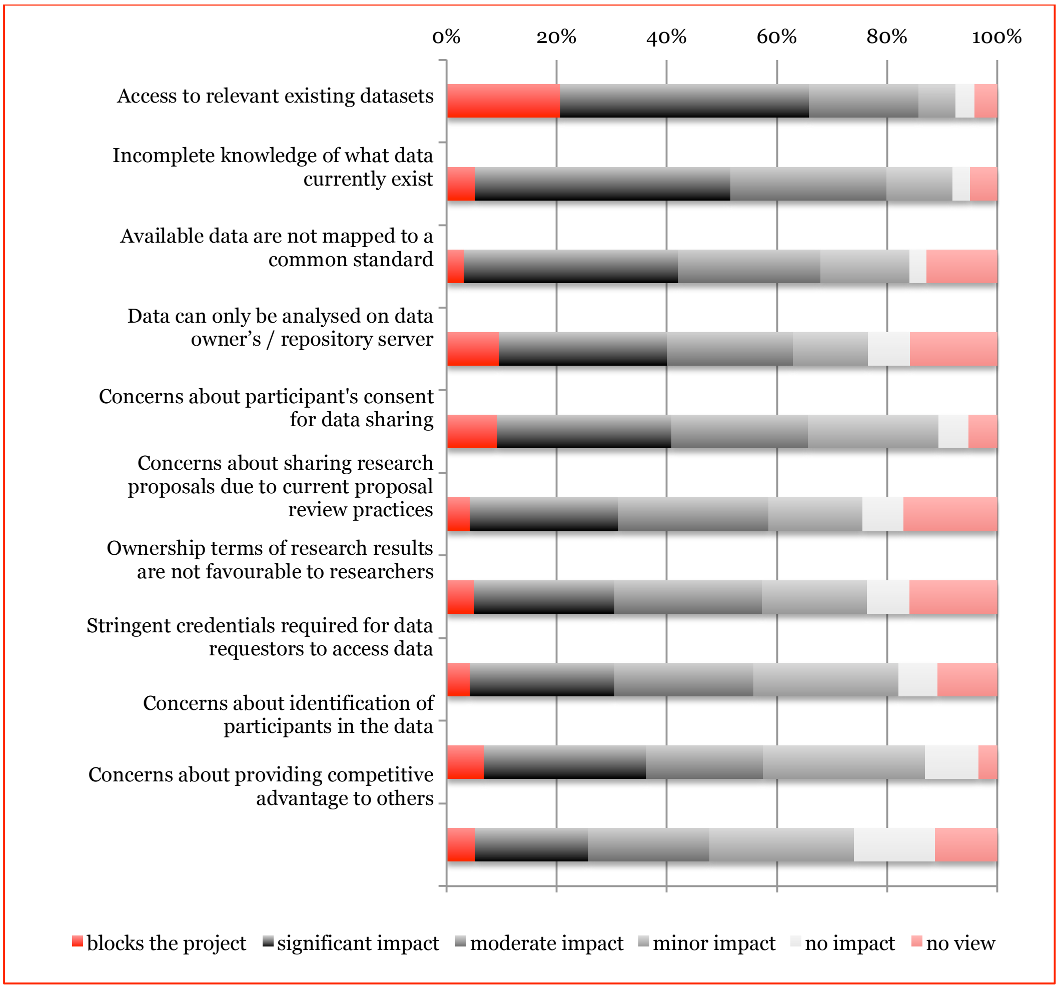
Individual barriers are discussed in more detail in Section 5.

Table 5 Current barriers to individual participant data research

Answer option	Score (all)	Score (industry)	Difference
Access to relevant existing datasets	2.8	2.6	0.2
Incomplete knowledge of what data currently exist	2.4	2.3	0.1
Available data are not mapped to a common standard (e.g. CDISC, MedDRA, SNOMED CT)	2.3	2.2	0.1
Data can only be analysed on data owner's / repository server	2.2	1.9	0.3
Concerns about participant's consent for data sharing	2.2	2.4	-0.2
Concerns about sharing research proposals due to current proposal review practices	2.0	1.6	0.4
Ownership terms of research results are not favourable to researchers	2.0	1.7	0.3
Concerns about identification of participants in the data	1.9	2.2	-0.3
Stringent credentials required for data requestors to access data	1.9	1.6	0.3
Concerns about providing competitive advantage to others	1.7	2.2	-0.5

*Survey question: "Based on your experience, please rate the extent to which the following current barriers have an impact on researchers conducting projects involving individual participant data."; n = 375 – 385; n industry = 42-45. Answers were converted into numerical values, assigning the value zero to 'no impact', one to 'minor impact', two to 'moderate impact', three to 'significant impact', and four to 'blocks project'. The values were multiplied by the number of responses, added up and divided by the total number of responses. 'No view' responses were not included. Answers are ranked by impact score.

Figure 5 Current barriers to individual participant data research



4.2 Preferred characteristics of a future data access model

The survey also explored respondents' views on characteristics of a potential future IPD access model (Figure 6). Answers were converted to numerical values and barriers ranked by average "importance score", as detailed in the legend of Table 6.

All characteristics were rated highly, with the lowest average importance score of 2.5 on a scale from 0 to 4 (i.e. between 'moderately important' and 'significantly important'), and a high score of 3.2 (i.e. between 'significantly important' and 'essential').

On average, respondents felt that it was most important a future model provide the researcher with technical information in relation to trials / data sets accessed. Respondents also rated highly that a future model include both commercial and academic trial data, that datasets could be downloaded for analysis, and that data were harmonised and presented in a single format. Respondents were least concerned about the inclusion of historical data, and the ability to analyse data with any software. A ranking of characteristics by percentages of respondents indicating 'significantly important' and 'essential' was the same as the ranking by importance score. Industry respondents assigned less importance to all characteristics, with the largest difference in importance score relating to the ability to download data for analysis (1.0 difference). Lower ranked were also the inclusion of both academic and commercial datasets, and historical data, and access of data via a central server, with the ability to use any software (0.4 difference, each).

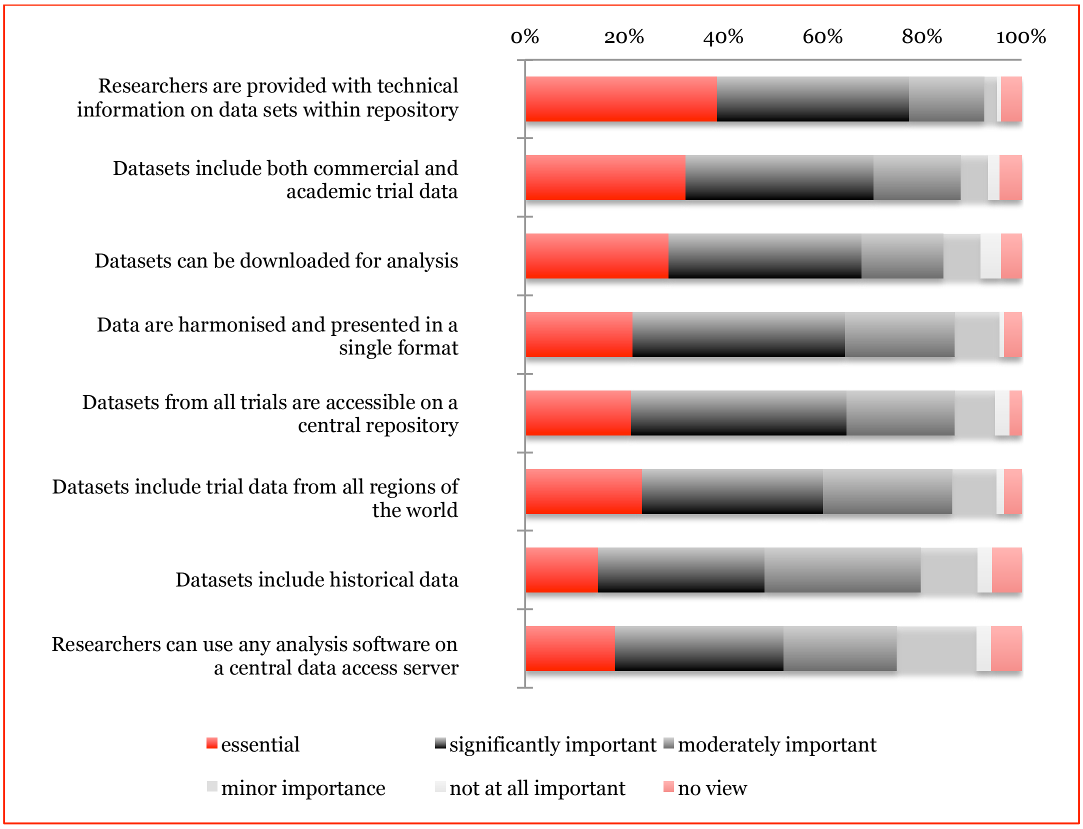
Individual characteristics are discussed in more detail in the Section 5.

Table 6 Preferred characteristics of a future data access model

Answer option	Score (all)	Score (industry)	Difference
Researchers are provided with technical information in relation to trials / data sets within the repository	3.2	2.9	0.3
Datasets include both commercial and academic trial data	3.0	2.6	0.4
Datasets can be downloaded for analysis	2.8	1.8	1.0
Data are harmonised and presented in a single format	2.8	2.6	0.2
Datasets from all trials are accessible on a central repository	2.7	2.3	0.4
Datasets include trial data from all regions of the world	2.7	2.7	0.0
Researchers can use any analysis software on a central data access server	2.5	2.1	0.4
Datasets include historical data	2.5	2.1	0.4

Survey question: "Please rate the importance of the following statements relating to the characteristics of a future data repository for the type of research you / your colleagues may want to conduct"; (n = 344 – 347). Answers were converted into numerical values, assigning the value zero to 'not at all important', one to 'minor importance', two to 'moderately important', three to 'significantly important', and four to 'essential'. The values were multiplied by the number of responses, added up and divided by the total number of responses. 'No view' responses were not included. n = 344-347 (all); n = 38-39 (industry). Answers are ranked by importance score.

Figure 6 Preferred characteristics of a future data access model



4.3 Main concerns about sharing and re-using individual participant data

When asked to describe "**the one thing that you believe would impede researchers' willingness to deposit data in a clinical trial data repository**", 40% of survey responses related to researchers' fear of losing control over how the data would be used. Within this group, 30% specifically mentioned risks to data protection and patient privacy, 16% the risk of misinterpretation or deliberate misuse of data, and 9% for each, potential lack of appropriate patient consent for secondary analysis, and fear of criticism of the original analysis. 50% of respondents from companies described "loss of control over data" as the main barrier.

The second most cited barrier was the risk that the data would be exploited without any benefit for the original researcher or study sponsor. Overall, 34% of responses listed this issue as their main concern. 63% of these responses specifically mentioned a fear of lack of recognition of the trialist's contribution, e.g. co-authorship; 27% cited potential competitive advantage to others, and 10% were concerned about loss of IP.

11% of all responses addressed the effort and cost associated with depositing data in a database.

When asked to describe "**the one thing that you believe would stop researchers from using a clinical trial data repository**", 34% of the responses were most concerned about issues with the quality of the deposited data, data format, and data structure, with 9% of all respondents citing specifically a concern about lack of data harmonisation or poor data structure. 20% felt that a heavy administrative approval process would be the main barrier, and 12% cited technical issues such as prescribed use of software or inability to download data. 11% thought researchers would be put off by the cost and effort involved in using the data, including potential access fees, and 7% listed a lack of understanding of the data, or not knowing what data were available, as the main barrier. While numbers were low, representatives from companies appeared to be concerned in particular by the level of harmonisation of datasets (18%, as compared to 9% of all responses), but were less concerned about burdensome approval processes (9%, as compared to 20% of all responses).

4.4 Potential future demand for individual participant data

Over half of our survey respondents indicated that incomplete knowledge of what data currently existed or difficulties in accessing relevant existing datasets were having a significant impact on research projects involving analysis of IPD - leading in some cases to blocking of the project as a whole (see Table 5). This might explain the enthusiastic upwards shift in anticipated data use when we asked survey respondents to estimate how many data requests they were likely to make in the coming year if IPD from commercial and academic trials were to be made available through a suitable data access model (Table 7). While 43% had indicated that they had *not* made any requests in the last year, only 14% thought they would not make any data requests over the next year should a new repository become available. Similarly, respondents from industry signalled a shift in the number of requests: 65% indicated they had not requested data over the last year, with this figure dropping to 23% for the next year should a repository become available.

As discussed in Section 3.2, the high proportion of industry respondents who had not made any requests for data over the past year may be due to the fact that companies predominantly use data generated and held by the company itself. The indicated shift in the number of data requests signals an interest in accessing data generated by other organisations more frequently. With the right data sharing model, providing benefits to all parties involved, this could support a shift in current research practices, from a siloed system to a more collaborative approach.

Table 7 Current and potential future demand for individual participant data

Estimated number of data requests per year	**0**	**1**	**2-5**	**6-10**	**10<**	**Response Count**
Last year with current access model	**43% (97)**	19% (44)	25% (57)	3% (6)	11% (24)	228
Next year with a potential new repository	14% (32)	17% (39)	**45% (102)**	10% (23)	14% (31)	227

Survey question: How many data requests do you think you would make in the next year to conduct new research projects if individual participant data from commercial and academic trials were made available through the most suitable data access model? Please also indicate the estimated number of requests you made in the past year to conduct research using IPD.

4.5 Summary

Survey respondents indicated that the most serious barrier to research projects involving IPD was current access to relevant existing datasets, and incomplete knowledge of what data exist. This was followed by concerns over a lack of common data standards, being restricted to data analysis on the data owner's or repository server, and concerns about participant consent. Respondents from companies tended to be more concerned about providing competitive advantage and about identification of trial participants than the overall survey population, and less concerned about sharing research proposals due to current proposal review practices, limitation of analysis to data owner's / repository server, and the stringency of credential required for access to data.

Referring to a potential future IPD access model, survey respondents felt that it was most important to provide researchers with technical information in relation to accessed trials / data sets. Respondents also considered it 'significantly important' that a future data holding include both commercial and academic trial data, that datasets could be downloaded for analysis, and that data were harmonised and presented in a single format. Industry respondents assigned less importance to all characteristics listed in the survey, with the largest difference in the importance attributed to the ability to download data for analysis. Lower ranked were also the inclusion of both academic and commercial datasets, and historical data, and access of data via a central server, with the ability to use any software.

Survey respondents' main concerns about sharing IPD were 'losing control' over the data (40%), and a fear that data would be exploited without benefit to the original researcher or study sponsor (34%). Views on what would stop researchers from seeking access to IPD in a repository covered a range of issues. The largest number of respondents cited concerns over the quality of deposited data (34%), and a cumbersome administrative approval process (20%).

Compared to the current situation, many more survey respondents were expecting to make requests for data should access be enhanced through a data repository. While 43% had not requested any data over the last year, only 14% thought they would not request any data from a database with a suitable access mechanism. Similarly, respondents from industry signalled a shift in the number of requests, with the proportion of those who requested data one or more times increasing from 35% last year to 77%.

5. Key considerations for a broader data access model

We investigated the implications of sharing IPD via a central access model in more detail. This section presents our findings on the potential benefits and drawbacks of a central repository or data portal, and describes the research communities' views on the scope of such a database, the process of data preparation and deposition, and the considerations around access to and use of IPD.

5.1 Transparency

One of the main arguments put forward in favour of making IPD accessible to the wider research community is the ability for independent researchers to re-analyse and confirm the findings of a study. This is seen as particularly important for those data that were used to support market authorisation for current treatments, to verify both effectiveness and safety of the intervention. In addition, as one interviewee put it: "The scrutiny of others raises the quality of clinical data and can offer new insights." The possibility that other people can access and re-analyse datasets was expected to raise the quality of research and reporting (if and where this is currently lacking). In support of this expectation, a study looking at published psychological research[38] found that researchers' willingness to share data for reanalysis was associated with the strength of the evidence (defined as the statistical evidence against the null hypothesis of no effect) and the quality of the reporting of statistical results (defined in terms of the prevalence of inconsistencies in reported statistical results).

Regarding calls for greater transparency of clinical trials that underpin commercial products, not all interviewees felt that a repository needed to make data available at the participant level: some interviewees considered Clinical Study Reports sufficient to provide full disclosure. If further analysis required IPD, this could then be requested from the original researcher. However, as described in Section 3.3, compared to IPD, the information provided in CSRs would limit what can be achieved.

In the following sections, we focus on the benefits and challenges of making existing IPD available for novel research uses.

5.2 Benefits of a central access model for individual participant data

5.2.1 Saves time and effort required for new analyses

Obtaining data at the level of the individual trial participant can be an arduous, time-consuming task. If data are not available in an accessible database, the researcher needs to identify the relevant studies and their datasets and then contact the individual researchers to request and arrange for access. Some investigators may not be responsive. For example, in one study[39], ten requests for raw data supporting publications in journals with a clear requirement for data sharing led to only one author sending an original data set. This issue is reflected in a comment made by a survey respondent: "As a relatively junior researcher (though one with reasonable technical skills) the major stumbling block to my research has been the political manoeuvring necessary to obtain data for analysis. I have only managed this by using the names of more senior researchers, and their influence, to encourage data sharing." Other investigators may be willing to share but are hampered by extensive institutional processes required prior to allowing external data access[40]. A review by Riley et

[38] Wicherts, JM et al (2011) Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PLoS One 6(11): e26828.

[39] Savage, CJ & Vickers, AJ (2009) Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PloS One 4: e7078

[40] Hrobjarsson, A (2013) Why did it take 19 months to retrieve clinical trial data from a non-profit organisation? *BMJ* 347:f6927.

al[28] gives the example of a study[41] involving IPD meta-analyses that required 2088 hours for data management, with 1000 emails exchanged between study collaborators and the data managers.

Our survey showed that respondents considered the impact of "incomplete knowledge of what data currently exists" and "access to relevant existing datasets" to be the largest barriers to current research using IPD. 53% of respondents felt that incomplete knowledge had a significant impact on research projects or completely blocked them, compared to 15% of respondents who thought it had little or no impact. 66% of respondents felt that current access to relevant data blocked or had significant impact on research, compared to 11% who felt this issue was of little or no importance. Responses from industry representatives were comparable to these overall findings.

Most respondents were also in favour of a central access point. When asked to rate the importance of being able to access datasets from all trials on a central repository, the majority of respondents indicated that this was 'significantly important' (43%), followed by 'moderately important' (22%) and 'essential' (21%). While a smaller proportion of respondents from companies (5%) felt a central access point was 'essential' (compared to 21% for all respondents), the majority still indicated it was 'significantly important' (44%). A single access point to IPD would address concerns voiced by a number of survey respondents about a "multiplicity of systems and processes", and "not knowing about [a particular repository] and what it contains".

The required costs and time for obtaining IPD will clearly vary depending on the complexity of the analysis and the number of studies involved, but it can be expected that researchers who routinely collect data for analysis across different trials will benefit substantially from central access to IPD through time (and hence cost) savings. On the other hand, the data providers will likely need to dedicate additional time and resource to making data available - and hence appropriate incentives will need to be in place (see Section 5.5).

One interviewee from industry pointed out benefits for the commercial sector. Companies tend to share data through collaborative groups (see Section 2.1.1). A central repository would decrease the time required and transaction costs associated with forming such groups, and would also provide a data infrastructure ready for use. However, this interviewee felt that overall, academia would benefit more from enhanced access via a central access point than industry. Other representatives from industry were cautious in predicting if, and to what extent, their companies would benefit. Referring to the ClinicalStudyDataRequest portal, the view was that we would have to "wait and see".

A survey respondent pointed out that having the data in a central repository would lower the bar for data access, since legal aspects around sharing (patient consent, data ownership, etc.) were already dealt with, freeing the user to focus on the research aspect.

5.2.2 Enhances data quality and value, and uncovers potential issues in data collection and interpretation

Central repositories receive data from multiple studies and multiple sources. If this is followed by curation, and potentially standardisation, by skilled database staff, the value of these datasets is significantly enhanced and oversight gained through the process allows potential issues in the analysis across datasets to be identified (and subsequently addressed). Box 8 provides two existing examples of central IPD databases illustrating these benefits.

[41] Ioannidis, JP et al (2002) Commentary: Meta-analysis of Individual Participants' Data in Genetic Epidemiology. Am J Epidemiol 156: 204-210.

Box 8 Benefits of data gathering via individual participant data repositories

Example 1: Cognitive test scores in Alzheimer's Disease

Cognitive test scores are key data points collected in trials addressing Alzheimer's Disease (AD) and other neurological conditions. These data are collected by administering a standard set of tasks and questions; individual scores assigned are subsequently combined into a single (total) score.

As part of the work of the AD consortium, the C-Path Institute gathered IPD from 27 clinical trials. In the process it became evident that slightly different variations of the test were used across trials, and that individual questions were interpreted differently. As a consequence, rather than only providing the final total score, data on each individual component of the test was captured in C-Path's database so that subsequent analysis could take account of any differences. In addition, the data are provided with extensive background information. One of the common analysis errors witnessed was around missing data. If a particular element of the cognition test was not entered, this could be down to two reasons: either the patient could not complete the task or answer the question at all, or the test had not been administered. For accurate analysis, it is important to distinguish between the two, and the background information provided by the database explicitly draws attention to this issue.

Example 2: Antimalarial reference standard and proficiency testing

The WorldWide Antimalarial Resistance Network (WWARN) has gathered data from 350 clinical trials on anti-malarial drugs, from 230 centres spread across the world. During the collection of these data, WWARN staff identified gaps and inconsistencies at various levels, from the type of data collected in trials, to large disparities in results for pharmacological assays between study centres (even if the same method was used). For example, WWARN found significant discrepancies in the drug testing results across a number of laboratories[42]. Accurate measurement of drug concentration in blood or plasma is essential to differentiate between actual drug resistance and inadequate drug exposure due to under-dosing, altered metabolism or poor absorption. In response to these findings, the WWARN team started to provide *in vitro,* pharmacology and drug quality testing laboratories with certified drug reference standards for their studies. Using reference standards ensures reliable, reproducible results that can be compared over time or location to identify trends that signal changes in malaria drug efficacy. The team also designed a proficiency testing programme to help participating pharmacology laboratories assess their ability to carry out accurate drug analysis, resolve any potential problem areas and to improve the quality of their results. A substantial improvement in antimalarial drug measurement performance was shown over the course of testing for nearly all laboratories that participated in the programme.

In addition, harmonising data for specific diseases can lead to the development of standards for future data collection across the community. This can be expected to significantly facilitate integration of future datasets into the existing body of data. Adoption of common data elements will reduce costs in the design of case report forms for new studies. For example, gathering of datasets on traumatic brain injury (TBI) as part of the IMPACT project (see Appendix A) led to the definition of common data elements, whose use is currently required in all observational studies and trials in TBI funded by the NIH as well as some calls by EU funding sources[43]. This facilitates future analyses across datasets and will assist in further optimising clinical trials, potentially reducing time and effort required to develop

[42] Lourens, C et al (2014) Benefits of a pharmacology antimalarial reference standard and proficiency testing programme provided by the Worldwide Antimalarial Resistance Network (WWARN). Antimicrob Agents Chemother. 58:3889-94. http://aac.asm.org/content/early/2014/04/22/AAC.02362-14.full.pdf+html

[43] Maas, AIR et al (2013) Advancing care for traumatic brain injury: findings from the IMPACT studies and perspectives on future research. Lancet Neurol 12: 1200-1210.

effective treatments. Furthermore, the development and uptake of core outcome sets may also make it easier to compare, contrast and combine studies in the same area[44].

Data portals, with data remaining on individual data providers' servers, would not present the benefits described above. However, a central data portal could support research communities in accessing and combining datasets relevant to their research fields to achieve these outcomes.

5.2.3 Avoids duplication of research

Many studies are undertaken that are too small to allow strong conclusions to be drawn (underpowered trials), are started but discontinued for a variety of reasons, or show a statistically insignificant, negative or null effect for the treatment being studied. While the original intent of the clinical trial could not be fulfilled, the data are nevertheless a potentially valuable source of information.

Unfortunately, results from these types of trials are frequently not published. A study[45] of 1017 trials found that 253 of these (25%) had been discontinued, and that up to 60% of these remained unpublished more than 8 years later. Other studies[46] have shown that trials with positive findings are nearly four times more likely to be published compared to trials whose findings were not statistically significant, were perceived as unimportant, or showed a negative or null direction of treatment effect.

The ability to access IPD from *all* clinical trials more easily would allow these data to be analysed in combination with other datasets, boosting the statistical power of the analyses. It would also allow separation of real effects from artefacts particular to a specific study, and validation of prognostic models against a second dataset.

Full, unbiased access to IPD and its appropriate analysis would enable clinicians to better design clinical trials and optimise research questions. As several interviewees explained, during the planning stages of a clinical trial, investigators conduct a systematic review of all evidence available, but they are often not able to assess all the parameters measured in a trial. For example, information on adverse events of a treatment may not be included in a publication or clinical study report, but the reader will not know if this is because the measure was not collected, or if it was collected but not reported. Access to the full set of IPD would clarify this and improve trial design.

5.2.4 Draws in new research communities

While the survey indicates that some new research methods are being applied to IPD, one interviewee felt that the data sharing landscape was still a long way from fundamental changes: "While current developments could be described by the term 'evolution', a 'revolution' is not imminent and will require at least another 5 years to emerge."

There was broad consensus among interviewees and workshop participants that enhanced access to IPD would allow a wider range of researchers to take advantage of the data, opening up the data sharing landscape for an influx of new expertise and creative ideas which could lead to the development of as-yet unpredictable novel methods. As one survey respondent put it: "Ease of access allows good ideas to emerge from perhaps unexpected sources."

[44] Gargon, E et al (2014) Choosing important health outcomes for comparative effectiveness research: a systematic review. PLoS ONE 9: e99111.

[45] Kasenda, B et al (2014) Prevalence, Characteristics, and Publication of Discontinued Randomized Trials *JAMA* 311: 1045-1051.

[46] Hopewell, S et al (2009) Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database of Systematic Reviews 2009, Issue 1. Art. No.: MR000006.

The potential for engaging research communities not traditionally affiliated with clinical trials, and the outcomes that can be achieved, is exemplified by the Prize4Life ALS Prediction Challenge (see Box 9).

Box 9 New user communities of individual participant data

The DREAM-Phil Bowen ALS Prediction Prize4Life Challenge

In 2012, ahead of the launch of the PRO-ACT database, the non-profit organisation Prize4Life, in collaboration with the DREAM Project, announced a prize competition in which participants used a subset of the PRO-ACT dataset to develop algorithms that predict the progress of Amyotrophic lateral sclerosis (ALS). The competition ran on the InnoCentive Prize platform and was sponsored by Nature, Popular Science, and the Economist. A prize of $25,000 was offered for the best algorithm(s) using 3 months of patient clinical data to predict the progression of a given patient's disease over the following 9 months. The prize was later increased to $50,000.

The challenge drew 1073 participants from 64 countries and resulted in the submission of 37 unique algorithms. Commissioning this piece of work through a prize competition, rather than a call for proposals, opened the field up to the "unusual" contenders: 80% of participants had no previous experience in the ALS disease field, and many did not have any form of medical research background.

Almost all of the final solutions used machine-learning techniques (random forests), which are not commonly employed in clinical research, indicating a potential missed opportunity / current disconnect between clinical statisticians and the broader data science community.

The first prize was split between two teams: a recently qualified lawyer and a mathematician, and a team from a scientific marketing company. Prize4Life is currently planning a second challenge using the PRO-ACT database.

5.3 Drawbacks of a central access model for individual participant data

5.3.1 "Disconnects" the original researcher from the data

Interviewees expressed significant concern over the risk that centralising data access and implementing an independent review mechanism for data requests would "disconnect" the original researcher from the dataset. Most interviewees felt that these types of datasets were very complex and required direct input from the original researcher, or at a minimum high-quality curation and extensive documentation, if they were to be used to conduct meaningful analyses. Beyond providing due recognition of the effort invested in primary data collection, having the data provider involved in projects re-using the data was vitally important for quality control and to avoid rogue analysis – as well as desirable for additional intellectual input.

The explanation of one survey respondent exemplifies shared concerns: "Simply centralising the data access sounds attractive but could actually harm the data quality. The danger here is that the [researcher requesting the data] no longer needs to engage with the individuals who generated the data (which you have to if you go to each individually). Hence, unless there is very high quality documentation, the [researcher] may use the wrong variables or misunderstand the quality of the data. I had full access to a (US) NIH funded dataset (after review and approval) but had real difficulties getting the [investigators] to engage. The documentation was voluminous but really hard to get through and even then I was concerned that I had chosen the wrong outcome variables. I really needed the help of a local data manager or [investigator] who understands the data better than anyone else. I finally got this. Centralisation is not an answer in itself and may disengage the data custodians." Another survey respondent put it quite bluntly: "Data sharing will work if the original generators of the data can be involved, but otherwise it will be wastefully expensive and inefficient and, most probably, produce misleading results (e.g., through failures to understand the data properly)."

The issue is addressed in different ways across database models. The data collaboration model, described as "Collaboration of trialists/trial sponsors" in Section 2 of this report, is based on continued engagement with the original researchers or data providers. Some repositories, e.g. the National Institute for Diabetes, Digestive and Kidney Disease (NIDDK) data repository have started to offer funding for workshops and networking activities between the groups that generated (or are generating) the data and potential users. In this way, the NIDDK intends to establish a cohort of junior investigators who can fully exploit the available data.

5.3.2 Represents significant cost

While a central access point would reduce the time (and costs) needed to assemble and prepare data for new research projects, the associated large-scale data preparation and deposition would represent a significant commitment of resources for the data provider. It would be necessary to balance costs and potential benefits of such an effort, and share the burden equitably between the different stakeholders.

One interviewee from industry felt that the cost burden on the pharmaceutical industry was "enormous" already, e.g. financing for the CSDR common data exchange platform. To be competitive in the global market, companies did not want to take on additional cost.

A number of interviewees drew attention to the fact that it was unclear what proportion of the datasets would actually be re-used. One interviewee explained that many datasets from his Clinical Trials Unit had never been requested by external researchers. Another pointed out that a large proportion of existing trials have never been included in reviews, and that it was therefore likely even fewer datasets would be requested at the IPD level. Hence, much effort invested in preparing many of the datasets for sharing or deposition could be "wasted".

The research impacts of a potential future access point for IPD are difficult to predict, e.g. in terms of increasing the numbers of people involved in research more broadly. At a minimum, indications are that at least some of the costs will be "recovered", e.g. through optimisation of the design of future clinical trials and reductions in waste in research, not to mention the costs saved through the development and identification of effective treatments that reduce the burden of disease. For example, a 2012 report[47] prepared for a large pharmaceutical company calculated average costs per subject for clinical trials in several European countries, ranging from €5,679 in Poland to €9,758 in the UK (£4,500 and £7,700, respectively). If research using existing IPD can help reduce participant numbers, e.g. by 50% or 25%, as suggested by the IMPACT and PRO-ACT database case studies respectively (see Appendix A), this would represent substantial time and cost savings.

5.3.3 Puts researchers in resource-limited countries at a disadvantage

Important research questions can be addressed using the large number of clinical trial datasets from around the world. However, care needs to be taken that researchers from resource-limited countries who generate these data are given the opportunity to exploit them as well.

A representative from the World-wide Antimalarial Resistance Network (WWARN) pointed out issues with sharing of data generated by institutions in low- and middle-income countries. While sharing data globally had clear value (and was at the core of WWARN's activities), he felt that there was a risk it would put researchers from these countries at a disadvantage. Hosting a central IPD repository required the presence of significant research infrastructure and resources, and as a result, they were most likely located at top institutions in high-income countries. Without an obligation to engage with the original researchers, most research publications re-using IPD would be published by groups in high-income countries, while investigators in lower-income countries, who collected data in local trials, at

[47] http://www.novartis.co.uk/downloads/europe-economics-clinical-trials-report.pdf (accessed 21 Oct 2014, € to £ exchange rate of same date used)

times over decades, were subsequently left without return. The interviewee pointed out that this would be a missed opportunity to promote equality in research. This view was mirrored by other interviewees, and a survey respondent who explained that while "access to a clinical trial data repository would definitely add value towards research progression, the capacity for high quality data management and statistics, especially in resource-limited countries like [those in] Africa, would require attention in order to avoid researchers in these settings being reduced to data collectors with no or limited ability to analyse these data". Furthermore, one interviewee pointed out that limited connectivity to the internet, or lack of a fast connection required in particular for remote analysis, may disadvantage researchers from resource-limited countries further, blocking access to data from these locations.

5.3.4 Increase in risks such as breach of patient privacy, rogue analysis, and (ab)use of data for competitive advantage

Making IPD from clinical trials available to a wider range of individuals, without the continued close involvement of the data provider, opens up the potential for data misuse – be it with malicious intent or due to lack of competence. These risks are discussed further in section 5.6.2.

5.4 Considerations around the scope of a central data access model

5.4.1 Access to academic/non-commercial trials alongside or in combination with those from commercial trials

Most of the survey respondents felt that a future data sharing model should provide access to both academic and commercial trial datasets. 71% of respondents considered this to be 'significantly important' (38%) or 'essential' (33%). Only a total of 8% gave it a 'minor' or 'no importance' rating. The 39 respondents from companies attributed slightly less importance to combining these data in a central repository, with 54% giving a 'significantly important' or 'essential' rating. 31% thought it 'moderately important', whereas a total of 13% felt it was of 'minor' or 'no importance'.

Interviewees considered the provision of academic alongside commercial trial data to be an important aspect of any future data sharing, and did not foresee any real barriers to combining the data for analysis. Most felt that both types of trial, academic and commercial, were following the same high standards in the current regulatory environment, and one interviewee pointed out that many clinicians were involved in both (at least in the UK). There were however two survey respondents who questioned if academic trial data would be of sufficient quality, and surmised that industry would not be interested in these datasets.

One interviewee from industry mentioned that while commercial trials followed the CDISC standards (as required for FDA market authorisation), many academic trials did not. This was seen as a missed opportunity to facilitate secondary analysis, but was explained by the scale of infrastructure required, which was not always available to academic researchers.

Another interviewee explained that academic trials and commercial trials tend to differ in the types of questions they address. While commercial trials were focussed on supporting drug approval by regulatory agencies, academic trials often addressed questions on the use of medicines or treatment strategies after they had been approved, to test if treatments work under real-world conditions. The interviewee felt that combining these datasets would provide important opportunities for additional research.

5.4.2 Access to trial data from all regions

Pooling data from trials conducted in different regions of the world, especially for rare conditions or for patients from underrepresented groups (e.g. children, ethnic minorities), could represent significant opportunities to gather additional data to increase the statistical power of the analyses. Additional benefits can be derived, e.g. pooling data from clinical trials on infectious diseases has allowed global monitoring of emerging drug resistance (see WWARN, Appendix A).

The survey results indicate that most of the respondents considered the inclusion of data from all regions in a future repository to be 'significantly important' (36%). 26% felt it was 'moderately important' to include these data, 24% deemed it 'essential', and 10% thought it of 'minor importance' or 'not important'. The results were broadly similar for respondents from companies.

Several interviewees recommended a targeted approach to limit costs, focussing on disease areas that would benefit most from global data.

5.4.3 Access to historical trial data

Access to historical data, data from clinical trials conducted prior to establishment of a data sharing initiative, allows not only re-examination of published results (transparency), but can be especially useful in research areas where long-term follow up data has been gathered (see the EBCTCG; Box 2), for rare diseases, or where relatively few studies have been conducted over a long time period (allowing scarce data to be pooled). However, much of this 'legacy data' exist in different formats, and are neither de-identified nor stored in a safe harbour environment. Hence, including these trials in a data sharing initiative would require significant input and "detective" work from highly skilled staff (who, as one interviewee from industry pointed out, are always in short supply).

The survey results indicate that the majority of respondents considered the inclusion of historical data in a future repository or data portal to be 'significantly' or 'moderately important' (66% and 33% respectively). 15% felt it was 'essential' to include this data, while 15% thought it of 'minor importance' or 'no importance'. Respondents from industry attributed slightly less importance to the inclusion of historical data compared to the entire population of survey respondents, with 26% indicating that it was of 'minor' or 'no importance', and only 8% indicating that this was 'essential'.

All interviewees saw value in providing access to historical data in principle, but most were concerned about the balance between cost and benefit. Some of the existing initiatives, such as the YODA project, are in the process of assessing the need for historical data and the time investment required.

A number of interviewees recommended that the decision to include historical data should be assessed on a trial-by-trial basis, and efforts prioritised by disease areas that would benefit the most. For example, historical data could provide important insights into neglected diseases, for which little research has been conducted and much of it several decades ago. Ebola was mentioned as a case-in-point: a disease area of particularly urgent need at this time. Old data gathered in the 1970s might be available and access to these data may help today's researchers to better understand the disease (e.g. its etiology) and identify potential points for treatment.

Another interviewee pointed to potential pitfalls when analysing data from trials from different time periods (e.g. different decades). In this case, researchers will need to be aware of and take into account changes in medical technology. For example, the concept of stage shift is well known in cancer research. In this field, "staging" refers to the practice of categorising a patient's cancer stage, based on the size of the primary tumour, and how far the cancer cells have spread. Stage shift occurs when new diagnostic technology improves the detection rate of cancer cells, e.g., a patient might have been categorised as 'early stage' 5 years ago because no metastasis were found, but today's technology would have found evidence of spread and classified the same patient as 'intermediate'. If the patient was enrolled in a clinical trial for a drug effective only against early stage cancer, this would have masked any effect the drug might have had, and would compromise analyses that pool data from clinical trials done across many years.

5.4.4 Inclusion of other types of data

While this study is primarily focused on tabulated clinical trial data, we also explored the need to make other types of data available alongside these datasets.

- Data from other types of clinical studies

Many of the databases profiled in Section 2 include datasets from observational or pre-clinical studies alongside data from clinical trials (for example all the NIH-funded repositories, some of the data collaborations co-ordinated by the C-Path Institute, and the IMPACT database). IPD from clinical studies other than randomised clinical trials can yield important insights. As an example, we have included a case study of a research project carried out with IPD from two observational studies and three patient registries (see Appendix A, *New surrogate outcome measure: Qualification of biomarker in polycystic kidney disease.*)

A number of interviewees and survey respondents highlighted the importance of including these types of data alongside data from clinical trials as a complementary source of information. This is in part because clinical trials are usually performed on selected, tightly focused populations, and other types of study may help to show whether their results might be applicable to more general situations.

Several workshop participants considered the current data landscape too fragmented, with insufficient integration of data from clinical trials, observational studies of interventions and cohort studies. All interviewees who commented on the usefulness of other types of clinical data were strongly in favour of allowing these to be integrated into a potential future data access initiative.

- Images

Many diseases are diagnosed or staged through imaging technology. A third of survey respondents (34%) indicated that they were involved in or aware of IPD research that had made use of radiology reports and images. Areas where this might be particularly important include cancer research, where images provide information on staging and pathology, Alzheimer's Disease and Polycystic Kidney Disease (see case studies in Appendix A), where images are used as a biomarker in clinical trials.

However, one interviewee pointed out that while images were important, the infrastructure requirements for an imaging repository were of a different magnitude to those for clinical data – and hence represented a higher cost. At the same time, while imaging data can be made available on cloud-based platforms, the cost of downloading these large data files from the cloud would represent a substantial cost to the user, which is often overlooked but might limit the ability to carry out secondary analyses.

Images can be instrumental in progressing understanding of some disease areas, as evidenced by the Alzheimer's Disease Neuroimaging Initiative (ADNI)[48]. ADNI is a longitudinal, multicentre study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease. ADNI's database includes raw, pre- and post- processed image files from MRI and PET scans obtained from more than 2000 individuals participating in the initiative[49]. Outcomes of the initiative to date include the development of methods for early detection of Alzheimer's Disease, the development of standardised methods for clinical tests (such as MRI and PET imaging), and the publication of more than 350 papers[50]. Hence, the inclusion of images in a future clinical data access initiative, or linkage of a clinical trial data repository to an image database, should be considered, at least for disease areas where images are of key importance to diagnosis.

[48] http://adni.loni.usc.edu/about/ (accessed 8 Oct 2014)

[49] https://ida.loni.usc.edu/services/Menu/IdaData.jsp?page=DATA&subPage=AVAILABLE_DATA (accessed 8 Oct 2014)

[50] http://www.adni-info.org/Scientists/ADNIOverview.aspx (accessed 8 Oct 2014)

5.5 Considerations around incentives for data providers to share data

In the absence (or even in the presence) of compulsory data sharing through regulators or funders, any future database has to take into account the motivations and concerns of data providers. Current indications are that both industry and academic research groups are at times reluctant to share their data. As presented in section 5.2.1, one study[39] found that ten requests for raw data supporting publications in journals with a clear requirement for data sharing led to only one author sending the original data set. Deposition of data in repositories has been found to be challenging even when a funder-mandated requirement was in place, as reported by NIH institute-run repositories. As a result, the NIDDK repository has recently focussed on enforcing the data sharing requirement (leading to the addition of 18 new datasets between January and August 2014).

Failure to deposit information is not limited to IPD. The FDA requires that summary results of registered trials be uploaded within one year of trial completion. Despite this, a database search of phase II, III and IV trials, which completed in 2009 and had been registered on ClinicalTrials.gov found that only 22% of trials had fulfilled the requirement in a timely fashion[51].

5.5.1 Incentives for researchers from academia and industry

Interviewees were concerned about ensuring that academic researchers had the right incentives for sharing their data. A survey respondent explained academic researchers' reluctance to sharing in these words: "Researchers are fiercely protective of their data driven by rivalry among research teams, concerns about protection of IP and the need to prove academic credentials through publication." A second survey respondent went further, and warned that mandating data deposition would turn researchers away from entering the clinical trials field as "it [was] hard enough in the present regulatory environment to do clinical trials and [we] should be careful not to add further to these disincentives. [Investigators will] lack the incentives to generate the data in the first place if it is just as easy to wait for someone else to do so." Another survey respondent concurred, saying that mandating data deposition would result in "a lack of incentives to generate the data in the first place", since it would be "just as easy to wait for someone else to do so." Another survey respondent pointed out: "There will be some analysis that could be very sensitive for the trial centre and/or the sponsor. [...] Uncontrolled or selective publication (before appropriate awareness and performance improvement) could create a defensive culture that inhibits participation in clinical research." Indeed, a recent study reported that investigators who receive industry funding withhold data because of restrictions on their control over the data[52].

At a minimum, research funders will need to consider covering the cost of data preparation for submission to a repository with a specific budget allocation in the research grant. The NIH has implemented such a system, where research proposals to some institutes of the NIH include a budget line to cover the cost of time and effort spent. To facilitate this, the NIMH has provided a simple calculator to assess the required funding[53]. One interviewee pointed out that effective tools for preparing and uploading datasets and the relevant documentation would need to be available to make this as effortless as possible. If the process were too difficult and arduous, it would add a significant burden and act as a strong deterrent.

Alternatively, a repository could take on most of the effort by accepting data in any format. Repository staff would then work with the data provider to prepare a well-curated, and potentially harmonised, dataset for deposition. Funding to cover staff salaries would need to be made available.

[51] Prayle, AP & Smyth, AR (2012) Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. BMJ 344: d7373.

[52] Rathi, VK et al (2014) Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey. Trials 15: 384.

[53] http://ndct.nimh.nih.gov/preplanning/#tab-2 (accessed 20 Oct 2014)

Many survey respondents and interviewees emphasised that data sharing does not align with the current reward system in academia, which is primarily based on publication record. This could be addressed by bolstering systems to ensure that datasets become legitimate contributions to scholarly communication. One interviewee pointed to DataCite as a possible solution. DataCite[54] are working with databases to assign persistent identifiers to datasets, which are similar to Digital Object Identifiers (DOI) for research articles, and provide information about the associated dataset and a direct link to the dataset itself. A wider use of this system would reward data producers by allowing the data to be cited, similar to research articles, making it easier for others to locate and reuse the data, and support easy tracking of the impact of the data. Further, national research assessment exercises and individual institutions could consider data sharing as a key indicator of contribution to science and, therefore, help career progression for junior faculty who do this.

Many interviewees and survey respondents mentioned that sufficient time was needed to be granted to the original researcher for analysis and exploitation of data. While views differed, generally a period of one year after completion of an academic trial was considered a reasonable timeframe, allowing the investigator to analyse the data and publish findings. Two respondents pointed out that academic trials are rarely "closed", so the dataset available for sharing would be a "snapshot", including data collected at the time of the primary publication but not beyond. This would also make it difficult to determine the required time point for sharing.

Addressing the fear of losing control over the data, a survey respondent stressed that "a critical component of data sharing must be reciprocity with respect to the research and results: if you take data out you must report back. And report back whether you did anything, nothing, or you're still working on it." Reporting requirements could be incorporated as a condition of access, not only alleviating concerns about "not knowing how data are being used", but also providing opportunities for learning and avoiding duplication of research. Researchers may also be incentivised to share if this becomes common practice for broadening collaboration networks, leading to an increase the number of co-authored publications. A summary of concerns, along with potential solutions, is presented in Table 8.

Table 8 Concerns and potential solutions to incentivise or de-risk data sharing

Concerns of data provider	**Potential solutions**
Time and cost to prepare the data If standardisation is required, necessary skills may be lacking within group	Include cost of preparing data in grant Develop and provide easy-to-use, effective tools for data preparation; potentially offer expert staff to assist Shift standardisation to repository or user; employ dedicated staff that support data standardisation Develop global standards to be applied during data collection, use potentially enforced as condition of funding
Lack of recognition of data contribution made	Institutions include data sharing as key indicator of contribution to science; e.g. DataCite citation record
Loss of control over data, fear of: • Misinterpretation of data • Being "beaten" to the publication • Loss of IP	Opportunity to review analysis before publication Reasonable timeframe for sharing Non-commercial uses only
Does not support researcher's career	DataCite metric included in research assessments of faculty Offer of collaboration (collaborative group model) – leading to co-authored publication and extension of collaboration network

[54] https://www.datacite.org/node/135 (accessed 20 Oct 2014); see also http://jlsc-pub.org/cgi/viewcontent.cgi?article=1035&context=jlsc

5.5.2 Benefits to academic researchers

While the majority of survey respondents and interviewees were concerned about the drawbacks of, and lack of incentives for, broader sharing of data via a central access model, there were also a few positive views. One interviewee pointed out that the process of cleaning and preparing the data necessary for sharing at the conclusion of a project (and/or publication) would ensure best practice in his own research group, making it possible to reproduce analyses at a later time even if the original researcher had moved on. A respondent involved in the FREEBIRD database welcomed that the effort of preparing the data had been accomplished during the final (funded) stages of the research project, rather than having to revisit the dataset at a later point after a data request was received. These views are mirrored in a recent review article[55], which reported that "promotion of well annotated datasets would occur with sharing of participant-level data. In an empirical study, investigators unwilling to share data often stated that doing so would be too much work, suggesting that researchers do not always develop a clean, well annotated dataset in a format that is easily understood by others. Along with enabling routine data sharing, proper annotation could help the researchers themselves to easily understand and use their datasets in the future." Presumably, preparing the data in such a way would also facilitate their integration with data from other studies, representing starting points for future collaborations.

5.5.3 Incentives and benefits for industry

Several interviewees from industry indicated that making IPD accessible via a central access point was the "right thing to do"; however, many were unclear about research benefits this would bring to their companies, and were unsure about if and how it may impact on research practice within their organisation. Some considered the current sharing arrangements, e.g. via collaborative consortia, sufficient to address needs, whereas a central IPD database was seen as a drain on resources. As one interviewee put it: "From [the companies'] perspective, data sharing via a repository is a huge risk without benefit."

There were a couple of positive views: One interviewee from industry welcomed the time and cost savings a central repository would bring in the form of a ready-to-use sharing platform (but saw this as a minor benefit compared to the drawbacks). Another interviewee surmised that enhanced access to data might help set in motion a shift in collaboration practices among industry. However, none saw an immediate benefit to commercial operations – one interviewee clearly stated that the benefit for academia could be expected to be much more substantial.

Box 10 presents an example of a data sharing network optimised for incentivising data providers, at the cost of ease of access for the data user.

[55] Chan, A-W at al (2014) Increasing value and reducing waste: addressing inaccessible research. The Lancet 383: 257 – 266.

Box 10 Example of a data sharing initiative optimised for incentives for data providers

Optimisation of data sharing incentives within WWARN

WWARN's model has been successful in gathering data from two thirds of academic and commercial clinical trials on anti-malarial treatments and drug resistance since 2000, from across the world, by involving the original researchers. The primary researchers retain the option to participate in the project or to limit access to the data[56]. This provides incentives for academic investigators to participate as it may benefit their careers through new international collaborations and co-authorship on resulting publications (the "currency" of academic research). The approach also alleviates contributors' concerns over potential misinterpretation of the data, and the ensuing work needed to disprove negative claims, which may be time consuming and costly. The approach has allowed WWARN to build a relationship of trust with industry, to the point that companies have started to share unpublished data (albeit with embargo until publication or submission for registration). Hence, while WWARN's model enables the data sharing principle to be fulfilled, it retains the primary investigators' involvement and trust.

5.6 Considerations around access models and potential risks

5.6.1 Access models

Current data sharing initiatives employ a wide range of access models, from completely open "downloadability" for anyone, including members of the public and non-researchers, to tightly controlled access only after approval by the data provider.

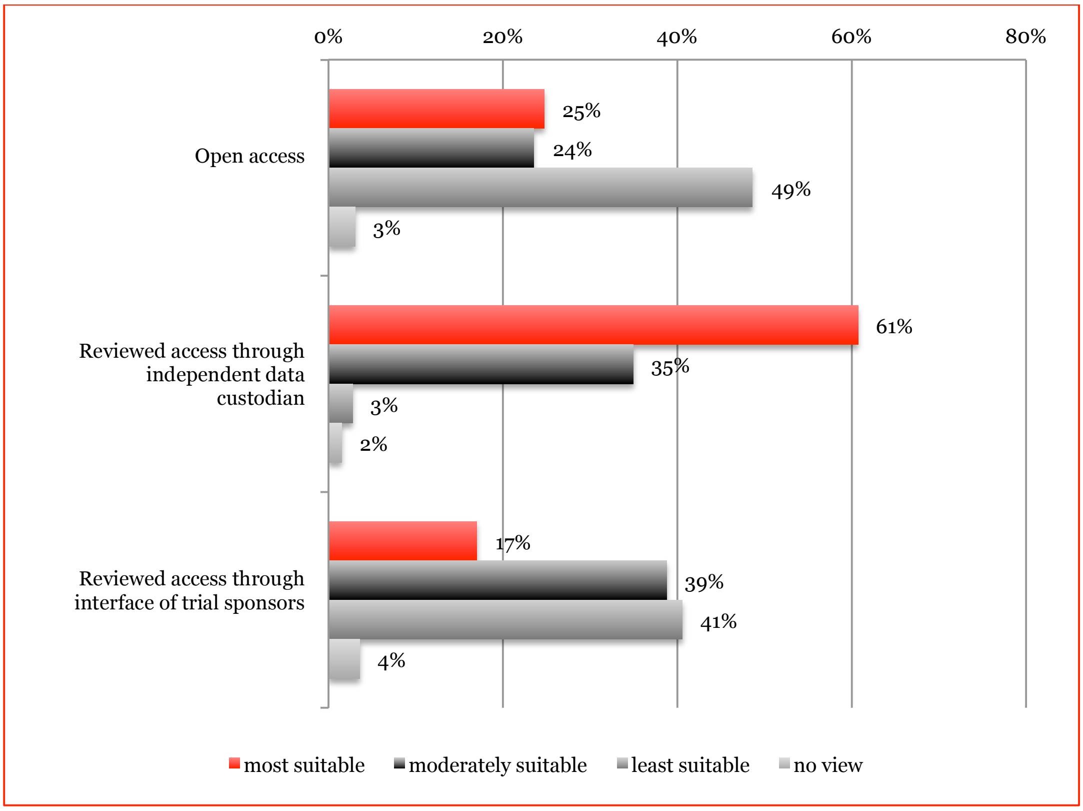
Survey respondents were asked to rank the suitability of different access and data storage models for their, or their organisation's, future research needs (Figure 7). The respondents provided the following answers:

- 78% considered reviewed access most suitable,
- 61% thought it would be most suitable if data were stored in a central repository secured by a trusted, independent data custodian, making this the preferred model among respondents,
- 25% considered open access to be most suitable, while 49% thought this access model least suitable.

Respondents from industry were more supportive of a reviewed access model (91% considered this most suitable), and less in favour of an open access model (78% considered this least suitable).

[56] Note, however, that they cannot veto the publication.

Figure 7 Preferred access and data storage models



(n range 326-335)

A range of justifications and concerns for each approach were put forward in response to open questions in the survey (see Box C 1 and Table C 6 in Appendix C).

It was notable that many views, especially around the issue of risk, were opposed to each other. For example, one survey respondent felt that “totally open access [was] likely to lead to a flood of poor quality, sensational analyses, based on little understanding of statistics and probability.” while another stated that: “There is a fear that no custody will result in bad use of the data. However I do not share this fear.”

Survey respondents were also asked about their concerns sharing research proposals due to current proposal review practices. While on average, respondents considered this to have a ‘moderate impact’, there was no broad consensus, with approximately 25% of responses for moderate, significant, and minor or no impact, each. Industry respondents tended to be less concerned about this issue.

Access to data through the interface of the trial sponsor was considered ‘most suitable’ by 17% of survey respondents, while around 40% considered this approach ‘moderately suitable’ or ‘least suitable’, each. Respondents from companies were less concerned about this access model, with only 17% considering it ‘least suitable’, 59% ‘moderately suitable’, and 23% ‘most suitable’.

Some survey respondents explained their concerns in more detail, which included:

- Concerns about trials sponsors holding back data if the proposed analysis could adversely affect them (potential data censorship or blocking of research projects),
- Increased difficulty in aggregating data if datasets are stored with each trial sponsor, i.e. in multiple locations, and
- Often restrictive nature of the trials sponsor’s data environment, forcing researchers to use analysis programs they may not be familiar with and reducing flexibility

Table 9 provides a summary of arguments made by survey respondents and interviewees in favour of or against three potential access mechanisms: open access (no review), access via review by an independent custodian, and access via review by the data owner.

Table 9 Rationale for and against different data access models

Open access, no review	**Review by independent custodian**	**Review by data owner**
Easy exploration of data possible, likely to encourage use	Delays use of data, difficult to get to point of analysis – hence researcher may not attempt	Delays use of data, difficult to get to point of analysis – hence researcher may not attempt
Ensures no bias regarding access	Carries some risk of bias	May carry higher risk of bias (e.g. conflicts of interest with data provider)
Allows access to patients and "new" researchers, including "citizen scientists" and students	Conditions may be too stringent, limits researcher / patient access	Conditions may be too stringent, limits researcher / patient access
High risk of rogue analysis due to malicious intent or incompetence	Controls risk of rogue analysis to some degree by monitoring qualifications of data requestors; ensures that the data are used to answer a scientific question, and that a properly formulated hypothesis is in place	Controls risk of rogue analysis to some degree by monitoring qualifications of data requestors; ensures that the data are used to answer a scientific question, and that a properly formulated hypothesis is in place
High risk of rogue analysis due to failure to understand data, unless direct contact with original researcher established	Risk of rogue analysis due to failure to understand data, unless direct contact with original researcher established	Controls risk of rogue analysis to higher degree as direct interaction with original researcher is required
Data may be used for research for which the appropriate patient consent is not in place	Ensures that data are used in a manner that is covered by patient consent	Ensures that data are used in a manner that is covered by patient consent

5.6.2 Risks of enhanced access to individual participant data

- Breach of patient privacy

Clinical trials capture personal data on participants, often of a sensitive nature. One interviewee pointed to the unease the public felt regarding disclosure of information from store loyalty cards, pointing out that if individuals felt very protective of data on their consumer habits, these concerns were likely to be much greater for information on their body and health.

Survey respondents expressed a spread of views on the impact of "concerns about identification of participants in the data" for current research using IPD. An equal number of respondents felt that this had a 'significant impact' (30%) or a 'minor impact' on research (30%). 7% indicated it blocked projects. Respondents from industry were on average more concerned about the impact of potential patient identification: 36% indicated this issue had a 'significant impact', and 14% felt it blocked research projects; while 26% indicated that the impact was minor.

A possible explanation for this range of opinions is that survey respondents had different access models in mind when answering this question. The risk of patient identification will depend on whether access is reviewed or not, and the amount of patient-specific data contained in the accessed dataset. As one survey respondent explained: "It is likely that an open access portal would not contain any patient identifiable data which may be essential [for the research project] and for which ethical approvals and patient consent may be available. In these cases, [access to] more detailed, identifiable data via an independent custodian would be preferable."

Most interviewees directly involved in clinical trials held the view that a repository needed to use a reviewed access model to protect patient privacy. While it could not be a full guarantee, interviewees felt that *bona fide* researchers requesting access to a database for a *bona fide* research project had little incentive to attempt to identify individuals within the datasets. In addition, some repositories require data users to sign an agreement that they immediately report accidental patient identification. If this contract were breached purposefully, the data user can be held liable.

Even with anonymised IPD (for example, in accordance with the US Health Insurance Portability and Accountability Act, HIPAA), the issue of patient identification remains a subject of concern. One workshop participant explained that it would not be difficult to identify specific individuals from the anonymised data, should the person accessing the data make a serious attempt to do so. An interviewee felt that while the risk of patient identification was manageable, it had to be addressed on a trial-by-trial basis (e.g. for rare diseases), bringing with it additional data management costs. Another stated that a "low risk of re-identification [was] acceptable".

Interviewees agreed that datasets could be further de-identified by removing additional parameters. This would be accompanied by a decrease in "research potential", limiting the research questions the data could address, or the methodology that could be applied. For example, in order to protect patient privacy, the PRO-ACT database of ALS trials removed the link between individual patients and the trials they were part of. Participants of the workshop discussed the limitations and concerns this brought for re-analyses of these datasets; many participants felt that this strategy should be avoided because limits the type of meta-analysis that could be conducted.

A number of interviewees and workshop participants recommended the development of a global privacy standard, which sets out the legal requirements for de-identification. This would facilitate the work of central repositories, by allowing them to work within a common legal framework. A recent paper commissioned by the US Institute of Medicine describes a high-level risk-based methodology that can be followed to de- identify clinical trial IPD[57], and may support the development of such a standard.

Interviewees from industry pointed out that companies were under constant scrutiny by a "litigation-happy" public. Regarding an open access database model, industry was particularly concerned about the threat of litigation following a breach of patient privacy by another member of the public. This would also have consequences for recruitment of participants for trials: if the public lost trust in the protection of personal data, it could seriously endanger future studies. Many interviewees felt however that these concerns could be successfully addressed by models of controlled access.

Several interviewees raised ethical considerations around patients' rights. One stated that patients and their families should be allowed to access data relating to their condition. A survey respondent felt that participants should be offered updates and notified of results of studies that re-used their data.

- Risk of providing competitive advantage for others

The ability to access existing data may confer a competitive advantage on users, as it is perceived to "give away" knowledge generated through the investments (and hard work) of the trial sponsor or original researcher. Academic data providers may be "scooped" to a publication, contributing to other research groups' successes without benefit to their own career.

[57] Emam, KE & Malin, B (2015) Concepts and Methods for De-identifying Clinical Trial Data. http://www.iom.edu/~/media/Files/Report%20Files/2015/SharingData/ElEmamandMalin%20Paper.pdf?la=en (accessed 16 Jan 2015)

On average, respondents ranked competitive advantage lowest among the potential barriers to current research presented in the survey. Similar numbers of respondents felt that this had a 'minor impact' (27%), a 'moderate impact' (22%) or a 'significant impact' (21%), and 15% thought it was of 'no importance'. (These views may refer to both, commercial competitive advantage, and academic competitive advantage.) Respondents employed by companies were more concerned about the impact of the threat of competitive advantage: 36% indicated this issue had a 'significant impact', and 7% felt it blocked research projects. Still, 23% indicated that the impact was (currently) minor.

Supporting this finding, the perception of this risk was also mixed among interviewees from industry. While some interviewees expressed serious concerns, others were not overly worried or advised that it was not possible to predict if, and to what extent, the recent initiatives to share commercial data would be used for competitive advantage. There were however particular concerns about protection of IP, especially for SMEs. As one interviewee put it: "Balance is needed: data sharing has to take account not only of scientific needs but also the competitiveness of those who put money into the research."

One interviewee explained that a particular concern regarding use of data for competitive advantage (in industry) was that it would be difficult to control: while patient identification could be uncovered and pursued through legal action (as a breach of the data transfer agreement), the use of data for competitive advantage may not become evident and could hence not be addressed.

- Risk of rogue analysis

While uncovering errors or issues with the original analysis would be in the public interest, rogue analysis, if through lack of understanding of the data, incompetence or malicious intent, poses a serious concern. Consequences include time and effort to re-dress the inaccurate research findings, as well as potential loss of income for the company and public trust in the research community.

A recent publication[58] describes some examples of incorrect analyses, which were published and had to be subsequently refuted. The burden of disproving an incorrect re-analysis is likely to fall on the researcher whose data were used, contributing to their reticence for open access databases.

Survey respondents and interviewees made the following suggestions to mitigate the risk of rogue analysis:

a) Reducing the risk of incorrect analysis by ensuring appropriate handling of data

The concern that a clinical trial dataset could be misunderstood, and hence analysed incorrectly, was evident from the survey. When asked to rate how important it was that a future repository provide researchers with "technical information in relation to trials / data sets within the repository", 39% of respondents chose 'essential' or 'significantly important' each. Only a total of 4% felt this was of 'little' or 'no importance'. This concern rated as the most important feature of a potential future repository.

The majority of interviewees felt that a link between users of datasets and the original researcher needed to be established to prevent publication of results based on a misunderstanding of data.

A number of the data sharing initiatives profiled in Appendix A address this issue by conducting a scientific review of the request for access to the datasets, which determines the validity of the proposed methodology, and tests whether the available data are able to support the proposed analysis (for example, the WWARN, the C-Path Institute consortia, and the IMPACT initiative). For most situations, this requires direct interaction with the data

[58] Berlin, JA et al (2014) Bumps and bridges on the road to responsible sharing of clinical trial data. Clin Trials 11: 7–12.

provider, or database staff who are familiar with the datasets and can point to any issues in the proposed research.

b) Reducing the risk of rogue analysis by controlling access

Any data access initiative will need to carefully define the term "qualified researcher" – too narrow a definition to the field may exclude experts from other research disciplines who could use the data to make an important contribution to research progress and health.

Many of the interviewees considered it important that researchers who accessed and used datasets had the right skills, and felt that researchers re-analysing clinical trial data should be subject to the same minimum requirements as researchers conducting the original research, e.g. be qualified statisticians. As one interviewee questioned: "Those directly engaged in clinical trials, gathering and analysing data, are held to a very high standard, why should this bar be "allowed to drop" for secondary use of patient data?"

Survey respondents held divergent views on the research impact of current "stringent credentials required for data requestors to access data". Similar numbers of respondents indicated that this had a 'minor impact' (27%), a 'moderate impact' (25%) or a 'significant impact' (27%). Respondents from industry tended to be even less concerned: half felt it had a 'minor' (40%) or 'no impact' (16%), while a total of 26% thought had a 'significant impact' or 'blocked' research projects.

In opposition to these views, a survey respondent advocated for a more flexible definition of credentials, explaining that: "Many physical therapists who do not have a traditional PhD (and are not MDs) are well-qualified to do database research but review boards are often biased in requiring a physician PI (or Co-PI) which is really unnecessary for someone with reasonable credentials. That is part of the value of increasing access - that individuals who are well-qualified but previously had difficulty getting funding to generate new data would be in a position to do much needed research in the field."

c) Reducing the risk of incorrect analysis by ensuring sound scientific practice

Several of the interviewed experts highlighted it was crucial for researchers requesting access to submit an analysis plan (prospectively), explaining how they will look at the existing (retrospective) data. Sound scientific practice would require a clearly defined hypothesis to be tested, formulated ahead of any secondary analysis, and was necessary to prevent the publication of spurious correlations uncovered by data dredging.

One interviewee also suggested a "stringency labelling system" for research publications, indicating the strength of evidence for the study results. This would allow readers to clearly distinguish between analyses of the primary question of a clinical trial, versus secondary (and hence weaker) analyses of data to answer questions the trials were not specifically designed to address.

Several novel approaches to limit the risk of harmful rogue analysis were put forward by interviewees and survey respondents. These included:

- A model whereby "newbies" are paired with "old hands", i.e. a requirement to show that someone with clinical trial experience is part of the team.
- A "two-tier", or graded, data system, whereby the lower, open access tier contains the datasets with any sensitive information removed (both, in terms of patient privacy and commercially-sensitive data). Access to the upper tier is regulated by a well-defined review process. As one survey respondent explained: "Open access is most appropriate with the caveats noted above [...], but if open access means that some data could never be shared because it is too sensitive, then graded access to more sensitive data depending

on the ability of the researchers to maintain privacy and confidentiality would be necessary and appropriate.” [59]

- A researcher certification process, whereby people wishing to gain access to and use clinical data have to pass a review process (to receive a “trusted” label), but are subsequently free to use all data contained within a database, without having to submit separate proposals for each project. This system would rely on a positive track record of correct use of IPD in the past, both scientifically and ethically.

5.6.3 Ambiguity around patient consent

Participants of clinical trials allow researchers to collect personal information, often of a sensitive nature, for the purpose of addressing the primary question of the clinical trial. To this end, patients sign a consent form that sets out what the collected data will be used for. Currently used forms are not standardised across trials, and often do not include specific mention of the possibility that data may be used for secondary analysis to address a question other than the primary question of the trial. This has led to different, often opposing, interpretations.

Most survey respondents (32%) indicated that “concerns about participant’s consent for data sharing” had a ‘significant impact’ on current research projects. However, a sizeable proportion (25%) felt it had a ‘moderate’ or ‘low impact’, each, while 9% thought that it ‘blocked projects’ and 6% considered it to have ‘no impact’. The responses from respondents from industry were comparable.

These differences in views were also evident from the interviews conducted: While some individuals felt that data could be used, in anonymised form, as long as the consent form did not explicitly exclude secondary use, others considered a lack of explicit consent a complete block for such research. An interviewee from industry explained that: “There is a difference in interpretation of the need to modify the consent form between companies. This ranges from an interpretation that a lack of consent means that data cannot be shared, to an interpretation that anonymisation of the data modifies the data from being a person’s to something that can be shared as appropriate.”

At the same time, some interviewees and survey respondents were concerned that the lack of clarity of current consent forms was used as an excuse by trial sponsors “to duck requests for data on the basis of this conflict”. One survey respondent felt that “this [issue] could probably be addressed by rewording of patient consent documents.” On the other hand, one survey respondent surmised that the additional consent required (for use of an individual’s data in secondary analysis) might have a negative impact on patient recruitment.

A representative from a patient group put forward the view that individuals only gave consent for use of their data to address the primary question of the trial, and hence that any secondary analysis addressing a completely different research question, unrelated to the primary one, was not covered. If the data were to be used for such research, the interviewee felt that patients needed to be re-consulted, or, at the very least, that the proposed research needed to be reviewed by an ethics committee. This was mirrored by a survey respondent’s explanation: “Patients need to be confident that the data they have given to a researcher are going to be appropriately used. [...] We need to be transparent about what we are doing with information - particularly who has access, what they can see, what they are using it for and how the results are being used. If we get it right, we can show the public that donating their data is as important as giving blood or becoming an organ donor.”

Other survey respondents who identified as patients voiced their concerns about the “sale of data”, for profit rather than the primary goal of helping patients. As one survey respondent put it: “On the whole I think patients wish their data to be used for the benefit of other

[59] For example, the FREEBIRD database has implemented a system along these lines, providing access to nearly the entire (anonymised) datasets but restricting access to the randomisation code to researchers who have submitted a proposal and had this approved.

patients, but do wish to be reassured that [...] the use is for *bona fide* medical research to improve outcomes, rather than research to improve profits." Other survey respondents explained: "As a patient and member of the public I would not want any of my data to be sold to any bidder or for unknown usage", and "We patients want our data used, but for research, not flogged off by the Government for profit".

5.7 Data format and analysis environment

5.7.1 Data curation and quality control

Datasets from clinical trials are generally large and complex. While a data warehouse, where datasets can be "dumped" in any format and without quality control, would ensure broad access to the data, a lack of data curation and accompanying documentation may increase the risk of faulty analysis and severely limit the use of data. Hence, a number of interviewees recommended that, at a minimum, data had to undergo strict quality control at the point of deposition in a central database. Many were concerned that otherwise, the data could quickly become unusable as the individuals who provided it could no longer be tracked down for clarification, or if the sponsoring company went out of business. At the same time, a number of interviewees felt that it was more important to have all the data available, in whichever shape and format it was in, to preserve datasets for later use[60]. The costs to clean up and harmonise the datasets, if these were warranted by the importance of the data, could be taken on by the research group who wished to use it for secondary analyses.

Concern over data quality was also evident through the fact that the most common response to the survey request "describe the one thing that you believe would stop researchers from using a clinical trial data repository" was a lack of data quality, structure or format (34%), and that the provision of technical information in relation to trials and data sets was considered the most important characteristic of a future IPD sharing model.

The NIDDK data repository presented one approach to data quality control: When investigators send their data to the database, NIDDK staff perform a data safety and integrity check (DSIC) which test that the data can be used to reproduce the published results of the study. If the dataset passes this test, it is added to the repository. The DSIC files are uploaded along with the data, so potential users can test and verify their understanding of the dataset.

5.7.2 Harmonisation

Harmonisation, or standardisation, of data represents a key step in secondary analysis of IPD, since it relates to preparing the data to be combined across different trials. However, it also represents a significant burden with regards to time and resources. As one interviewee put it: it is "a balancing act between the amount of time and effort spent and the usability of the database".

The survey investigated if the lack of harmonisation ("available data are not mapped to a common standard") presented a barrier to current researchers. While the largest group of respondents (39%) felt that this issue had a 'significant' effect, views on this issue diverged, with 26% indicating that it a had 'moderate impact', and 16% attributing a 'minor impact'. The breakdown for responses from industry representatives was comparable to these overall findings.

Referring to a hypothetical future repository, the survey also asked about the importance of making data available after harmonisation to a common format. 65% of respondents indicated that harmonisation was of 'significant importance' or 'essential', while a total of 10% considered it to be of 'little' or 'no importance'. The breakdown of responses from companies was similar.

[60] This requires a central repository where data are stored by an independent data custodian, rather than by the data provider / trial sponsor as in the data portal model.

One interviewee recommended that data be more extensively harmonised in priority areas for research (e.g. a specific disease), which are currently under-served by the research community. Many survey respondents and interviewees felt that such harmonised datasets would be considered very "attractive" for secondary analysis, drawing researchers into the field. Another interviewee explained that data harmonisation represented a "de-risking of the data" for researchers, who might otherwise shy away from spending extended periods of time with "unharmonised" datasets due to concerns that these may ultimately turn out not to be suitable for the intended use.

A key question is therefore at which point harmonisation should take place. For a data repository, which receives datasets for storage in a central location, four options can be considered:

1. The data are harmonised prior to submission by the original researcher.
2. The data are harmonised by data repository staff on receipt.
3. The data are harmonised by the data user after data have been received.
4. The data are harmonised by the data repository as and when the data are requested for a research project.

Option 1 would constitute a significant strain on the trial group's resources, and group members may not have the necessary time, or potentially skills, to transform the data. In addition, current incentives and reward mechanisms are not aligned with such a task. On the other hand, it was pointed out that trial investigators would eventually become unavailable (e.g. after retirement), and that it was hence imperative to translate their knowledge into an accurate map for the data to preserve the future value of the trial dataset.

Option 2 shifts this burden to repositories, which would need to employ experts who can handle the data after submission (potentially, in any format). Staff will also need to check back with the trial group to verify inconsistencies, identify errors, and document all necessary trial details. While this is the most thorough approach, and generates datasets ready for analysis, it is probably the most expensive option and may not represent value-for-money because some datasets may never be requested. This option provides good "data stewardship", as both, the original and standardised data along with supporting documentation can be stored and preserved in a central location.

Options 3 and 4 move the effort of standardisation to the *time of use*. Datasets are stored as received, and standardisation is carried out only if the dataset is requested. This could be carried out by either the end user (Option 3), or by dedicated repository staff (Option 4). This option is likely to be the most cost-efficient because data are only standardised if they are to be used. However, as the lag between data deposition and analysis may be quite large, there is a risk that the original researchers who were involved in generating the data are no longer available for questions, possibly rendering the entire dataset unusable. In addition, the large amount of work that would need to be put into understanding and using the datasets may deter potential data users.

Existing repositories make use of options 1 through 3 (see also Section 2.1). For example:

Option 1: The NDCT repository of the NIH National Institute for Mental Health (NIMH) requires the data provider (i.e., the NIMH grant holder) to harmonise the data to the NDCT's standard prior to submission. A cost model spreadsheet[61] is available to assist researchers in defining an estimated data sharing budget. Researchers are expected to include the results of this cost model in their application budget.

Option 2: The PRO-ACT database contains data harmonised across datasets from 17 clinical trials. Data were received in any format and transferred unmodified to the Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital, who cleaned,

[61] http://ndct.nimh.nih.gov/preplanning/#tab-1 (accessed September 2014)

harmonised, anonymised, and imported the data into the database. The estimated cost of soliciting, cleaning, and harmonising data for import into PRO-ACT was $500,000.

Option 3: BioLINCC is a data and biospecimen repository run by the NIH National Heart, Lung, and Blood Institute (NHLBI). Researchers funded through NHLBI grants submit their datasets to the repository in any format, and users receive the data in this format. BioLINCC does not offer custom data solutions.

By nature of their set-up, with datasets remaining with the individual data providers, data portals sit within option 3 (end user harmonises data). Research communities with an interest in preparing all available data relevant to their area of interest could take on the harmonisation task to enable broader use of the data.

5.7.3 Analysis on remote server

In order to safeguard the data from misuse, many existing data sharing initiatives do not allow users to download data to their own computers. Instead, users access and manipulate data on a remote server, with software provided in a locked data environment. Examples of this type of data access mechanism are the ClinicalStudyDataRequest data portal, and the Sylvia Lawry Centre for MS research repository.

The survey investigated if the *inability to download data* ("data can only be analysed on data owner's/repository server") presented a barrier to current researchers. The respondents had divergent views on this issue, with the largest group (31%) indicating that this inability had a 'significant impact' on current research. 23% felt it had a 'moderate impact', 14% and 8% that it had a 'minor' or 'no impact', respectively - but 10% felt it blocked current research projects. Industry respondents tended to be less concerned about this issue: 25% of respondents chose 'minor impact', and 16% felt this issue had 'no impact'.

However, referring to a hypothetical future repository, two-thirds of survey respondents felt it was 'significantly important' (39%) or 'essential' (29%) to be able to download datasets for analysis. 17% thought this ability was 'moderately important', and 8% indicated 'little' and 4% 'no importance'. As was the case for current perceived barriers, respondents from companies were less concerned about this aspect of a future repository: only one third thought it was of 'significant importance' (20%) or 'essential' (13%), whereas 25% felt this was of 'no importance', and 15% attributed 'little importance'. This characteristic showed the highest divergence of opinion between the overall population of respondents and respondents from companies only.

The survey also demonstrated that the majority of respondents considered the ability to use "any analysis software" to be important, should a future repository require analysis of data within a locked environment. Half of respondents indicated this ability was 'significantly important' (34%) or 'essential' (18%), 23% considered it 'moderately important', and 16% 'of little importance'. A smaller proportion of respondents from companies were concerned about the ability to use analysis software of the researcher's choice: 33% indicated that it was of 'significant importance', 10% that it was 'essential', whereas 23% and 10% felt this was of 'little' or 'no importance', respectively. Views may differ depending on the background of the respondents. For example, while medical statisticians participating in the ALS Prize Challenge (see Box 9) used predominantly SAS or R, participating computer scientists employed other programmes, which required the ability to download the datasets as a text file so they could be imported.

Many existing initiatives provide specific analysis software and tools for analysing data accessed via their secure data platform. This may limit the types of analysis, and the number of researchers who may want to access and use the data. As one interviewee stated, he would not use data without having full control over the computational environment and his choice of tools to analyse it, noting that the time investment to implement even simple tests with provided software would be too great. There were also concerns about the later reproducibility of an analyses conducted in a locked environment. For example, if there was a need to revisit the work months after publication, the data environment might have been updated (e.g. operating system and analysis software), making it impossible to reproduce the analysis.

An important consideration is also the type of data handling required for specific analysis methods. For example, one-stage meta-analysis (used in more than a third of the projects that survey respondents reported on, Figure 2) requires the ability to combine IPD from multiple trials, rather than combining the analysis results of individual trials. Unless all datasets reside in the same repository or can be transferred to the same remote environment, this is difficult to achieve and in many cases may eliminate the ability to conduct this type of analysis.

It should be noted that some of the existing data sharing initiatives have made, or are considering, exceptions to the data lock where a strong scientific case can be made. For instance, trial sponsors signed up to the CSDR data portal have in some instances approved transfer of data to users' machines and the YODA project is considering this.

5.8 Summary

Survey respondents', interviewees', and workshop participants' views were largely positive regarding a central access model for IPD and the research opportunities afforded through such an initiative. However, it was evident that there were substantial concerns about the practicalities and potential risks of sharing initiatives.

The **benefits of a central access model for IPD** were that it could:

- Increase transparency
- Save time and effort required for new analyses, by providing a single / a small number of access points to data, with legal aspects already taken care of
- Enhance data quality and value, and uncover potential issues in data collection and interpretation
- Increase data discoverability
- Avoid duplication of research
- Draw in new research communities, by lowering the effort required for researchers external to the core clinical trials community to access data.

The **drawbacks regarding of a central IPD access model** were that it would:

- Disconnect the original researcher from the dataset, and hence increase the potential risk of incorrect analysis.
- Represent significant cost to data providers and repositories, with the possibility that many datasets will never be re-used.
- Put researchers in resource-limited countries at a disadvantage, by placing data at the hands of researchers in highly-funded research institutions without research benefit for those who collected the data.

In addition, survey respondents and interviewees highlighted the misalignment between the benefits of data sharing and rewards for the original researchers / trials sponsors. This ranges from the cost and effort of preparing datasets for sharing, the lack of recognition of the data contribution made, and a loss of control over the dataset leading to potentially increased risks such as misuse of data, giving competitive advantage to other researchers or companies, and loss of IP.

Regarding the **scope of a central IPD access model** suitable to maximise research benefits, survey respondents and interviewees broadly agreed that:

- Data from academic and non-commercial trials should be provided alongside commercial trial data, as these often addressed complementary research questions. Respondents did not foresee any real barriers to combining the data, but more effort may be needed to harmonise data from academic trials (as not all use the CDISC standards).

- Access to trials from all regions was desirable but not practically achievable. Data from disease areas that would especially benefit from access to global data should be prioritised, rather than trying to gather all data from the outset.
- Access to historical data was desirable, especially in research areas where long-term follow-up data exist, but not practically achievable across all disease areas given the cost implications. Data from disease areas that would especially benefit from historical data should be prioritised. Researchers conducting secondary analysis needed to be made aware of potential pitfalls when analysing these data, such as differences in data collection due to changes in medical technology.
- Other types of data should be combined with, or at least linked to the numerical data from clinical trials. This includes data from observational studies, which provide important long-term datasets complementary to the shorter-term clinical trial data, and images, which are essential in some disease areas.

Regarding **access mechanisms for a future IPD repository**, to enable the broadest possible use of the data while keeping risks at an acceptable level, most survey respondents and interviewees felt that reviewed access to datasets held by a trusted custodian was most suitable. However, while half of survey respondents (49%) considered the open access model least suitable, a substantial proportion (25%) chose this as the most suitable model, indicating that the scientific community does not currently agree on this point.

Concerns about data continuing to be held by the original research or trial sponsor included potential data censorship, increased difficulty in aggregating data if datasets are stored in multiple locations, and the often restrictive nature of commercial trials sponsors' data environments.

Potential risks of enhanced access to IPD included:

- Breach of patient privacy. This could be limited by removing additional data parameters from the trial dataset, and / or by limiting access to *bona fide* researchers.
- Providing competitive advantage for others. For academic research groups, this could be limited by allowing sufficient time for the original researcher to exploit the data before external access is granted, or requiring the original researcher to be informed of, or potentially involved in, any subsequent projects. This risk is difficult to address in a commercial setting.
- Rogue analysis, either through lack of knowledge or malicious intent. Suggestions for how this risk could be limited included:
  - extensive data curation of deposited data, and availability of detailed technical information alongside the dataset(s)
  - limiting access to research teams with the right skills and credentials
  - requiring submission of a clearly outlined research proposal along with the request for access
  - and/or requiring the original researcher to be informed of, and potentially involved in, any subsequent projects.

In addition, it was evident that the lack of clarity on patient consent forms concerning secondary use of data needs to be addressed, with some interviewees calling for the development of a standard question addressing this issue to be included on all forms going forward.

Regarding **data format and the analysis environment**, survey respondents and interviewees broadly agreed that data needed to be curated to a very high standard. Respondents also thought it important that researchers could download data to their servers, or at least use any analysis software they wanted on the remote desktop provided by the repository. Respondents from industry assigned lower importance to this. Harmonisation of datasets was desirable, but not practically achievable on a large-scale. A number of views were put forward as to when data should be harmonised (at the point of deposition or when data is requested) and by whom (data provider or data user), to optimise capturing the full value of the data while keeping this effort to a reasonable level. Existing databases use a range of models, which may account for different levels of data requests made by the research community. Some interviewees suggested that data could be harmonised in priority research areas, lowering the bar for data use, to accelerate research and draw in new research communities.

6. Conclusions and recommendations

Clinical trials are conducted to test the efficacy and safety of medical interventions, producing large volumes of well-characterised individual participant data (IPD). Following a number of high-profile cases where companies stood accused of failing to provide access to safety data, there is currently a general push for clinical trial data to be made publicly available, at least at the summary-level, to address issues of transparency. As a testimony to these developments, the European Medicines Agency recently announced the adoption of a policy to publish parts of clinical study reports on all authorised medicines, and re-confirmed plans to make available individual patient data in the future[62].

In addition to their primary purpose of testing the efficacy and safety of a medical intervention, clinical trial datasets represent a research asset in their own right. Data from individual trial participants can be used to address a range of important additional (secondary) research questions, such as comparing the effectiveness of different interventions, identifying subgroups of patients and new biomarkers, and aiding the design and methodology of future clinical trials. These additional research opportunities, together with mounting legislative and public pressure for greater transparency, have initiated a trend among some research organisations, publishers and trial sponsors to make IPD from clinical trials more broadly accessible via data repositories.

While enhanced access to, and pooling of, IPD may open up exciting new avenues for research, there are legitimate questions around the need for changing current data sharing practice. Any future repository will also need to consider questions around the scope, technical parameters, and suitable access mechanisms.

This study examined the history and set-up of existing data sharing initiatives, their current research use, and impacts achieved. It also gathered the views of members of the research community regarding the need for broader access to IPD, via a central access point, and the characteristics a potential future repository or data portal should have in order to allow researchers from the academic, non-profit, and commercial sectors to contribute data and share research benefits, while protecting patient privacy and respecting the wishes of trial participants regarding re-use of their data.

During the publication process of the present study, the US Institute of Medicine (IoM) published their independent report “Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk” in January 2015, which sets out guiding principles and a practical framework for the responsible sharing of clinical trial data. The report concludes that sharing data is in the public interest, but that a multi-stakeholder effort was needed to develop a culture, infrastructure, and policies to foster responsible sharing. The recommendations formulated in this study and those in the IoM report are broadly in line and complementary to each other. The authors of the present study hope that information presented in these studies will contribute to further the thinking of international stakeholders around the issues at hand.

[62] http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2014/10/news_detail_002181.jsp&mid=WC0b01ac058004d5c1 . Announced on 15 October 2014, policy to take effect from 1 Jan 2015 for products for which the application was submitted in 2014 and later. (accessed 11 Dec 2014)

In this section, we provide a synopsis of findings on different aspects of sharing IPD from clinical trials, as set out in the objectives of this study, and present recommendations for future sharing based on the evidence we gathered.

6.1 Current models and support for novel research

The past 30 years has seen a substantial increase in studies re-analysing existing IPD from clinical trials, often involving datasets from multiple trials (e.g. see [29,29]). For many of these studies, researchers had to spend substantial amounts of time to discover the existence of, and gain access to, existing datasets.

In recent years, an increasing number of IPD repositories and data portals have become available. We identified a cross-section of 18 clinical trial sharing initiatives, in the UK and internationally (Section 2.1). We found several distinct types of databases and repositories, but in simple terms they can be thought of along two dimensions – access and focus – with data holdings exhibiting varying degrees of openness (e.g. data only available to members of the research consortium that are depositing individual clinical trial datasets; data openly available to any third party) and focus (e.g. datasets held and harmonised in order to support research on a single disease; datasets deposited to allow access for any research question, without a specific focus). There are various other important characteristics, such as the size of the database, whether data from the treatment arm of trials are included, and whether it contains data from academic / non-commercial trials, commercial trials, or both.

On the basis of these clinical trial datasets, substantial novel research is being carried out that would not be possible by access to summary-level data only. We identified several notable research-enabled developments, which are having a positive impact on medical practice and patient health outcomes, ranging from informing clinical guidelines on the efficacy and safety of treatments in patient subgroups, through identifying early (disease-specific) clinical endpoints for future clinical trials, to aiding the design of clinical trials, and the development of a clinical trial simulation tool.

In gathering information about these data sharing initiatives, and looking at the evolution over time (i.e. date of formation of repositories), we expect to see a continued growth in the number of such repositories and data portals in the coming years in the absence of wider agreements in place. It is also clear that most of these 18 repositories and data portals are steadily expanding their individual data holdings as new relevant datasets become available. We found that, at present, the majority of existing data sharing initiatives do not coordinate efforts, likely owing to the different access models, disease areas covered, and funding arrangements. Increased information exchange could strongly support the development of best practice, and start to explore options of how to enable linking of existing databases, and possibly new repositories, in the future. We recommend that such information exchange across existing data sharing initiatives be enhanced.

Ultimately, these existing repositories and data portals are capturing data from just a small fraction of all clinical trials globally. There is an opportunity, in principle, for independent and trusted global actors to accelerate the process by supporting the creation of larger holdings of data from both non-commercial and commercial trials, and creating the infrastructure to facilitate the flexible linking of the many existing data repositories. In order to enable optimal use of data to answer new research questions, we consider it essential that further fragmentation, e.g. through launches of new, unlinked initiatives, is avoided. Any new effort hence needs to aim to establish a central access point that incorporates both non-commercial and commercial data – if not from the outset, then with the clear goal of enabling these datasets to be combined at a future point.

Recommendations:

- Promote establishment of larger data holdings, with the clear aim of incorporating IPD from both, commercial and non-commercial clinical trials
- Initiate enhanced information exchange between existing data sharing initiatives and support linking of existing repositories and data portals

6.2 Future access to individual participant data from clinical trials

This section presents a synopsis of findings on the different elements of sharing IPD from clinical trials, and presents potential future avenues should a new central IPD repository be established, as favoured by survey respondents, taking account of both, the expected benefits and challenges.

On average, experts consulted saw benefits to all characteristics of a potential future IPD repository we investigated[63], including respondents from companies. The preferred characteristics of a future sharing model for IPD from clinical trials, as informed by the survey, are presented in Box 11 .

Box 11 Preferred characteristics of a future sharing model for individual participant data from clinical trials, based on survey responses

- One central repository
- Repository includes data from academic / non-commercial and commercial trials
- Data are held by a trusted third party
- Datasets are curated to a high standard
- Access is reviewed by an independent review board
- Data can be downloaded to user's server
- Datasets are harmonised
- Historical data are included
- Data from all regions are incorporated

While Box 11 outlines the "ideal" repository scenario, this may not be practically achievable, given barriers such as costs and differences in national legal frameworks.

In the following sections, we synthesise our main findings on the need and demand for enhanced IPD sharing and the key mechanisms and practicalities that would need to underlie a broader data access model, and present our recommendations based on the evidence gathered as part of this study.

6.2.1 Is there demand for individual participant data, and a need to change current data sharing practices?

It was clear from the information we gathered that current issues with access to data and simply "not knowing what data exist" are having a significant impact on research, with much of the data difficult to access or altogether inaccessible. The majority of experts we consulted expected that access to IPD via a central repository or data portal would improve on the current situation, enhancing the quality of research by providing not only more data points but even enabling new research directions, with a subsequent increase in demand for IPD datasets.

We found that many of the existing data sharing initiatives were relatively unknown outside the immediate relevant research community. To enhance discoverability and information exchange, we recommend a central information website (e.g. a "IPD research hub") be established, or existing registries adapted, offering weblinks to existing repositories and data portals. In addition, appropriate and targeted funding streams need to be available in order for researchers to share existing IPD from clinical trials as well as re-use IPD for secondary analysis.

[63] Average scores in the survey above 2 ("moderately important"), on a scale from 0 ("not at all important") to 4 ("essential")

Enhanced discoverability and access to funding will likely boost demand for research using IPD, and contribute to solving new research questions.

Recommendations:

- Establish a central information website, or consider adapting current clinical trial registries, with profiles and links to existing repositories and data portals
- Ensure that funding streams for sharing and secondary analysis of existing clinical trial data are available to facilitate generation of new knowledge
- Monitor actual demand and research outcomes following promotion of available repositories and data portals

6.2.2 Is there a need for a central repository or data portal?

Our study found that a central repository or data portal, providing a single point of access, was expected to greatly enhance access to IPD, and address barriers presented by the currently fragmented database landscape, which requires researchers to navigate a multiplicity of access points while shouldering the associated administrative burden. In particular, a central IPD repository would provide the necessary infrastructure for organisations not able to physically host the data and facilitate researchers to form worldwide collaborative working groups with all the associated benefits, e.g., harmonisation of data and development of common data standards.

While a single central repository is the ideal case solution, we recognise that this may be difficult to achieve in practice, especially across different jurisdictions. A potential solution would be to set up a small number of repositories or data portals in different regions of the world, hosted on compatible platforms that allow data linking when necessary. This would require coordination of potential funders and data providers from relevant countries. It may be most cost effective to extend an existing data platform to encompass, or link to, new data holdings, rather than setting up a new repository.

It should be stressed that all stakeholders, i.e. industry, funders of non-commercial trials, and disease-specific groups need to be strongly encouraged to discuss and agree on the smallest number of access points possible, rather than further increasing the number of individual initiatives. Such discussions should also explore sustainable business models for large-scale IPD repositories.

Recommendations:

- Establish a central repository or data portal to facilitate access to IPD from clinical trial data. Such an effort may need to take the form of a small number of regional repositories on compatible data platforms
- Establish a global discussion forum of potential funders of IPD sharing initiatives to develop global support and a joined-up approach leading to the implementation of globally “linkable” IPD repositories and data portals

6.3 Scope of a central data access model

6.3.1 Is there benefit to combining data from academic / non-commercial and commercial trials?

Experts consulted were broadly in favour of combining datasets from academic / non-commercial and commercial trials in order to maximise data coverage and research opportunities. These datasets are expected to conform to similarly high quality standards even if the data formats used often differ.

We recommend the implementation of a flexible central data access model open to non-commercial and commercial datasets, drawing on best practice of existing data sharing initiatives. This could potentially involve extending a suitable existing data platform in which both physical data repository and portal functions co-exist to allow for differences in data providers' approaches to data sharing. We believe it is essential to establish a close collaboration between the different data providers with the common goal that combined access be provided at a later point and meet (as far as possible) the "ideal scenario" set out in Box 11. We note that as industry is gathering experiences from recently implemented repositories and portals, the current review and data access mechanisms may change, potentially facilitating a future merging with non-commercial data initiatives.

Recommendations:

- Evaluate current data sharing platforms against desired characteristics (Box 11), and for suitability for expansion, to develop and implement a data sharing platform drawing on best practice from existing repositories
- In case different data sharing requirements prevent some data providers from joining the "new" repository or data portal from the outset, continue dialogue to allow data linkage or merging of data holdings at a future point

6.3.2 Is there benefit to including historical data, and data from all regions of the world?

Experts consulted broadly agreed that there would be benefit in including historical data and data from all regions of the world in a central repository or data portal.

While "complete" data would be ideal, the costs associated with preparing all historical data may be too high and potentially outweigh the research benefits that can be derived. Rather than opting for a blanket inclusion of all historical datasets, specific priority research areas could be targeted, with the ensuing costs covered by relevant funding bodies, such as disease-specific charities. We recommend including this "top-up" option for a future repository, with clear processes for how individual funders can contribute. Targeted efforts may also cover the cost of data harmonisation (see Section 6.4.3).

Inclusion of data from all regions is likely to be important for some research areas, but may be hampered by differences in legal frameworks across borders. We recommend the initial establishment of a repository in one region, such as the EU and North America, as a testing ground to develop a robust, cost-efficient solution, acceptable to data providers, data users, and patients alike. This could then function as a model for efforts in other regions of the world, to be rolled out at a later point.

Recommendations:

- Rather than aiming to incorporate all historical data from the outset, adopt a case-by-case approach, e.g. only in research priority areas, or as mandated by individual funders
- Establish clear processes for deposition of historical data in priority research areas
- Implement a pilot repository in one or a small number of regions to develop a robust, cost-efficient solution that could function as a model for future efforts in other regions

6.3.3 Should other types of data be included in a central individual participant data repository?

Experts consulted generally agreed that a repository for IPD from clinical trials should include, or at least have the ability to link to repositories of, other types of data such as observational study data, genomic data, medical images, and public records and registries.

A future IPD repository is unlikely to integrate every type of data from the outset; however, options and challenges for future linkage across databases should be considered as the repository develops. In addition, it will be helpful to learn from other data sharing initiatives from further afield, such as public health or e-government initiatives, as many of the issues around data sharing will cut across.

Recommendations:

- Identify options for future linkage across databases

- Support information exchange with existing IPD sharing initiatives from other disciplines (e.g. public health)

6.4 Access mechanisms and technical characteristics

6.4.1 What is the most suitable access and data storage mechanism?

Clinical trials capture personal data on participants, often of a sensitive nature. To protect patient privacy, and to mitigate against inappropriate use of the data, the majority of experts consulted considered reviewed access, with data held by a trusted third party, the most suitable access mechanism.

Existing databases employ a mix of various levels of administrative and scientific review for data requests. The majority of experts consulted felt that data requests should undergo some kind of review, including a screen for suitable scientific objectives and adequate research capability of the requestor, to safeguard against misuse or misinterpretation of data. We recommend a more detailed comparison of review parameters of existing data sharing initiatives to identify best practice and challenges, leading to the development of an effective, streamlined process.

While reviewed access was deemed to be most suitable by the majority of survey respondents, a substantial proportion felt that open access was the most suitable access model. In addition, a number of open access databases already exist (e.g. FREEBIRD, International Stroke Trial, and ITN TrialShare). We hence recommend that a future IPD repository should consider integrating open access options, through which some data providers can make their (suitably de-identified) data available without review, if they wish to do so. This open access option will further enhance discoverability of all IPD datasets. Monitoring of demand and use (or misuse) of these open access datasets will provide important information to steer future decisions. For example, should demand via the open access track prove to be much higher than via the reviewed access route, with interesting research outcomes, and in the absence of misuse of data, options for broadening open access to “restricted” datasets ought to be considered. This could involve a higher level of de-identification of datasets held under reviewed access (e.g. by removing additional, potentially sensitive parameters from the dataset) in line with trial participants’ informed consents.

Recommendations:

- Develop a repository model with reviewed access and data held by a trusted third party
- Carry out a detailed comparison of review parameters of existing data sharing initiatives to identify best practice and challenges, and develop an effective, streamlined process
- Incorporate open access options to allow data providers to make suitably de-identified data available without review, should they wish to do so, and monitor demand, actualised risks, and research outcomes to inform further efforts

6.4.2 What are the minimum requirements regarding data format?

Experts consulted were strongly in favour of providing extensive technical information and documentation along with deposited datasets to enable repository users to understand and use these correctly. Similarly, “lack of data quality” was the primary reason given for why researchers may not want to request data from a future IPD repository.

Hence, datasets should be curated to a high standard, e.g. by checking that the data deposited reproduce the published analysis of the trial. In order to lower the burden on data providers, this requires appropriate tools for data handling, and dedicated staff at either the data provider’s organisation (particularly universities) or at the repository to support and monitor these activities.

Recommendations:

- Adopt or develop, and test data handling tools to facilitate data deposition

- Investigate staffing needs and “data manager” roles to provide support at the repository or academic institutions to assure high data quality

6.4.3 Should repository data be available in a common format?

The majority of survey respondents considered the lack of common data standards a barrier to current research using IPD, and indicated a strong preference for harmonised data in a future repository. However, practically, harmonisation of data across the entire repository from the outset is unlikely to be achievable given the costs and effort involved.

One option would be to only carry out data harmonisation in priority research areas by collaboration-type working groups, as and when needed, and funded, using the stock of well-curated datasets within the repository as the starting point. This will require the availability of support staff that can be “recruited” to these efforts, with costs covered by funding from relevant funding bodies, e.g. disease-specific charities (see also Section 6.3.2). Such harmonisation will also help to adopt and further develop common data standards, which can be applied in future trials, reducing the effort of harmonising data at the point of data deposition.

Recommendations:

- Adopt a case-by-case approach to harmonisation, e.g. only in research priority areas, rather than aiming to harmonise all data from the outset
- Establish processes for harmonisation of IPD across trials in priority research areas that offer individual funders the option of carrying out these activities

6.4.4 What is the appropriate data analysis environment?

Most survey respondents from non-commercial backgrounds felt strongly that a future repository should allow the user to download datasets, and that researchers should be able to use any analysis software. Industry respondents assigned less importance to these factors.

Ultimately, it is currently up to individual trial sponsors to specify if data can be downloaded or only accessed within a secure environment. In order to maximise research benefits, it will be important to implement a repository model that can incorporate as many datasets as possible. This study did not investigate the proportion of data providers who would not be able to provide data to a repository that allows downloading of data. In addition, given the size of some data holdings, the cost of transferring and storing data elsewhere may not be practicable. In order to develop a suitable data platform, understanding these requirements is necessary.

This point should also be revisited after the recently initiated commercial data sharing initiatives have gathered more experience and the actual level of risk of downloading data can be assessed with more confidence.

Recommendation:

- Implement an IPD repository model that allows the user to download data when permitted by the data provider
- Investigate the need for a secure data environment for analysis, as determined by the proportion of data providers who would not be able to deposit data if it were downloaded by repository users

6.5 Areas for further investigation

The study identified a number of important areas to consider further should a new IPD repository be established in the future. These areas are described below.

(We did not establish if evidence to inform these areas already exists, or if current initiatives are developing solutions, as this falls outside the scope of this study.)

6.5.1 Incentives for academia

Data generation is a long and arduous process and sharing datasets with others is not currently recognised as a particularly important or valuable activity in many academic settings[64]. Many interviewees from academia agreed that there was an element of complacency by not depositing and sharing the data, letting datasets "sit" on individual investigators' computers, without adequate care for their preservation.

However, we found that even when deposition of data was mandatory as a condition of funding, as is the case with all repositories of the US National Institutes of Health we profiled, repositories are experiencing substantial issues with compliance. This highlights a potential lack of incentives and / or compliance monitoring mechanisms. Any future data repository for academic trials will need to take these issues into consideration.

The question of suitable incentives should be examined in more detail with the academic community, seeking to balance the burden of depositing data with potential benefits to the data provider (see Table 8) and the research community. As one interviewee explained "A gradual approach is needed: first make it possible, then encourage it, finally mandate it. Bring people with you".

In order to achieve this evolution, without creating additional barriers to research, continued consultation and dialogue will help building consensus and support around this issue, and contribute to a mind-set change in the community. A survey respondent considered "[a central IPD repository] would be a very helpful mechanism for enabling the sharing of data, but the culture needs to change – providing the mechanism is just one part of it". Consultations could take place through a widely distributed request for comments, or a series of workshops. We recommend these consultations also address the role of publishers, currently the "gatekeepers" of formal scholarly research, in incentivising (or potentially monitoring for compliance of) sharing of and access to research data. This could link in with existing efforts in this community (for example, see [65].)

6.5.2 Connecting the data requestor with the original researcher / trial sponsor

We found that most potential data providers were deeply concerned about the "loss of control" over datasets deposited in a central repository. One option to address this issue, and to mitigate data misuse would be to inform the original research / trial sponsor as and when their data have been requested and by whom. It would then be up to the data provider and user to make contact. This would be in the interest of both parties, with the former able to retain some level of oversight over how the data are used, and the latter able to ask questions to avoid misinterpretation of the data. Another model would be to offer the original researcher the option to view and comment on, but not veto, proposals (alongside an independent review board). Arguments for and against this approach, as well as alternative approaches, should be discussed further with the research community.

6.5.3 Data curation tools and support

The survey results make a clear case for the need to deposit data that has been curated to a high standard, to ensure the deposited datasets are of sufficiently high quality for use by researchers who were not involved in the original trial. Should a new repository be established, it will be important to define the minimum steps required for appropriate data preparation, to gauge the suitability of existing tools to facilitate this process, and to determine needs and associated costs in this area to lower the barrier to deposition as far as possible (e.g. development of additional tools, or introduction of dedicated data managers, at

[64] For example, see the report "Establishing incentives and changing cultures to support data access", May 2014, Expert Advisory Group on Data Access.

[65] Lin, J & Strasser, C (2014) Recommendations for the role of publishers in access to data. PLOS Biol 12(10): e1001975.

a central repository or at major research institutions, to support data preparation on request).

6.5.4 Cost of different data access models

An understanding of the costs of different data access models, such as required IT infrastructure and data platforms, data curation and standardisation efforts, assistance to the data user, and any other components will be important should options for a central repository be considered. Existing information can feed into this process (e.g. from existing databases, from clinical trials as well as other disciplines). This may allow the assembly of a "menu" of budget options, including a cost-benefit analysis for different scenarios, and serve as a basis for discussion of sustainable funding models. The cost of including or linking to other types of data (such as data from observational studies and medical images) may be included in this analysis.

6.5.5 Public information and creating support

Clinical trials are conducted with the involvement of many members of the public, both healthy subjects and patients. There is a consensus across stakeholder groups that the collected data has to be preserved and used to the maximum for the advancement of health sciences and benefits to current and future patients. Any future repository model must take into account the views and concerns of those who participated in the trials. While this study did not investigate on-going efforts, and any potential gaps, it was evident from the information gathered that public communication and discussion will continue to be an important area.

6.5.6 Data ownership and legal responsibility

Questions around legal ownership of data and results of analyses, as well as legal responsibility for the data (e.g. who carries the legal responsibility in case of breach of patient privacy) were raised repeatedly in interviews. These legal issues can be expected to have a substantial impact on future data sharing initiatives, and will hence need to be understood in detail. This could include the development of a globally-agreed standard for de-identification, setting clear guidelines that data providers can adhere to, and provision of appropriate protection for trial participants' privacy, as well as legal protection for the data provider in case a breach of privacy takes place.

6.5.7 Patient consent forms

Many consent forms do not cover secondary use of data. Survey respondents were concerned about the lack of clarity if use of anonymised data for secondary analysis required explicit consent from the trial participant. Appropriate wording covering secondary use of data should be developed and become a standard part of all consent forms, in order to remove this barrier going forward. There are initiatives underway to develop draft wording that can be included in consent forms for any sponsor (commercial or non-commercial), e.g. the Harvard Multi-Regional Clinical Trial (MRCT) group[66].

The findings of this study provide an overview of current research uses of individual participant data from clinical trials, as well as potential future uses and possibilities to enhance access. Alongside the above areas for further investigation, we hope that these findings will form part of ongoing global discussions and efforts to realise the full potential of clinical trial data to inform research practice, generate new findings and, ultimately, benefit patients.

[66] http://mrct.globalhealth.harvard.edu/informed-consent-language (accessed 11 Dec 2014)

Appendix A Individual participant data sharing initiatives and research impacts

Overview

This Appendix profiles a range of case studies on existing initiatives for sharing of individual participant data (IPD) from clinical trials, and examples of research combining IPD from multiple sources. It reflects the most up-to-date information publicly available on the various initiatives at the time of writing (December 2014).

Table A 1 provides a summary of the key characteristics of the "families" of data sharing initiatives covered in this section. For further detail on database families, see Section 2.1 in the main report; additional detail on individual databases is provided in a comparison table in Appendix B. Table A 2 presents an overview of profiled research projects and objectives.

Table A 1 Data sharing initiatives

	Collaboration of trialists/trial sponsors	**Disease-specific data repository**	**Funder-mandated access (NIH)**	**Commercial trial repository or data portal**	**Open data sharing by individual research units**
Data sharing initiative	EBCTCG, C-Path consortia, WWARN, IMPACT	PRO-ACT, C-Path CODR AD, Sylvia Lawry Centre	NIH: NIDDK, BioLINCC	ClinicalStudyDataRequest, YODA	FREEBIRD
Disease-specific data	Yes	Yes	No	No	Yes
Data source	Academic and commercial	Academic and commercial	Academic	Commercial	Academic
Treatment arm / control arm	Both	Control arm only[a]	Both	Both	Both
Data harmonised by:	Database staff	Database staff	Data provider/ Database staff / User	Data user	n/a
Access approved by:	Data provider	Database staff (scientific)[b]	Database staff (administrative)	Independent review board	None
Data held by:	Data custodian	Data custodian	Data custodian	Trial sponsor	Original research unit
Funding source	Public funders, industry, foundations	Public funders, industry, foundations	Public funders (US NIH)	Industry	Public funders

Table A 2 Case studies of research using individual participant data from multiple sources

Research category	Research topic	Data gathered by:	
Efficacy and safety of therapies	Tamoxifen in treatment of early breast cancer	EBCTCG	Covered in EBCTCG database section
Modelling disease progression Identification of new biomarker candidates	Algorithms to predict ALS disease progression	PRO-ACT	Covered in PRO-ACT database section
Informing policy (driving standards)	Prognostic model for epileptic seizure recurrence	Individual research group	
Aiding design and methodology of clinical trials	Alzheimer's Disease clinical trial simulation tool	C-Path Open Data Repository	Covered in C-Path / CAMD CODR section
Dose optimisation in a patient subgroup Assessment of parasite drug resistance levels	Antimalarial combination therapy in young children	WWARN	Covered in WWARN database section
New surrogate outcome measures	Qualification of biomarker in polycystic kidney disease	C-Path consortium	Covered in C-Path consortium section
Identification of an earlier clinical endpoint	Approved use of 12 week endpoint, rather than 24 week, in chronic Hepatitis C trials	FDA study	
Early detection of emerging drug resistance	Molecular markers of malaria parasite resistance	WWARN (use of clinical and molecular data)	Covered in WWARN section
Prognostic models Common data standards Improved trial design	Dealing with heterogeneity in causes, pathophysiology, treatments and outcomes of traumatic brain injury	IMPACT, FREEBIRD	Covered in IMPACT database section
Comparison of efficacy and safety profile of different treatments Aiding design and methodology of clinical trial	Anti-epileptic drugs	Individual research group	
Treatment efficacy in patient subgroups	Surgical interventions	Individual research group	

The following sections, A 1 – A 5, describe examples of data sharing and research outcomes for each cluster. Section A 6 presents case studies of research successes employing IPD datasets gathered by individual investigators.

A.1 Collaborative groups of trialists / trial sponsors

A.1.1 Early Breast Cancer Trialists' Collaborative Group (EBCTCG)

The Early Breast Cancer Trialists' Collaborative Group (EBCTCG) overview is a major collaborative endeavour that investigates the treatment of women with early (or operable) breast cancer. The collaboration first came together in 1983 to discuss the combination of the results of RCT of tamoxifen and chemotherapy. Currently, the collaboration involves around 300-400 research groups across the world - essentially all groups conducting randomised trials on treatments of women with early-stage breast cancer, where a main outcome is mortality.

The EBCTCG overview takes place in cycles lasting approximately 5 years, going through the stages of study identification, data collection, processing and analyses, presentation and discussion of the results by the collaborating researchers, and publication of these results. Following extensive searches for published and unpublished trials, investigators from academia and industry who conduct randomised trials on early breast cancer with survival as the major outcome are invited to join the group. Trial data that address one of EBCTCG priority research questions are collected in the central database, located at the Clinical Trial Studies and Epidemiological Studies Unit (CTSU) at the University of Oxford in England, the base for the EBCTCG Secretariat. The database staff select the variables relevant to the EBCTCG in discussion with the EBCTCG Steering Committee and converts the submitted data to a highly structured format (excluding data that might be submitted but are not required for the overview, such as data on quality of life measures or some toxicity effects, as these are outside the remit of the group).

While the data are held in Oxford, the contributing investigators retain ownership of their data. Other researchers can request access to datasets in the database to conduct their own analyses, but have to contact the data owner for approval before it can be transferred by the EBCTCG Secretariat[67].

The EBCTCG database currently holds data from around 700 clinical trials.

Efficacy and safety of therapies: Tamoxifen for women with early breast cancer

Combining data from multiple trials has allowed the EBCTCG to reliably assess moderate treatment effects. For example, the collaborative group analysed data from 37,000 women with early-stage breast cancer in 55 trials of immediate tamoxifen treatment, comprising about 87% of the worldwide randomised evidence[68]. The IPD meta-analysis provided strong evidence that tamoxifen treatment substantially improved the 10-year survival of women with endocrine receptor positive (ER+) tumours, irrespective of other patient characteristics or co-treatments.

As data on long-term outcomes become available, the EBCTCG carries out updates of their meta-analyses. Following on from the 1998 paper, a study published in 2011[69] looked at the long-term outcomes of around 21,500 women with early-stage, ER+ breast cancer who had received more than 5 years of tamoxifen treatment. This encompassed data from 99% of all women known to have been randomly assigned into trials of about 5 years of adjuvant tamoxifen, with a median follow-up of 13 years. The findings demonstrated that rather than simply delaying an inevitable event, 5 years of tamoxifen treatment prevented a high proportion of recurrences even 10 years after treatment, potentially curing many patients. These results allow clinicians and women to make well-informed decisions about treatment, with confidence about the likely effects of tamoxifen on breast-cancer events and overall survival.

Findings published by the EBCTCG have been embedded into clinical practice and guidelines for treatment of women with early breast cancer across the world, and have informed the design of planned clinical trials. The results have been incorporated into clinical decision and survival prediction tools, and fed into clinical treatment guidelines, not least because members of the EBCTCG sit on guideline committees and can contribute first-hand knowledge of the analyses. The collaboration has given rise to an extremely well-networked research community, facilitating information exchange between groups and avoiding potential duplication of efforts.

[67] This, however, happens rarely, as the data have already been exhaustively analysed through the EBCTCG overview.

[68] Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. Lancet (1998) 351: 1451-67.

[69] Early Breast Cancer Trialists' Collaborative Group (EBCTCG) et al (2011) Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet 378: 771-784.

A.1.2 The Worldwide Antimalarial Resistance Network (WWARN)

WWARN, the Worldwide Antimalarial Resistance Network[70], was established in 2009. Its overarching aim is to develop a collaborative network to assess the impact of antimalarial drug resistance, and to generate reliable evidence necessary to inform malaria control, elimination and eradication efforts. WWARN is primarily funded by the Bill and Melinda Gates Foundation, but also receives support from USAID (US), DfID (UK), ExxonMobil Foundation, and the French Foreign Ministry.

WWARN collaborates worldwide with 230 partners engaged in clinical trials on anti-malarial drugs, including academic research institutions as well as industry. To date, WWARN's database holds records from approximately 100,000 individual patients, generated in 350 clinical trials (phases 2 - 4). In addition to patient data from clinical trials, the database includes pharmacological, molecular, and *in vitro* data (the latter are mainly from routine surveillance rather than clinical trials).

Once received, the WWARN platform transforms submitted data to a common format, curates the information available, and re-analyses the study. This allows datasets to be harmonised so that they can be combined. The WWARN website also offers an interactive mapping application, WWARN Explorer[71], to allow custom queries of the large number of studies held within the repository (in aggregate format), and to visualise summary results using dynamic mapping. While the WWARN data platform provides the infrastructure for sharing data, data ownership remains with the primary data provider.

Two populations of researchers use the data:

1. The majority of work is carried out by WWARN itself, through the network's six scientific groups located around the world. Each group specialises in different aspects of malaria drug resistance. To date, the WWARN Scientific Advisory Committee and executives have formulated priority research questions and scanned the data held within the database to identify suitable datasets for inclusion in the study. They then contact each of the primary investigators to discuss their participation in the project. The primary investigators have the right to refuse, but to date all investigators have agreed to participate.

5. WWARN also enables data access for external research groups that approach WWARN with a specific research question. WWARN uses a data mining system to identify the best data to use, and approaches the primary investigator to discuss their participation in the project. The data may not be used for commercial purposes.

The balance between data sharing and retention of data ownership is considered key to WWARN's data sharing model. Important research questions can be addressed using the large number of standardised clinical trial datasets via WWARN's data sharing platform, which is a complex (and costly) piece of research infrastructure that could not be developed by individual researchers. At the same time, researchers are kept apprised of how the data may be used, and retain the option to either participate in the resulting projects, or to block access[72]. This incentivises participation as it may provide benefits to the careers of data providers, through new international collaborations and co-authorship on resulting publications (the "currency" of academic research). In addition, the approach alleviates contributors' concerns over potential misinterpretation of the primary ("their") data. This system has allowed WWARN to build a relationship of trust with industry, and companies have started to share unpublished data (with embargo until publication or filing with a regulatory agency). WWARN is therefore more than a data repository – it offers a platform for collaborative research, and access to expertise through its own researchers.

[70] The initiative's name reflects its initial aim of collecting all available surveillance data of antimalarial resistance, across jurisdictions, in order to develop a single data sharing platform. However, WWARN's focus shifted from surveillance to clinical trials data over the course of the project.

[71] http://www.wwarn.org/resistance/explorer (accessed 5 September 2014)

[72] Note however that researchers cannot veto publications resulting from the projects.

WWARN is currently in discussion to implement the same concept for visceral leishmaniasis in collaboration with the Drug for Neglected Diseases Initiative (DNDi) and parasitic helminths with WHO/TDR, as the system is sufficiently flexible to be adapted for clinical trial data from other diseases for a marginal cost, compared to the cost of building a similar platform for each disease from scratch. This captures an economy of scale by using a similar platform, which is especially important for neglected diseases where resources are extremely limited.

Dose optimisation in a patient subgroup: antimalarial treatments in young children

Widespread use of suboptimal drug treatment is known to lead to the emergence of resistant malaria parasites, particularly in Southeast Asia. To ensure that the dosing regimens remain as effective as possible, for as long as possible, it remains critical that all antimalarial drugs are monitored to assure optimal dosing for all patients. The artemisinin combination therapy (ACT) dihydroartemisinin plus piperaquine (DP) is an effective, widely administered drug combination prescribed for patients with malaria caused by the parasite *Plasmodium falciparum*. While ACTs are highly effective in most settings, resistance to some drug combinations has been reported, with the malaria parasite re-emerging in the patient.

A study carried out by WWARN, published in December 2013[73], provided new information about the extent of the dosing problem and its consequences. For this project, WWARN conducted a systematic search of the literature to identify all studies published between 1960 and February 2013 in which patients were enrolled and treated with DP. They then approached all principal investigators and invited them to share the IPD. Ultimately, 24 published and two unpublished studies were included in the analysis, from 26 clinical study sites in Asia, Africa and South America. In total, data from more than 7,000 patients were analysed, representing almost 80% of all available published data on DP.

Data were pooled using a standardised methodology. A series of dosing impact pooled analyses were conducted to assess how well the ACTs were performing. The results showed that although DP was still a highly efficacious drug, curing more than 97% of all patients, children under the age of 5 years were at a greater risk of treatment failure. The analysis also found that a third of these young children received a dose of piperaquine below the level recommended by the WHO, and piperaquine dosing was shown to be a significant predictor of re-appearance of malaria. Finally, the researchers estimated that increasing the target dose of piperaquine in children aged 1–5 years would halve the risk of treatment failure and cure at least 95% of young children. The data provided an indication that such an increase was not associated with gastrointestinal toxicity.

One problem with DP in its current formulation is that the approved weight-based dosing schedules led to some children not receiving the WHO-specified minimum daily doses. In addition, whilst clinical trials are rigorous and children are weighed carefully to dose them accordingly, most routine health clinics use age as a proxy for weight to determine the dose. Since the relationship between weight and age varies among children and localities, this increases the chances of under-dosing. The study strongly suggested that further detailed dose optimisation studies in young children were essential, to cure these patients, and to prolong the useful therapeutic life of DP by preventing/delaying the emergence of resistance.

The study provided evidence for public health policy-makers to review current dosing recommendations for DP for the under-5 age group, ensuring that the ACT remained therapeutically useful for as long as possible. The study was reviewed by the WHO Technical Expert Group on Malaria Chemotherapy who will base new recommendations on its findings.

Using the datasets assembled by WWARN, similar pooled analyses can now be done to assess the efficacy of other drug combinations currently in use.

[73] The WWARN DP study group (2013) The Effect of Dosing Regimens on the Antimalarial Efficacy of Dihydroartemisinin-Piperaquine: A Pooled Analysis of Individual Patient Data. *PLOS Med* 10: e1001564.

Early detection of emerging resistance: molecular markers of parasite resistance

In a second study[74], WWARN researchers used a large pooled clinical trials dataset to investigate the relationship between patient outcomes after treatment with two ACTs and genetic variation in the malarial parasite. Such an assessment is a critical step in validating molecular changes in parasite populations as useful markers of early signs of changing parasite susceptibility to commonly used antimalarial drugs.

The two most commonly used ACTs worldwide are artemether-lumefantrine (AL) and artesunate-amodiaquine (ASAQ), whose therapeutic efficacy remains high in most regions of the world. However, there have been some reports of decreasing AL cure rates in Africa and Asia, and reports of high levels of treatment failures of ASAQ.

Resistance to the long-acting component of ACT has been associated with specific changes in genes encoding *P. falciparum* drug resistance transporters. However, individual studies generally lack sufficient statistical power to assess the association between parasite genotypes and outcomes of clinical treatment. In a recent study, a WWARN research group investigated this link. IPD from AL or ASAQ treatment along with molecular markers of *P. falciparum* from 31 clinical trials were standardised and pooled. The drug treatment response of more than 7000 patients was analysed to determine whether patients infected with parasites of a certain genotype were more at risk of treatment failure. The pooled analysis showed that the genotypes of infecting parasites indeed influenced the outcome of AL treatment. However, this was not statistically significant for ASAQ treatment, probably because there was too little data available for this combination compared to AL. The study also provided evidence of strong selection of particular alleles by both drugs in recurrent parasites.

The results demonstrate that tracking these molecular markers can signal early decreases in susceptibility to two commonly used ACTs, confirming that application of molecular tools can offer cost-effective methods for early detection of parasite drug resistance. This would enable policy makers to manage emerging threats of resistance before clinical failure of a drug has occurred, and preserve the useful therapeutic life of these valuable antimalarial drugs for as long as possible.

A.1.3 The Critical Path Institute (C-Path) – collaborative consortia

The Critical Path Institute (C-Path) is a non-profit, public-private partnership with the US Food and Drug Administration (FDA). Based in Tucson, Arizona, USA, C-Path creates collaborative consortia where scientists from academia, industry, government agencies, and non-profit organisations formally agree to share information, develop scientific consensus and make findings available for public use. C-Path's aim is to accelerate the pace and reduce the costs of medical product development through the creation of new drug development tools and methodologies such as biomarkers, clinical outcome assessment measures, in-silico models and the development of data standards to support these efforts, which aid the scientific evaluation of the efficacy and safety of new therapies.

To date, C-Path has initiated seven global consortia that work on product-independent tools for drug development, including some for specific disease areas: Alzheimer's Disease (AD), Parkinson's Disease (PD), Tuberculosis (TB), Multiple Sclerosis (MS), and Polycystic Kidney Disease (PKD). Other C-Path consortia apply measurement science methods to develop safety biomarkers that are indication-agnostic and patient-reported outcome measures for specific indications.

[74] Venkatesan, M et al (2014) Polymorphisms in *Plasmodium falciparum* Chloroquine Resistance Transporter and Multidrug Resistance 1 Genes: Parasite Risk Factors That Affect Treatment Outcomes for *P. falciparum* Malaria After Artemether-Lumefantrine and Artesunate-Amodiaquine. Am J Trop Med Hyg 91: 833-843.

C-Path is funded through a variety of mechanisms, depending on the project. For example, support is provided by FDA grants, membership fees from consortia members, and grants from private foundations.

C-Path holds data from over 50 clinical trials, contributed by its consortium members. In many cases, clinical data are converted to a common data standard, the Clinical Data Interchange Standards Consortium's Study Data Tabulation Model (CDISC SDTM) standard, and integrated into a comprehensive dataset. Data providers retain ownership of the data, with C-Path establishing a data use agreement with the data-contributing organisation for each contribution of data. This specifies the types of data that are to be transferred to C-Path, how broadly the data can be made available for access, and the anticipated uses of the data. For the most part, data are used only internally, to support the regulatory objectives of each C-Path consortium. For the C-Path CAMD Alzheimer's disease database, the data contributors agreed to make the data available to external researchers via the C-Path Online Data Repository (CODR).

Each consortium is supported by C-Path's data management team. After the consortium outlines a research project, subject matter experts in the consortium identify the most important sources of data needed to provide the evidence, which supports the development of the proposed tools. Once data contribution agreements have been negotiated between C-Path and the contributing organisations, C-Path's data managers process and curate the data, including conversion to CDISC SDTM standard format if applicable. The data are then loaded into CODR. Analysis data extracts are provided to the consortium working groups, which carry out the analysis.

New surrogate outcome measure: Qualification of biomarker in polycystic kidney disease

Autosomal Dominant Polycystic Kidney Disease (PKD) is a genetic disease affecting around 12 million people worldwide for which there is currently no known cure or effective treatment. Traditional endpoints of renal function only show changes very late in the course of the disease, making it difficult to assess the effectiveness of new treatments.

The PKD Outcomes Consortium is a collaboration between the C-Path Institute, the PKD Foundation, Clinical Data Interchange Standards Consortium (CDISC), four academic medical centres (Tufts University, University of Colorado, Emory University, and the Mayo Clinic), and three pharmaceutical companies. In addition, a representative from the FDA serves as an active advisor to the consortium.

The consortium used standardised data from 3 patient registries and 2 observational studies to prepare for qualification of an imaging biomarker, total kidney volume (TKV), as a surrogate outcome measure in clinical trials. TKV can be used to assess disease progression at an earlier stage than the currently used measure (glomerular filtration rate), when patients may be *more likely to respond to new therapies*, and before irreversible damage has occurred.

At the time of writing, the biomarker was in the final stages of review for qualification as a surrogate outcome measure with the FDA and EMA. Once qualified "fit for use" in evaluating the efficacy of new treatments for ADPKD, its use can accelerate drug development, without drug developers having to re-confirm its utility.

A.1.4 The IMPACT project

The International Mission on Prognosis and Analysis of Clinical Trials (IMPACT) project was a project aimed at optimising clinical trial methodology in the field of Traumatic Brain Injury (TBI), to maximise the chance of demonstrating benefit of effective new therapies. IMPACT represented a collaboration between the University Hospital Antwerp, the Erasmus University Medical Center Rotterdam, the University of Edinburgh, and the Virginia Commonwealth University and was funded via a grant from the US NIH National Institute of Neurological Disorders and Stroke (NINDS) from 2003 until 2011. The collaboration assembled a database of IPD from twelve randomised clinical trials and five observational studies. This data, along with data relevant to the design and analysis of pragmatic clinical trials, including pre-hospital, admission, and post-resuscitation assessments, information on the acute management, and short- and long-term outcome, were merged into a top priority data set. Of the 12 clinical trials datasets, 11 were from commercial trials (all of which had been negative).

Prognostic models, common data standards, and improved trial design: bringing together research in Traumatic Brain Injury

Traumatic brain injury (TBI) is a leading cause of death and disability worldwide. Each year more than 1.5 million people die and about 10 million people are hospitalised after traumatic brain injury[1]. Prognostic models with admission data are essential to support early clinical decision-making, and to facilitate reliable comparison of outcomes between different patient series and variation in results over time. It also allows for appropriate counselling of patients' families, and can play a role in the design and analysis of future clinical trials.

Clinical trials of TBI provide significant challenges: trauma is a neglected research topic worldwide, consent in unconscious patients requires careful ethical consideration, and the injuries are very heterogeneous in terms of mechanism and pathology.

In 2008, two studies were published, both of which developed prognostic models to predict clinical outcome six months after TBI. The first study[75] used data from the CRASH trial, with a patient population of over 10,000 (see also Section A.5.1). The second study[76] used data from the IMPACT database, which - at the time - combined patient data from eight clinical trials and three observational studies to give a patient population of over 9,000. Each study used the other data set to cross-validate their conclusions.

Before these studies, prognostic models for TBI had been developed from small samples of patients, had poor methodology (for example, in over half of the models, loss to follow-up was not reported), were rarely externally validated, and were not clinically practical. Using data from much larger patient populations enabled the investigators to develop more robust prognostic models. In addition, the two datasets allowed the models to be validated externally (against each other), improving confidence in the results.

The resulting prognostic models were made accessible to clinicians via a Web-based calculator and have since been validated against new datasets[77] and explored in different local contexts[78] [79].

By 2011, the IMPACT database contained data from 12 clinical trials (including CRASH) and 5 observational studies. Other impacts achieved by this project include the examples profiled below (see also [80]). By 2014, (re)analysis of the IMPACT data had produced 62 publications[81].

1) Common data elements for TBI trials

During the data-gathering phase of IMPACT, the lack of common data elements for research in TBI emerged as a major challenge to combining data for cross-trial analysis. In response

[75] Perel, PA et al (2008) Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. BMJ 336:425.

[76] Steyerberg, EW et al (2008) Predicting Outcome after Traumatic Brain Injury: Development and International Validation of Prognostic Scores Based on Admission Characteristics. PLoS Med 5: e165

[77] Roozenbeek, B et al (2012) Prediction of outcome after moderate and severe traumatic brain injury: External validation of the International Mission on Prognosis and Analysis of Clinical Trials (IMPACT) and Corticoid Randomisation after Significant Head injury (CRASH) prognostic models. Critical Care Medicine 40:1609-1617.

[78] Honeybul, S et al (2012) Access to Reliable Information about Long-Term Prognosis Influences Clinical Opinion on Use of Lifesaving Intervention. PLoS ONE 7: e32375.

[79] Olivecrona, M & Olivecrona, Z (2013) Use of the CRASH study prognosis calculator in patients with severe traumatic brain injury treated with an intracranial pressure-targeted therapy. J Clin Neurosci. 2013 20:996-1001.

[80] Maas, AIR et al (2013) Advancing care for traumatic brain injury : findings from the IMPACT studies and perspectives on future research. Lancet Neurol. 12: 1200-1210.

[81] Ferguson, AR et al (2014) Big data from small data: data-sharing in the 'long tail' of neuroscience. Nature Neurosci. 17, 1442–1447.

to this finding, IMPACT investigators initiated a process of standardisation of data collection in TBI studies, including proposals for definitions and coding of demographic characteristics, basic clinical data, biomarkers, neuroimaging, and outcomes. To address the needs of different trial designs, three versions for coding data elements were developed: a basic, an advanced, and an extended format with the greatest level of detail (which could be collapsed into the basic version), facilitating comparison across studies. As a result of these efforts, common data elements are now required in all observational studies and trials in TBI funded by NIH-NINDS, and a database using these standards has been instituted – the Federal Interagency Traumatic Brain Injury Research (FITBIR) informatics system, sponsored by the U.S. Army Medical Research and Materiel Command (USAMRMC)[82]. Some EU-funding calls in this area have also mandated the data standards' use.

2) Improvement in trial design

Findings of the IMPACT initiative included that reduction in trial sizes of up to 50% could be achieved with covariate adjustment and by applying innovative statistical approaches[83]. Subsequently, these recommendations for trial design have been adopted in many completed and on-going TBI studies. In addition, several acute stroke trials have been published that have used different aspects of the method described in the IMPACT recommendations (stroke trials also have to cope with a similar prognostic heterogeneity of patient populations).

3) Identify best practice and gaps across study centres

Gathering data from across different trials and study centres uncovered large between-centre and between-country differences in management and outcome. For example, the IMPACT investigators found a 3·3-fold difference in the odds of having an unfavourable outcome at 6 months between very good and very poor centres. This comparative effectiveness research offers opportunities to exploit the existing heterogeneity and differences between countries, centres, and patients in TBI to identify best practices.

A.2 Disease-specific data repositories

A.2.1 The C-Path Online Data Repository (CODR) for Alzheimer's Disease

The Coalition Against Major Diseases (CAMD) is a public-private-partnership coordinated by C-Path. CAMD's mission is to accelerate the development of therapies for neurodegenerative diseases, such as Alzheimer's disease (AD) and Parkinson's disease (PD), by advancing tools for use in drug development for regulatory approval. The coalition includes 21 industry members, 3 non-profit organisations, 2 institutes of the US NIH, and the FDA and EMA.

CAMD members have contributed clinical trial datasets from AD trials to the C-Path Online Data Repository (CODR) (see also case study A.1.3). CODR's datasets from AD trials were made available to external researchers starting in June 2010. At the time of writing, CODR contained IPD from 6,500 patients across the control arms from 24 clinical trials of AD and mild cognitive impairment (MCI). All data were remapped to a common data standard, CDISC SDTM v1.2/ SDTM Implementation Guide v3.1.2, which was further developed by CAMD in collaboration with CDISC to include AD-specific data types. The remapped data were subsequently combined into a single database. As part of this effort, C-Path and CDISC jointly developed the first SDTM Therapeutic Area User Guide, for AD, based on the additional AD-specific content developed for CAMD.

The data are openly available to CAMD members as well as to external qualified researchers following approval of a request for access. External researchers apply via the CODR website and have to provide their name, credentials, the name of the organisation they represent, and their reasons for requesting access. The application is reviewed internally by C-Path's

[82] https://fitbir.nih.gov/jsp/about/index.jsp, accessed 3 Oct 2104

[83] Murray, GD et al (2005) Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. J Neurotrauma. 22:511-7.

CAMD Executive Director. Data providers are updated regularly on access requests but they have delegated their role in the review process to C-Path.

Once approved, external researchers can access the entire AD dataset encompassing 24 clinical trials. The data in the database are de-identified to ensure privacy of study participants. Clinical trial source, company and development programme are also de-identified.

Over the 3-4 years since its launch, CODR has seen steady external demand for access from approximately 200 researchers worldwide.

C-Path launched a second external access database for tuberculosis (TB) in September 2014[84], at the request of the US Centers for Disease Control (CDC). (The CDC is a member of C-Path's Critical Path to TB Drug Regimens consortium, CPTR.) The database contains IPD from three clinical trials carried out by the CDC.

Aiding design and methodology of clinical trials: the AD clinical trial simulation tool

Pharmaceutical companies typically run simulations using in-house data before deciding on the various components of the trial design, such as sample size and optimal trial duration and treatment effect measurement times. Integrating data from numerous sources can provide a much fuller picture and help to further optimise clinical trial design.

The C-Path/CAMD's Online Data Repository (CODR) for Alzheimer's Disease (AD) was one of three data sources used to develop a quantitative clinical trial simulation tool, which can be used to optimise clinical trial design for mild and moderate AD trials[85]. The tool is based on a drug-disease-trial model that describes disease progression, drug effects, dropout rates, placebo effect, and relevant sources of variability. While it cannot replace actual clinical trials, the model is expected to yield useful information that can be incorporated into trial design such as dose selection, population inclusion, sample size estimates, and study duration. The AD trial simulator tool was endorsed by the EMA and FDA in June 2013[86]. Since this database became available to the research community in the middle of 2013, there has been a steady level of requests for the tool, with up to 10 requests per week from a wide cross-section of users, both academic and commercial. Access to AD clinical trials data via CODR has also led to at least seven publications.[87]

A.2.2 The PRO-ACT database

The PRO-ACT[88] database is a project coordinated and implemented by the non-profit organisation Prize4Life, whose mission is to accelerate the discovery of treatments and a cure for ALS (amyotrophic lateral sclerosis, also called motor neuron disease), in partnership with the North Eastern ALS Consortium and the Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital.

The database project received funding from the ALS Therapy Alliance in January 2011 with the purpose of obtaining as many patient data records from completed trials as possible.

PRO-ACT currently houses around 8,500 ALS patient records from 17 completed Phase II/III ALS clinical trials (10 commercial and 7 academic trials). The estimated cost of

[84] http://c-path.org/programs/cptr/ (accessed October 2015)

[85] Rogers, JA et al (2012) Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis. J Pharmacokinetics and Pharmacodynamics 39, 479-498.

[86] FDA:
http://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ucm180485.htm;
EMA:
http://www.ema.europa.eu%2Fdocs%2Fen_GB%2Fdocument_library%2FRegulatory_and_procedural_guideline%2F2013%2F10%2FWC500151309.pdf&ei=LeoHVNP_IIO8ggSm7IDQBQ&usg=AFQjCNGK1Nbr2TernqL4V4MUl1AZ22MP6Q . (accessed September 2014)

[87] More information can be found at: http://c-path.org/category/publications/camd-publications/

[88] The acronym PRO-ACT stands for Pooled Resource Open-Access Clinical Trials.

soliciting, cleaning, and harmonising these 17 datasets for import into PRO-ACT was $500,000.

For most trials, both treatment and control arm data are included, whereby trials generally "failed" (i.e. results were clinically and statistically not significant), with only one 'modestly effective' treatment currently available on the market (extending patients' life by an average of 2 months).

PRO-ACT went live in December 2012, after 2 years of discussions with sponsoring companies, followed by a period during which the NCRI cleaned, harmonised, aggregated and de-identified the data, in accordance with HIPAA regulations (removal of trial dates, medications tested, and study centre information). NCRI also hosts the data on its secure servers. Information currently available includes demographics, clinical information, family history data, functional measures, vital signs, and lab data (blood chemistry, haematology, urinalysis, adverse events, and concomitant medications).

The database is open to anyone with an acceptable research proposal. Eligibility guidelines were agreed by a data access committee, which includes representatives of the companies contributing the data. Prize4Life staff review individual requests for fit with these guidelines, and keep the access committee informed through update reports at overview level. If the request is approved, researchers can download all or some of the data types in the database, as Excel or text files, and can run their analyses as needed.

By July 2014, Prize4Life had received over 350 requests from researchers from industry and academia.

Analyses using data in the PRO-ACT Database have allowed identification of novel variables correlating with survival, and enabled initial stratification of patients by progression slopes. By May 2014, 3 scientific papers had been published, and at least 4 additional papers were in preparation. As an example, researchers in one study were able to distinguish two discrete subpopulations of patients: slow progressors and fast progressors[89]. This distinction can now be used to implement a population enrichment strategy to control the level of heterogeneity in the patients included in new trials.

Modelling disease progression: Algorithms to predict disease progression in ALS

Amyotrophic lateral sclerosis (ALS), also referred to as Lou Gehrig's Disease or Motor Neuron Disease (UK) is a progressive neurodegenerative disease that affects nerve cells in the brain and the spinal cord, leading to paralysis. There is currently no known cure for the disease. Following the onset of symptoms, patients live for another 3-5 years on average; however, the disease progresses at markedly different rates – a long-surviving well-known patient is Professor Stephen Hawking who was diagnosed more than 50 years ago. ALS affects one in 1000 individuals.

In 2012, ahead of the launch of the database, the non-profit organisation Prize4Life, in collaboration with the DREAM Project, announced a prize competition in which solvers used a subset of the PRO-ACT dataset to develop algorithms which predict the progress of ALS[90]. A second set of data was used by the challenge managers to validate the algorithms. The challenge resulted in the submission of 37 unique algorithms. By way of an online scoring system, participants could measure the accuracy of their own algorithm versus a blinded dataset, and compare their performance to that of competitors.

The six best performing algorithms were able to identify common ALS predictive features (e.g. age, sex, site of onset) as well as several novel features, such blood pressure, pulse, phosphorus, and creatinine. None of these are currently routinely assessed in clinical

[89] Gomeni, R et al (2014) Amyotrophic lateral sclerosis disease progression model. Amyotroph Lateral Scler Frontotemporal Degener. 15:119-29.

[90] Kueffner R et al (2014) Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nature Biotechnology doi:10.1038/nbt.3051 (published online 2 Nov 2014)

practice, and all represent potential new lines of inquiry as ALS biomarkers. In addition, the winning algorithms predicted disease progression better than each of the 12 top clinicians (when provided with the trials data, rather than patient examination). Modelling suggests that use of this tool could reduce the number of patients needed for a new clinical trial by 23%, representing a significant reduction in trial cost. The algorithms are currently being tested by companies re-visiting their clinical trials data to determine if patient stratification can explain the (negative) study results, as well as by companies planning new clinical trials.

A.2.3 The Sylvia Lawry Centre for Multiple Sclerosis Research

The Sylvia Lawry Centre (SLC), a not-for-profit charitable organisation, located in Munich, Germany, was founded to promote research and education in the fields of medicine and natural science, in particular in relation to multiple sclerosis (MS). The Centre was launched in 2001 by the Multiple Sclerosis International Federation with co-funding from the National MS Society and other partners. The mission of the Centre is to improve human health by developing evidence based decision support tools for better clinical choices.

As part of this mission, the SLC has gathered data from the control arm of 29 randomised clinical trials addressing MS, conducted over the last 20 years. The datasets represent information on more than 26,000 individual participants, nearly all patients who have participated in the placebo arms of major clinical trials. Data were provided by pharmaceutical companies, universities, clinicians and researchers, including Bayer, Biogen, Idec, Serono, Schering, the Mayo Clinic, the University of California Los Angeles, and Addenbrookes Hospital, Cambridge UK. In addition, clinical centres have provided longitudinal data from a large natural history cohort.

The SLC built, hosted and maintained the entire MS database and also employed around 20 internal research staff for data analysis. However, the full dataset can only be accessed locally, i.e., at the centre in Munich, due to access restrictions imposed by data owners. (Metadata and analysis results can be transferred.) Between 2003 and 2008, 15 guest researchers visited the SLC to access the data.

In order to enable wider use of the data, the Centre has developed an open ‘synthetic dataset’ which was constructed to reflect the statistical properties of the entire database. This can be used to develop stable and robust analytical models, explore hypotheses, and confirm the plausibility of published research findings. The synthetic datasets can be downloaded and analysed with the Online Analytic Processing tools developed at SLC or with any other statistical software packages from external researchers’ desktop. Commercial organisations also use the datasets mainly to generate “virtual placebo groups” and increase sample sizes in clinical trials. The closed part of the dataset is then used for validation purposes only, as a joint project of the researchers and the SLC staff. The validation request by the researchers is checked for medical relevance and alignment with the Centre’s mission, and normally approved within two weeks. This approach ensures that datasets and the corresponding sponsors cannot be re-identified.

Recently, using the available data, a web-based prognostic calculator in MS has been developed that can serve as a decision support tool and is capable of delivering individualised estimates of disease progression[91]. It was shown that the tool was consistent in its predictive accuracy with low variability.

Analyses of the large body of data within the SLC database conducted by the research staff challenged some generally accepted beliefs in the field. Most MS clinical trials have used MRI derived parameters as surrogate marker to predict relapses and long-term evolution of disability. Multivariate analysis however provided a predictive statistical model using clinically relevant relapse and disability outcomes that were not improved by MRI measures.

[91] Galea, I., Lederer, C., Neuhaus, A., Muraro, P. A., Scalfari, A., Koch-Henriksen, N., Heesen, C., Koepke, S., Stellmann, P., Albrecht, H., Winkelmann, A., Weber, F., Bahn, E., Hauser, M., Edan, G., Ebers, G. and Daumer, M. (2013), A Web-based tool for personalized prediction of long-term disease course in patients with multiple sclerosis. European Journal of Neurology, 20: 1107–1109. doi: 10.1111/ene.12016

Furthermore, common clinical outcome measures for disability scoring in MS, which were routinely used in drug trials, could not be validated for the therapeutic efficacy of an intervention. These findings were considered in the development of the new draft guidelines on clinical investigation of products for the treatment of relapsing-remitting MS, issued by the EMA in 2012[92] and further discussed in a workshop in 2013[93].

A.3 Public-funder mandated repositories

A.3.1 Data repositories of the US National Institutes of Health (NIH)

The US National Institutes of Health (NIH) is a major public funder of medical research, with an annual expenditure on research grants of around $30 billion (2014). It is divided into 27 topic-specific institutes, who distribute research funds through their respective grant programmes. As set out in the NIH Data Sharing Policy, all investigator-initiated applications with direct costs greater than $500,000 in any single year are required to incorporate data sharing features in the application.

To fulfil this mandate, a number of NIH Institutes have (individually) set up databases where investigators can deposit their data. Two examples, the BioLINCC database of the National Heart, Lung and Blood Institute (NHLBI) and the repository of the National Institute for Diabetes, Digestive and Kidney Diseases (NIDDK) are described in more detail below.

1) BioLINCC - National Heart, Lung, and Blood Institute (NHLBI)

BioLINCC was set up by the National Heart, Lung, and Blood Institute (NHLBI) in 2000 to facilitate sharing of datasets and biospecimens from NHLBI-funded research. It contains treatment and control arm data from 82 clinical trials and 33 observational studies on heart, blood, and lung diseases (excluding cancer). The most “famous” dataset included is the (observational) Framingham Heart Study, which has been running since 1942. Where available, BioLINCC also provides access to biospecimen collections associated with these studies, which are stored in the BioLINCC biorepository.

Researchers request data by submitting information on the study protocol or proposed research plan and the data security measures to be used. They also have to provide an IRB approval/waiver statement for any level of access to the data. The request goes through an administrative review by the Repository Allocation Committee (NHLBI staff), confirming that the proposed use of data is consistent with the data agreement. After approval, datasets are transferred to the researcher in the format that they were received in (the NHLBI does not offer custom data solutions). Harmonisation of datasets within the repository is being considered, with its potential advantages being balanced against the high burden of cost.

Since 2000, approximately 640 investigators have received data. Nearly 35% of the requested datasets include data from clinical trials, i.e. around 220 requests over 14 years. (It should be noted that the actual re-use frequency of the datasets may be masked by the fact that most studies supported by the NHLBI share data readily with outside investigators, and do not require the involvement of BioLINCC.)

2) NIDDK Data Repository - National Institute for Diabetes, Digestive and Kidney Diseases

The NIDDK Data Repository was established in 2003 and is composed of three linked repositories for data, biospecimens and genotyping data from genome-wide association studies (GWAS) and sequencing studies. It includes 58 clinical study datasets from 43 distinct clinical studies on endocrine and metabolic diseases such as diabetes, digestive diseases, nutritional disorders, and obesity; and kidney, urologic, and hematologic diseases. Data are typically submitted in SAS format or converted to SAS upon receipt, but the

[92] http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/10/WC500133438.pdf

[93] http://www.ema.europa.eu/docs/en_GB/document_library/Other/2013/10/WC500151676.pdf

(both accessed Oct 2015)

repository is able to accommodate other data formats. Data are then curated, including a data safety and integrity check which involves reproducing the published results of the study.

In order to access the individual-level clinical data, requesters must provide a description of the proposed project, present proof of IRB approval or exemption, and sign a Material Transfer Agreement. Data requests are reviewed by NIDDK extramural scientific staff (for consistency with the data agreement) and, if need be, by an NIH bioethicist. Between 2005 and 2013, 309 data requests were approved (55 in 2013). At the time of writing, none had asked for datasets from different studies with the intention to combine these.

A.4 Commercial trial repositories and data portals

A.4.1 Clinical Study Data Request data portal

The Clinical Study Data Request (CSDR) portal enables researchers to view available studies conducted by a number of clinical trial sponsors and to request access to the underlying IPD. The repository was initiated by GlaxoSmithKline (GSK), and launched in May 2013. By December 2014, 1599 distinct datasets were available via the CSDR portal, from eight pharmaceutical companies: Boehringer Ingelheim (190), GSK (1058), Lilly (81), Novartis (6), Roche (60), Takeda (145), UCB (21), and ViiV Healthcare (38). These companies made a range of commitments regarding access to historical datasets, e.g. GSK will include all global interventional clinical studies that were ongoing or started after the formation of GSK in 2000, Roche will add studies going back to 1999, and Boehringer Ingelheim will list trials from 1998, with a likely 500 additional studies available by the end of 2015. These data concern medicines that had been marketed but also those from terminated research programmes. Both raw and analysis-ready datasets are provided, along with supporting documentation including the protocol, data specifications, annotated case report forms and clinical study reports[94]. Another three companies (Astellas, Bayer, and Sanofi) will make their datasets available via the CSDR portal in the future.

Clinical trials of rare diseases or very small number of subjects are not currently listed on the site, as anonymisation of these data is more difficult to achieve. However, if requested, the feasibility of anonymisation is individually assessed. In addition, sponsors may accept enquiries from researchers about the availability of data from other studies not currently listed on the website. These enquires are first reviewed by the sponsors who assess the feasibility to provide data from the study. The outcome of the enquiries are listed on the website.

Researchers may gain access to requested datasets by submitting a research proposal. The first step is a “requirements check” by study sponsors to ensure the information is complete and meets the requirement for informed consent. In historical trials, in general, informed consent was given by patients to use their data to study a specific medicine or disease. Secondary analyses of data are therefore restricted to the context of the original studies. More recently, patients are asked to give permission for broader research so other research may be possible with data from these studies.

In a next step, the proposal is vetted by an Independent Review Panel, currently consisting of 4 individuals from a range of backgrounds (clinical researcher, biostatistician, legal/ethics expert, and a patient representative). Proposals are screened for overall feasibility, scientific rationale, and relevance of the proposal’s approach and aims, qualification of the team, but do not undergo a detailed scientific evaluation. A decision from the Panel is normally communicated within 30 working days of a research proposal being submitted. Some sponsors may decline access to their data in exceptional circumstances, for example, where there is a potential conflict of interest or a competitive risk. In these cases, the reason for failing the requirements check would be listed on the CSDR website.

[94] Hughes S. et al (2014) Preparing individual patient data from clinical trials for sharing: the GlaxoSmithKline approach. Pharmaceutical Statistics 13: 179-183

Once the request has been approved and a Data Sharing Agreement is signed, the relevant sponsor(s), who holds the data, provide access to anonymised data along with study documents, in a password-protected, private workspace. This workspace is not accessed by study sponsors or others unless requested by researchers in order to resolve problems. Controls are in place to prevent researchers downloading the data to their computer.

Researchers can access the data via a secure internet connection, combine data from different studies, conduct research using statistical software provided (SAS and R), and finally download their analyses. Access to data is provided for 12 months, but extensions may be given up to 24 months. Access is free to researchers - all costs related to making data available on the CSDR website, including the internal staff and review panel, are paid for by the study sponsors.

In the spirit of transparency, the name and affiliation of the lead researcher, a declaration of sponsorship, the title of the research, the requested studies, a lay summary, and other relevant information are posted on the CSDR website. There is a requirement for the results to be submitted for publication in a peer-reviewed journal. The publication citation and statistical analysis plan are also posted on the website after the research has been published.

CSDR received a total of 58 requests over the first 12 months (May 2013 – May 2014); 45 of these met requirements; 36 were approved or approved with conditions; 23 signed data sharing agreements, and for 13 approved requests access to data has been enabled (status on 31 May 2014). All approved requests with signed data sharing agreements can be viewed on the CSDR website[95]. Of the 23 projects with signed data sharing agreements by May 2014, 12 have gained access to data from one trial only, and 9 gained access to data from 2 or 3 trials. The remaining two projects gained access to 8 and 11 datasets. Only one of the projects approved by May 2014 involves data from more than one trial sponsor; however, the website indicates that 4 such proposals have been received. The majority of these projects originated in the US (10) and the UK (5). Other countries account for a smaller number of approved projects: Australia (2), and one each from Belgium, Canada, India, Jordan, Spain and Switzerland. One researcher from the US has had two projects approved.

These approved proposals cover a range of research objectives, including the identification of predictive markers or risk factors in individual patients, development of prognostic models, comparison of effectiveness and safety of drug combinations (especially with newer treatments), dose optimisation, identification of improved clinical endpoints, and improvement of future trial design. Table A 3 below presents a short summary of proposal objectives and characteristics.

Many researchers interviewed felt that CSDR was an important step toward enhanced access to IPD through a single contact point, at no cost to the user. However, they also felt that the current system had weaknesses in that the user interface is not flexible enough. For example, researchers would like to combine CSDR data with datasets not available through the system, and use more diverse software packages for complex analyses. A survey respondent highlighted the need for sharing analysis results among co-investigators, and exporting/importing files and analysis scripts. One researcher reported[96] some of the cumbersome processes associated with the ‘remote desktop’ environment and working with case study reports in a ‘cramped’ environment.

It appears however that CSDR is taking the initial feedback from the research community into consideration, and a small number of requests to download specific datasets needed for the research have even been granted where risks to data privacy were minimised; this option may alleviate some of the reported obstacles going forward. The establishment of CSDR is a long-awaited first step in the right direction. It will be exciting to see the first scientific publications based on access to CSDR data and follow the developments of this commercial portal for the benefit of the clinical research community.

[95] https://clinicalstudydatarequest.com/Approved-Requests.aspx (accessed 26 October 2014)

[96] Jureidini JN & Nardo JM (2014) Inadequacy of remote desktop interface for independent reanalysis of data from drug trials. BMJ 349:g4353

Table A 3 Data request proposals to CSDR with signed Data Transfer Agreements (status: 31 May 2014)

Number of trials requested	Sponsor(s) of requested trial(s)	Requesting researcher country	Title of proposal	Type of study	Disease	Study aim
11	GSK	UK	Long-acting beta2-agonists for chronic obstructive pulmonary disease	Systematic review of treatment efficacy and safety	COPD	Assessment of the effects of long-acting beta2-agonists compared with placebo for patients with chronic obstructive pulmonary disease (COPD), upon clinically-important endpoints
2	GSK	Canada	Predictive and Prognostic Value of Blood Cell Ratios in Patients with Metastatic Renal Cell Carcinoma Treated with Pazopanib	Biomarker validation and predictive value assessment	kidney cancer	Determination of the predictive value of a readily available and inexpensive marker (based on blood counts determined in routine practice), for the likelihood of response to pazopanib in advanced kidney cancer
1	GSK	USA	Predictors of Benign Prostatic Hyperplasia (BPH) Progression	Identification of risk factors and predictors of disease progression; effect of medications and co-morbidities in disease progression	prostate cancer	Determinaton of variables that are most predictive of developing BPH (also known as benign prostatic enlargement) / having BPH progression, including patient demographics, medications and laboratory tests
1	GSK	USA	Relationship of Activated Clotting Time and Bleeding Related Outcomes in the FUTURA OASIS-8 Trial	Biomarker calibration for optimal dose of blood thinners balancing effectiveness and risks; effect of combined use of medications	arterial disease - coronary artery	Investigation of the relationship of Activated Clotting Time (ACT) value to bleeding and ischemic events and identification of an optimal ACT at which bleeding is minimised without any excess ischemic endpoints.
1	GSK	USA	Low-Dose vs Standard-Dose Unfractionated Heparin for Percutaneous Coronary Intervention in Acute Coronary Syndromes Treated With Fondaparinux in Patients with Peripheral Arterial Disease	Comparing effectiveness and safety of the combined use of newer treatments	arterial disease - peripheral arteries	Identification of an optimal anticoagulation strategy in patients with Peripheral Arterial Disease undergoing coronary angioplasty
1	GSK	USA	The Influence of Age, Gender, and Race on the Effects of Combination Therapy with Carvedilol plus Lisinopril Versus Lisinopril Monotherapy for Systemic Hypertension	Effect of combined use of medications, including new treatments, on minority patient groups	hypertension	Comparison of the effects of beta-blockers plus ACE inhibitor versus ACE inhibitor monotherapy in three main subgroups (race, age, gender)
2	GSK	UK	Factors influencing immune response and patient outcomes following influenza vaccination	Effect of combined medications, co-morbidities, and lifestyle factors on treatment outcome, in minority patient groups	influenza vaccination	Assessment of the influence of identified factors ie immunomodulatory drugs, comorbidities and lifestyle factors on immunogenicity following vaccination; effect of age (65+)

Number of trials requested	Sponsor(s) of requested trial(s)	Requesting researcher country	Title of proposal	Type of study	Disease	Study aim
1	GSK	USA	Assessing and Reporting Heterogeneity of Treatment Effect in Randomized Clinical Trials	Development of a new modelling technique, risk models for patient subgroups	method development	Development of mathematical risk models to help better understand which patients might get the most benefit from a new treatment, by examining effects on patient groups that differ in multiple characteristics simultaneously
1	GSK	Belgium	Descriptive evaluation of SAE reporting rates in vaccine clinical trials among Latin American children	Aiding the design of a future clinical trial on a related treatment to be conducted in a specific geographic location	vaccines	Estimation of the reporting rate of any serious adverse events and specific SAEs among Latin American infants, overall and by country, season, gender, age, and vaccine dose, to inform future clinical trial in vaccine development in same geographical location
3	GSK	USA	Functional Estimation of Interventional Effects	Development of a new statistical method for the determination of distribution data of drug effects	method development	Profile of the entire possible distribution of a bio-marker change in response to treatment
2	GSK	AUS	Associations between antihypertensive drugs and patterns of blood pressure changes: a strategy to reduce the burden of anti-VEGF induced hypertension	Understanding of side effect of combined medications, in patient subgroups	cancer	Understanding of 1) patterns of blood pressure changes with anti-VEGF drugs and 2) whether specific antihypertensive drugs or drug classes might be better than others in preventing and managing anti-VEGF induced hypertension and proteinuria
1	GSK	AUS	A multi-center, double-blind, placebo controlled study of paroxetine and imipramine in adolescents with unipolar major depression - efficacy and adverse outcomes	Re-analysis of data to check validity of trial conclusion	depression	Re-analysis of data to check validity of trial conclusion
2	GSK	USA	The Development of Toxicity Prediction Tools to Assist Oncologists in the Management of Adverse Events in Patients Receiving Treatment with Lapatinib	Development of a prediction tool for individual patient risk of side effects	breast cancer	Development of accurate prediction of side-effects of Lapatinib treatment for breast cancer patients at a higher than average risk for toxicity effects
1	GSK	UK	Antiepileptic drug monotherapy for epilepsy: an overview of systematic reviews and network meta-analysis	Comparison of effectiveness of different treatments	epilepsy	Overview of effectiveness of ten Anti-Epilectic Drugs currently licenced and used in clinical practice for use as monotherapy (efficacy, side-effects - for 2 different types of seizures)
2	GSK	UK	MASTERMIND: Stratification of glycaemic response the ADOPT and RECORD studies	Prediction of treatment response and potential side effects for individual patients, and identification of suitable biomarkers	diabetes	Identification of predictors of response to 3 types of glucose lowering therapy with different mechanisms of action. Confirmation that different responses are due to medication, rather than different disease progressions.

Number of trials requested	Sponsor(s) of requested trial(s)	Requesting researcher country	Title of proposal	Type of study	Disease	Study aim
3	ViiV Healthcare	Spain	The CD4/CD8 Ratio as a Predictor of Non-AIDS Events During Antiretroviral Therapy: Confirmation of its Predictive Value and Impact of Maraviroc in Treatment-Experienced Patients	Identification of prognostic markers, subgroups of patients	HIV	Confirmation that low CD4/CD8 ratio could indicated subtype of patients, in need of novel treatment course
8	Roche/GSK	UK	Assessing models for changes in tumour size over time and how they relate to survival times	Development of a disease progression model, determination of correlations between disease characteristic and survival, determination of effect of different treatments on survival in different disease scenarios	cancer (melanoma)	Building and testing of different mechanistic models of tumour growth; Assess if there is a relationship between empirical tumour size changes and survival, Assess if there is a relationship between parameter values from the different growth law models and survival
2	GSK	Jordan	Optimized Eltrombopag Treatment of Hepatitis C virus-related thrombocytopenia	Development of a dose optimisation algorithm for individual patients	Hep C infection	Development of an optimal dosing algorithm of eltrombopag, a platelet-raising medication (which is essential for allowing HepC treatment)
1	GSK	USA	A study of subgroup identification and micro aggregation	Statistical methods in subgroup identification	not disease-specific	
1	GSK	Switzerland	The implication of central adjudication of COPD exacerbations by experts for treatment effect estimates and sample size calculation	Validation of clinical endpoints; improvement of clinical trial design	COPD	Demonstration of the implications of endpoint assessment by the patient or individual experts, versus centrally by 'blinded' experts (trial endpoint: COPD exacerbations) by experts. More accurate assessment may reduce the necessary sample size and hence cost of future trials.
1	ViiV Healthcare	India	Pharmacokinetic modeling of dolutegravir in HIV patients	Development of an improved pharamcokinetic model to optimise patient treatment	HIV	Development of a highly established pharmacokinetics model of the anti-viral drug dolutegravir
2	GSK	USA	Assessing at the participant level the applicability of clinical trials to a specific patient	Prediction of individual patient outcomes vs average outcome for the entire trial population	method development	Development of a method comparing individual patients to a study population based on all coded patients' characteristics, to enhance clinicians' ability to tailor treatment options
1	GSK	USA	Predictors of Prostate Cancer Progression Among Men on Active Surveillance	Determination of predictors of disease progression; creation of a predictive tool	Prostate cancer	Definition of the variables, and combinations of variables, most predictive of prostate cancer progression; including factors such as age and body-mass index, PSA levels and kinetics, Gleason score, percent of positive biopsy cores, comorbidities and medication

A.4.2 The Yale University Open Data Access (YODA) Project

The Yale University Open Data Access (YODA) Project is an initiative of the Yale Center for Outcomes Research and Evaluation (CORE). The YODA Project was developed with the intention of offering:

2. A cost-effective, sustainable data sharing model, enabling all organisations, large and small, to disseminate their trial data to the larger research community;

6. An independent, academic, third party without interest in the data, removing the perception of influence over access; and

7. An established process for reviewing requests and associated registration materials to ensure that all required information is completely submitted, and the use of these data is intended to create or materially enhance generalisable scientific and or medical knowledge to inform science and public health.

The YODA Project started in August 2011 through a partnership with Medtronic, Inc. At the time, several law suits had been filed against the company relating to a controversial product, recombinant human bone morphogenetic protein-2 (rhBMP-2), and the company was under strong public scrutiny. In order to restore confidence, Medtronic provided Yale with a grant to lead an independent, systematic review of the entire body of scientific evidence on rhBMP-2. The grant also supported dissemination of these datasets to external researchers. Data from 17 rhBMP-2 clinical trials are now available via the YODA Project application process. From September 2013 to September 2014, the YODA Project received applications from 16 academic research groups; and data had been transferred to 4 of these (applications were incomplete for the remaining 12 requests).

The project's second partner, Janssen/Johnson & Johnson (J&J) approached Yale in mid-2013, and an agreement was signed in January 2014. Data sharing arrangements through an external request system started in October 2014. At the time of writing, the YODA website listed 81 trials. In the first instance, data from drug trials conducted by Janssen was made available. By September 2014, J&J had received nearly 100 expressions of interest for access to the data via the YODA Project. In January 2015, YODA and J&J announced that data from medical device and diagnostic trials was being made available[97].

The YODA Project's review board, consisting of Yale Faculty members, make all decisions regarding data access. J&J receive the information needed to conduct a feasibility assessment, but the YODA Project will ultimately determine which applications are approved. Researchers must inform the YODA project of publications arising from use of the data, which will be included on the project's website.

Unlike the Medtronic data, which is held by YODA and can be downloaded by approved researchers to their own computers, J&J makes its data accessible via a SAS data sharing platform. When the YODA review board approves a project, Janssen uploads the data to the SAS platform and YODA provides the data user with access to the particular data set. The researchers will then be able to run their analyses within a secure environment, and subsequently download the results of their work. However, if requests require data dissemination outside of the secure data sharing platform, these will be considered and evaluated.

[97] http://www.modernhealthcare.com/article/20150114/NEWS/301149949/johnson-johnson-becomes-first-devicemaker-to-broadly-share-trial-data (accessed 20 Jan 2015)

A.5 Open data sharing by individual research groups / unit

A.5.1 The FREEBIRD database

The FREEBIRD database was set up in 2011 by the Clinical Trials Unit at the London School of Hygiene and Tropical Medicine (LSHTM). It currently consists of two large clinical trials, CRASH and CRASH-2, which investigated the effect of treatments for adult trauma patients. Together, the studies involved more than 30,000 patients from across 49 countries. The database set-up up was funded by the UK's National Institute for Health Research (NIHR); running costs are absorbed by the Clinical Trials Unit budget. It is strongly supported by the consumer network and includes consumer testimony about the importance of data sharing.

FREEBIRD is available to any member of the public. After filling in a simple registration form, the dataset can be downloaded in CSV format, without an approval process. Identifiable personal information about participants, such as patient name, initials and the hospital ID number, are removed. In addition, the randomisation code is withheld, i.e., the data do not show which treatment was allocated to which patient. This was done to prevent users from drawing inappropriate conclusions about treatment effects, such as the effect of the treatment limited to a specific country, which the trial design would not support. However, users can request the randomisation code, accompanied by a detailed proposal for the study team to review for suitability. To date, this has occurred twice; for one project, the protocol is in preparation and for the second, the requester did not respond to the study team's additional questions.

One of the underlying premises for making the CRASH and CRASH-2 data widely available is that the LSHTM investigators do not consider themselves "owners" of these data: it was generated in more than 300 hospitals around the world, by numerous researchers.

Prior to FREEBIRD, the study team shared their data when contacted by external investigators. However, this represented a continuous data management effort (and expense), which was particularly difficult to accomplish once the LSHTM team had moved on to new projects. Doing the work "up front", while the study was still funded, facilitated broader long-term sharing. It also allows external researchers to peruse the data to see if it is suitable for their purposes. The team at the LSHTM Clinical Trials Unit intends to add data from currently on-going trials to the FREEBIRD database in the future, when these studies have been completed.

A.6 Individual participant datasets gathered by individual investigators

A.6.1 Identification of an earlier endpoint for clinical studies of chronic hepatitis C

Experts from the US Food and Drug Administration (FDA) have made use of safety and effectiveness data from multiple studies to address hurdles in drug development, such as identification of potentially valid endpoints for clinical trials, understanding of the predictive value of preclinical models, clarification of how medical products work in different diseases, and development of novel clinical designs.

One of these studies was the identification of an earlier endpoint for clinical studies of chronic hepatitis C (HepC) [98]. The primary endpoint for chronic HepC trials was based on detection of the virus at week 24 of follow up ("sustained virologic response"). Evidence suggested that assessing the response at earlier time points could provide an equivalent measurement of drug response. FDA scientists conducted an analysis of the combined data from 15 phase II and III clinical trials, 3 paediatric trials, and 5 drug development programmes to determine whether assessments conducted at earlier time points could provide results that were predictive of the outcomes at 24 weeks of follow up. The analysis determined that the sustained virologic response at 12 weeks of follow up was suitable as a primary endpoint for regulatory approval in clinical trials. This would also allow for HepC virus treatment options to be available earlier for patients suffering from this disease. In

[98] Chen, J et al (2013) Earlier sustained virological response end points for regulatory approval and dose selection of hepatitis C therapies. Gastroenterology 144:1450-1455.

addition, the study found that the sustained virologic response at 4 weeks of follow up could be used to guide dose and treatment strategies in trials.

As a result of these findings, the FDA has instructed pharmaceutical companies that they can use the response measurement at 12 weeks as a primary endpoint in clinical trials. The use of earlier time points for key regulatory decisions and dose selection may facilitate drug development for additional therapeutics under investigation.

A.6.2 Validation of a prognostic model for seizure recurrence following a first unprovoked seizure and implications for driving

In the UK and other European Union countries, the majority of people who have had a first unprovoked seizure are allowed to return to driving a car following six months without a subsequent seizure. This driving guideline is in part informed by prognostic modelling of data from a clinical trial, the Multicentre Study of Early Epilepsy and Single Seizures (MESS). The model included data from more than 600 participants, and estimated after 6 seizure-free months, the risk of a subsequent seizure within the next 12 months had dropped below 20%. In addition, data from MESS was used to develop a more detailed prognostic model allowing stratification of patient groups.

Before a predictive or prognostic model can be introduced into routine practice, it should be externally validated, i.e. tested for satisfactory performance in datasets that are fully independent of the development data. A subsequent study[99] to MESS used three external datasets of IPD to validate the prognostic model for seizure recurrence: 2 observational studies from the US and UK and a clinical trial from Italy, with a total of more than 1400 individuals. The analysis demonstrated that the prognostic model generalised relatively well, confirming its validity for predicting risk of seizure recurrence following a first seizure in people with various combinations of risk factors.

Following this external validation, the model was fitted to a pooled population comprising all three validation datasets and the development dataset. Again, the model fit well, providing support for a single, worldwide overall prognostic model for risk of second seizure following a first, which will enable driving regulations worldwide to be harmonised.

A.6.3 Comparison of efficacy and safety profile of different anti-epileptic drug therapies

Epilepsy is a neurological disease, characterised by recurrent seizures that are caused by abnormal electrical discharges in the brain. There are over 40 known forms of epilepsy, together affecting around half a million people in the UK alone.[100] The severity of symptoms varies greatly between people, and may change over time for individual patients. In most cases, the cause of the disease is unknown. If not properly controlled, epilepsy can have a major impact on a person's health and wellbeing.

Fortunately, most seizures can be controlled with anti-epileptic drug (AED) monotherapy. According to The National Society for Epilepsy in the UK, there are currently 26 different compounds with anti-epileptic activity on the market[101]. In the UK, the National Institute for Health and Care Excellence (NICE) prepares guidelines for physicians on which AED to prescribe, taking into account patient and disease characteristics. These guidelines require continuous updating to ensure that they provide the best available recommendations, as based on the most up-to-date evidence.

The evidence on the efficacy and safety profile of different AEDs originates largely from randomised clinical trials, often including trials in which the outcomes for patients receiving the drug of interest are compared to those of patients receiving a placebo. In the case of AED drug trials, however, it is more common to compare different AEDs. Data from multiple such

[99] Bonnett, LJ et al (2014) External Validation of a Prognostic Model for Seizure Recurrence Following a First Unprovoked Seizure and Implications for Driving. PLoS One 9:e99063.

[100] http://www.epilepsysociety.org.uk/what-epilepsy#.VCAiB-euMuo, accessed 22 Sep 2014.

[101] http://www.epilepsysociety.org.uk/list-anti-epileptic-drugs#.VCAghueuMuo, accessed 22 Sep 2014.

trials can be combined through meta-analysis to inform the development or revision of clinical guidelines. Three main arguments in favour of using IPD meta-analysis of comparative AED monotherapy trials[102] were presented:

1. To undertake a more complete analysis of time-to-event outcomes: The efficacy and safety (or risk of adverse events) of different AED treatments is compared using various 'time-to-event' outcomes (e.g. time to withdrawal, representing the moment when adverse events outweigh potential treatment benefits). Although methods have been developed to synthesise time-to-event data using summary information, it is unlikely that all trials fully report the necessary data for the outcomes and subgroups of interest.

2. To investigate the interaction between anti-epileptic drug and type of epilepsy: IPD meta-analysis offers an opportunity to investigate the effects of treatments in different subgroups of patients, particularly those with generalised epilepsy versus those with partial epilepsy, with higher power, due to increased patient numbers.

3. To undertake re-analysis to obtain results for all relevant outcomes: When only a subset of all outcomes is reported, this could signal a publication bias, which is likely to favour significant results. If used in an aggregate data meta-analysis without appropriate adjustment, this could lead to biased results.

Therefore, an IPD approach that can draw on the data for patients in all the trials allows a more thorough analysis of time-to-event data and treatment covariate interactions, while minimising bias. This approach was applied to a series of eight connected Cochrane Reviews of epilepsy monotherapy trials with meta-analysis (published between 2000 and 2007). In each of these trials one AED was compared against another AED.

The use of IPD in these reviews allowed standardisation of outcomes and analytical approaches across the included trials. Using IPD from around 4,000 patients and 19 separate trials, the eight reviews were able to provide important guidance on the comparative benefits and risks of different AEDs. They also informed the design of the NHS-funded SANAD trial, the largest ever trial in epilepsy patients, which compared several different AEDs in two separate trial arms.[103]

An inherent limitation of these eight reviews, however, is that they were only able to compare between pairs of two individual drugs. In practice, clinicians want to know how the drugs compare to all other drugs. To address this shortcoming, a network meta-analysis was undertaken, based on the same IPD data sets as the original 8 reviews and supplemented with data from the SANAD trial.[104] Together, the included trials investigated 8 different AEDs, so that a network meta-analysis approach allowed for 28 different pair-wise comparisons (e.g. if trial 1 compares drug A versus B, and trial 2 compares drug B versus C, the network meta-analysis approach also enables comparison between drug A versus C). The availability of IPD allowed the standardisation of outcome definitions, required for indirect comparisons across trials. Furthermore, IPD permitted assessment of outcomes for patients with either partial or generalised onset seizures, and enabled an examination of comparability of trial characteristics. The findings from this study supported the existing guidelines that recommended the use of the AED valproate as drug of first choice for generalised onset seizures.

[102] Williamson, PR et al (2000) Individual patient data meta-analysis of randomized anti-epileptic drug monotherapy trials. Journal of Evaluation in Clinical Practice, 6: 205-214.

[103] Marson, AG et al. (2007) The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. The Lancet 369: 1000-1015.

[104] Tudur-Smith, C, et al (2007) Multiple treatment comparisons of epilepsy montherapy trials. Trials 8:34.

A.6.4 Differences in surgical treatment effects across patient subgroups

Surgeons sometimes deem results from randomised clinical trials irrelevant to their practice because of concerns about generalisability of findings across patient populations. Therefore, many surgical clinical trials also include subgroup analyses to examine whether certain patients may benefit more from a specific treatment than others. However, there are two interrelated concerns regarding these analyses: failure to detect a relevant subgroup effect, and unjustified claims about subgroup effects that do not exist. Both could lead to suboptimal patient care. IPD meta-analyses of these clinical trials have the advantage that they provide increased statistical power, and the ability to examine the consistency of subgroup effects across studies.

To determine how many IPD meta-analyses on surgical interventions performed subgroup analyses, and whether the outcomes of these analyses have changed decision-making in clinical practice, a systematic review was conducted[105]. 18 relevant IPD meta-analyses were identified, looking at a variety of surgical interventions. Subgroup selection in these studies had been done mainly on patient and disease characteristics, on the basis of reports from scientific literature.

The study found that in half of the IPD meta-analyses that reported non-significant overall effects, the results became significant for at least one subgroup, corresponding to 14% of all subgroups in these studies. In addition, for the majority of meta-analyses that reported a significant overall effect, estimate results remained significant in one or more subgroups, but that for most subgroups the effect was non-significant. Together, these findings illustrate that IPD meta-analyses can reveal effects in particular subgroups that are significantly different from the effect on the overall study population.

So far, the findings from 8 of the 18 significant subgroups appear to have been translated into appropriate treatment guidelines. However, the authors caution that most of the included studies were only recently published, and that it could take years before findings get converted into guidelines or clinical practice.

[105] Hannink, G et al (2013) A systematic review of individual patient data meta-analyses on surgical interventions. Systematic Reviews 2:52.

Appendix B Comparison of existing individual participant data sharing initiatives

Status: public information available, December 2014

Database comparison table – Group 1

Category	Name of repository / data sharing network	Launch year	Focus area	Initiated by	Data hosted by	Number of clinical trials datasets	Number of individual patients	Contributing organisations	Treatment arm included?	Other types of data held
Group 1: Collaborative groups of trialists / trial sponsors	EBCTCG	1985	Early breast cancer, survival	variety of public funders and non-profit organisations	University of Oxford, UK	Approx. 700		academic and commercial	Y	
	C-Path consortia (CODR)	2005	Alzheimer's, Parkinson's, Tuberculosis, Multiple Sclerosis, Polycystic Kidney Disease, safety biomarkers, patient-reported outcome measures	FDA, membership fees, foundations	C-Path Institute, US	50 (27 AD, 7 PD, 10 TB, 6 MS, 5 PKD)		academic and commercial	y	pre-clinical data, observational studies
	WWARN	2009	Antimalarial treatments	Gates Foundation, USAID, DfID, Exxon Mobile Foundation	University of Oxford, UK	350	100,000	academic and commercial	Y	pharmacological, molecular, and *in vitro* data
	IMPACT	2003 - 2011	Traumatic brain injury	NIH- National Institute of Neurological Disorders and Stroke	U Hospital Antwerp, Erasmus U Medical Center, U of Edinburgh, Virginia Commonwealth U	12		academic (1) and commercial (11)	Y	observational studies (5); data derived from images (but no raw images)
	EORTC Headquarters	1962	Cancer	EORTC	EORTC			Academic and commercial	Treatment and placebo	

Category	Name of repository / data sharing network	Data standardised by	Access request review	Access modality	Data format	Demand	Future plans	Weblink
Group 1: Collaborative groups of trialists / trial sponsors	EBCTCG	database staff	original researcher / data provider	analysed by collaboration; transfer to user	text file			http://www.ctsu.ox.ac.uk/research/meta-trials/ebctcg/ebctcg-page
	C-Path consortia (CODR)	database staff	original researcher / data provider	shared within consortium	text file, in late 2014 will move to SAS platform			http://c-path.org/programs/
	WWARN	database staff	original researcher / data provider	shared within consortium; if approved, transfer to external user		<5 third party requests; studies have mainly been coordinated and led by WWARN researchers, working with the primary investigators	in discussion to implement the same concept for visceral leishmaniasis and parasitic helminthes	http://www.wwarn.org/partnerships/data
	IMPACT	database staff	scientific review board	via on site facilities in the participating institutes, researchers can use own software	R or SPSS compatible	only members of the collaborative group	The IMPACT project completed in 2011	http://www.tbi-impact.org/?p=impact/db
	EORTC Headquarters	database staff	study coordinator, group Leadership, and headquarters staff; company if involved	Analysis by internal staff, (preferred) or transfer to user	as needed			

Database comparison table – Group 2

Category	Name of repository / data sharing network	Launch year	Focus area	Initiated by	Data hosted by	Number of clinical trials datasets	Number of individual patients	Contributing organisations	Treatment arm included?	Other types of data held
Group 2: Disease-specific data repositories	PRO-ACT	2013	ALS / MND	Prize4Life, (US) Northeast ALS Consortium, NCRI at MGH	Neurological Clinical Research Institute (NCRI) at Massachusetts General	17	8,500	academic (7) and commercial (10)	Y (partial)	none

Category	Name of repository / data sharing network	Launch year	Focus area	Initiated by	Data hosted by	Number of clinical trials datasets	Number of individual patients	Contributing organisations	Treatment arm included?	Other types of data held
					Hospital, USA					
	C-Path AD repository (CODR)	2010	Alzheimer's Disease (AD)	C-Path Institute, USA	Cloud hosting provider	24	6,500	commercial	N	none
	Project Data Sphere*	2014	Cancer	CEO Roundtable on Cancer's Life Sciences Consortium (LSC)		14 (another 20 in planning)	9,000	academic (≥1) and commercial (≥ 8)	N	none
	Sylvia Lawry Centre for MS Research – The Human Motion Institute	2001	Multiple Sclerosis (MS)	Multiple Sclerosis International Federation, co-funding from other non-profits; research grants	Sylvia Lawry Centre	29 (2008)	26,000 (2008)	academic/non-profit (8) and commercial (20)	Y (partial)	MRI images, registries, and observational studies

Category	Name of repository / data sharing network	Data standardised by	Access request review	Access modality	Data format	Demand	Future plans	Weblink
Group 2: Disease-specific data repositories	PRO-ACT	database staff	staff reviews for fit with agreed access guidelines	transfer to user	Excel or text file	350 data transfers in 18 months	additional datasets being prepared for addition to database	https://nctu.partners.org/ProACT
	C-Path AD repository (CODR)	database staff	staff reviews for fit with agreed access guidelines	transfer to user	Text file; from late 2014 also SAS option	approx. 200 researchers in 3.5 years	launched a second data repository in Sep 2014: 3 TB trials from US CDC	http://c-path.org/programs/camd/
	Project Data Sphere*	database staff		access within SAS platform; or transfer to user	SAS, other?		additional datasets being prepared for addition to database; aim is to include data from 25,000 patients by April 2015	https://www.projectdatasphere.org/projectdatasphere/html/home.html
	Sylvia Lawry Centre	database staff	staff check requests for compatibility with SLC vision/mission and	data accessible on site; part of data (synthetic dataset)	data may be submitted in any format, but mostly	15-20 requests per year for		http://www.slcmsr.net/en/abo

Category	Name of repository / data sharing network	Data standardised by	Access request review	Access modality	Data format	Demand	Future plans	Weblink
			project feasibility	is downloadable	SAS format is received	collaborative research		ut/start.html

Database comparison table – Group 3

Category	Name of repository / data sharing network	Launch year	Focus area	Initiated by	Data hosted by	Number of clinical trials datasets	Number of individual patients	Contributing organisations	Treatment arm included?	Other types of data held
Group 3: Public-funder mandated repositories	NIH: NIDDK	2003	Endocrine and metabolic diseases such as diabetes, digestive diseases, nutritional disorders, and obesity; and kidney, urologic, and hematologic diseases	NIH-NIDDK	IMS Inc, US	58 clinical study datasets from 43 distinct clinical studies		academic (funded by NIH-NIDDK)	Y	biospecimen, genotyping data from GWAS, cohort/observational studies (58 datasets in total)
	NIH: BioLINCC	2000	Heart, blood, and lung diseases (excluding cancer)	NIH-NHLBI	IMS Inc, US	82	> 600,000 (trials and observational studies)	academic (funded by NIH-NHLBI)	Y	biospecimen, observational studies (33)
	NIH: NDCT	new platform: Aug 2014	ADHD, Alzheimer's Disease, Anxiety Disorders, Autism Spectrum Disorder, Bipolar Disorder, Depression, Pervasive Developmental Disorder, Schizophrenia, a.o.	NIH-NIMH (leverages platform developed for autism in 2007, the National Database for Autism Research)	NIH-NIMH	NIMH migrating the current (19) restricted access datasets		academic	Y	the system supports omics, clinical, imaging and neurosignal recordings data and results.
	NIH: NIDA*	2006	treatments of drug abuse	NIH-NIDA		31 (including pilot trials)		academic	Y	linked biospecimen and genetics data available

Category	Name of repository / data sharing network	Launch year	Focus area	Initiated by	Data hosted by	Number of clinical trials datasets	Number of individual patients	Contributing organisations	Treatment arm included?	Other types of data held
										from separate database
	Immune Tolerance Network – TrialShare*	Oct-12	Autoimmune disease (incl. Type 1 diabetes), allergies and asthma, organ transplantation	NIH-NIAID		35 (includes single arm trials, registries)	3,200	academic	Y	biospecimen

Category	Name of repository / data sharing network	Data standardised by	Access request review	Access modality	Data format	Demand	Future plans	Weblink
Group 3: Public-funder mandated repositories	NIH: NIDDK	user (database staff curate data)	Administrative review by extramural scientific experts	transfer to user	SAS, other formats possible	309 data requests in 9 years	Continue adding data from NIDDK-funded studies; enforce timely deposition of data; build capacity to use existing datasets	https://www.niddkrepository.org/home/
	NIH: BioLINCC	user	Administrative review by internal staff	transfer to user	same format as the data was received in	approx. 640 investigators received data. 35% of requested datasets included data from clinical trials, i.e. around 220 requests in 14 years.	Continue adding data from NHLBI-funded studies; cataloguing and indexing; marketing more widely	https://biolincc.nhlbi.nih.gov/home/
	NIH: NDCT	data provider harmonises to existing standard	NDCT Data Access Committee (NIH-NIMH Program staff)	transfer to user, or push to cloud	CSV download, or hosted database on the cloud allowing for access to rich datasets for imaging, neurosignal recordings and omics	average of 8 requests per month over the past six years (556 requests total) - for restricted datasets, which will be migrated into the NDCT	Reporting of findings, both positive and negative are expected through DOI initiated study definition (see http://ndct.nimh.nih.gov/results/)	http://ndct.nimh.nih.gov
	NIH: NIDA*	data provide		transfer to user	SAS, text file			http://datashare.nida.nih.gov

Category	Name of repository / data sharing network	Data standardised by	Access request review	Access modality	Data format	Demand	Future plans	Weblink
		r						
	Immune Tolerance Network – TrialShare*		no review	analysis on platform with provided analysis tools, or transfer to user	R	Jan 2013 – May 2014: over 600 registered public users; 159 datasets downloaded		https://www.itntrialshare.org

Database comparison table – Group 4

Category	Name of repository / data sharing network	Launch year	Focus area	Initiated by	Data hosted by	Number of clinical trials datasets	Number of individual patients	Contributing organisations	Treatment arm included?	Other types of data held
Group 4: Commercial trial portals and repositories	Clinical Study Data Request	May-13	all	GSK	trial sponsors physically hold data but user access provided via a secure workspace	1299, from study sponsors: GSK, Boehringer Ingelheim, Lilly, Roche, ViiV, Novartis, UCB, Takeda		commercial (multiple)	Y	none
	Yale University Open Data Access (YODA) Project	2013	all	Yale University and Medtronics	Medtronics data: Yale University, USA; J&J for J&J data	17 from Medtronics; 81 from Janssen/J&J		commercial (Medtronic; Jenssen/J&J)	Y	none

Category	Name of repository / data sharing network	Data standardised by	Access request review	Access modality	Data format	Demand	Future plans	Weblink
Group 4: Commercial trial portals and repositories	Clinical Study Data Request	data user	Independent Review Board	access within SAS platform with SAS or R; results can be downloaded	SAS and R readable formats	45 requests that met requirements: 36 approved (or approved with conditions); 3 rejected and revised to re-submit; 6 in process (May 2013-May 2014)	further datasets and trial data from Astellas, Bayer, and Sanofi, as well as additional datasets from other sponsors being prepared for addition to portal	https://clinicalstudydatarequest.com
	Yale University Open Data Access (YODA) Project	data user	Independent Review Board, members from Yale faculty	Medtronics: transfer to user; J&J: access within SAS platform with SAS or R; results can be downloaded	SAS	Medtronics: 20 requests in 12 months, data released to 4	Janssen: launched drug trials data in Oct 2014; data from medical devices in Jan 2015	http://yoda.yale.edu

Database comparison table – Group 5

Category	Name of repository / data sharing network	Launch year	Focus area	Initiated by	Data hosted by	Number of clinical trials datasets	Number of individual patients	Contributing organisations	Treatment arm included?	Other types of data held
Group 5: Open data sharing by individual research unit	FREEBIRD	2011	Traumatic Brain Injury / Injury and emergency research data	London School of Hygiene and Tropical Medicine, UK	London School of Hygiene and Tropical Medicine, UK	2	30,000	academic	Y	none
	International Stroke Trial (IST) database*	2011	Stroke	University of Edinburgh, UK	University of Edinburgh, UK	1	20,000	academic	Y	follow-up

Category	Name of repository / data sharing network	Data standardised by	Access request review	Access modality	Data format	Demand	Future plans	Weblink
Group 5: Open data sharing by individual research unit	FREEBIRD	Original researcher / data provider	no review for part of the data; requests for randomisation code reviewed by trial team	transfer to user	text file	Dataset has been downloaded 47 times	the two trials in FREEBIRD are also integrated into the IMPACT database	https://ctu-web.lshtm.ac.uk/freebird/
	International Stroke Trial (IST) database*	(curation: original researcher / data provider)	no review	transfer to user	text file			http://www.trialsjournal.com/content/12/1/101/

* desk research only

Appendix C Survey results

C.1 Methodology

An online survey was conducted over a 7-week period between 28 July 2014 and 8 October 2014. The survey targeted the research community globally in academic, commercial, and non-profit environments with prior experience of individual patient level data (IPD) from clinical trials. Beyond the core target group, views of clinical trial data managers, funder organisations, and medical journal editors were sought. The survey was distributed by e-mail either to relevant umbrella organisations, with the request for further dissemination via internal mail lists, or directly to relevant individuals[106]. The survey population hence represents a non-random sample of self-selecting respondents: it cannot be assumed that the views reported in this study represent the views of all relevant stakeholders.

In total, 628 survey responses were received. 75 respondents who indicated they had not been "involved in or aware of specific research using individual participant data" were filtered out, in order to ensure respondents had sufficient background and experience to answer the survey questions. We did however take note of interesting contributions in response to open questions.

Of the remaining 553 respondents that answered questions in Section 1 ("About me"), only 446 started Section 2 ("Current Uses of Individual Participant Data From Clinical Trials"), and 386 respondents started Section 3 ("Current barriers"). As the primary focus of this study is current barriers to IPD research and future perspectives on a IPD repository (survey sections 3 and 4), the analysis of respondent demographics and profile (see C.2) only takes account of the subset of 386 respondents who started Section 3. The analysis of Section 2, however, includes a larger number of respondents (up to 446).

For all questions, respondents who chose "no view" remained within the total number of respondents when calculating percentages unless stated otherwise.

In this appendix, we provide the full results of the survey in graphical format.

C.2 Demographics and respondent profile (Survey section 1)

Survey results mainly reflect the experience and views of the public (and not-for-profit) sector with over two-third of the respondents based in university, hospital and healthcare settings (Figure C 1). Private pharmaceutical companies, data analytics companies, and contract research organisations represented approximately 10% of responses (45). This group is referred to as "industry" or "companies" within this report. Two thirds (30) of respondents from companies were from large pharmaceutical enterprises. Companies may assign a single individual to complete the survey representing the position of the entire organisation (rather than one individual, which is generally the case for universities and hospital researchers). We therefore report survey results from this group separately, termed "industry" in figures and tables, as taking the average across the entire population of respondents would "drown out" the industry view. However, we cannot be certain how many of the 45 responses reflect the position of an entire company, or the personal views of individuals within a company[107].

[106] This included industry associations, non-governmental funders such as charities, governmental funders, professional societies and other relevant associations, regulators, patient groups, research and clinical trials coordination networks, individual researchers, and staff of existing data sharing initiatives and repositories.

[107] In two cases, respondents explicitly stated that they were expressing their personal opinions rather than the position of their employer.

The geographical coverage achieved is global with responses from Europe and USA representing around 90% of the total, and lower numbers of responses from other parts of the world[108]. A sample of 17 respondents from Japan (8), India (4), Thailand, Bangladesh, Mexico, South Africa and Argentina (1, each), all from non-commercial organisations, was analysed separately to gauge views of researchers based in different cultural settings and/or in middle- and low income countries and outside the established research networks of the North America, Europe, and Australia / New Zealand. This group is referred to as "other countries" respondents in this Appendix; while the respondent number in this group is small, we have included it in the analysis as an indication of potential differences.

Of respondents from companies (45), 17 were based in the UK (13 of which were from large pharmaceutical companies), 18 in other European countries (11 from large pharma), and 9 in the United States (6 from large pharma).

Around two thirds of the respondents confirmed that their main role was in delivery or management of research (both clinical and non-clinical), and the remaining respondents were distributed among roles related to research funding and dissemination.

80% of the respondents included in this analysis had direct involvement with research using IPD. This proportion was the same for respondents from companies only.

[108] The majority survey respondents were based in the UK (57%). We investigated differences between the responses of UK-based respondents versus those of respondents located outside the UK, to see how a potential over-representation may have affected the average of survey results.

While UK respondents expressed comparable views to non-UK respondents for most questions (e.g. views on current barriers to access and characteristics of a future repository were very similar), there were appreciable differences in the following two areas:

1) When asked if "the ability to access a clinical trial data repository, containing IPD from industrial and academic trials, [would] change your/your organisation's current research", 48% (75) of non-UK respondents indicated "it would significantly influence the direction of research, and it would likely lead to new research approaches and outcomes in areas of unmet need", whereas only 28% (59) of UK respondents felt this way. Conversely, 9% (14) of non-UK respondents thought "it would not change the research, but a central data access point and process would represent significant time- and cost-savings", compared to 19% (40) of UK respondents.
2) UK and non-UK respondents different in their approach to data storage and access models, with UK respondents being less in favour of an open access model, and non-UK respondents less in favour of reviewed access through the interface of trial sponsors:
    - While 35% (49) of non-UK respondents considered "open access" the most suitable model, only 17% (32) of UK respondents chose this option. A higher proportion of UK respondents felt "open access" was the least suitable model (57%; 106), compared to 38% (53) of non-UK respondents.
    - The majority of UK respondents (44%; 85) considered "reviewed access through the interface of trial sponsors" 'moderately suitable', compared to 32% of non-UK respondents (32%, 45). The majority of non-UK respondents felt this storage and access mechanism was 'least suitable' (51%, 73), compared to 33% (63) for respondents based in the UK.

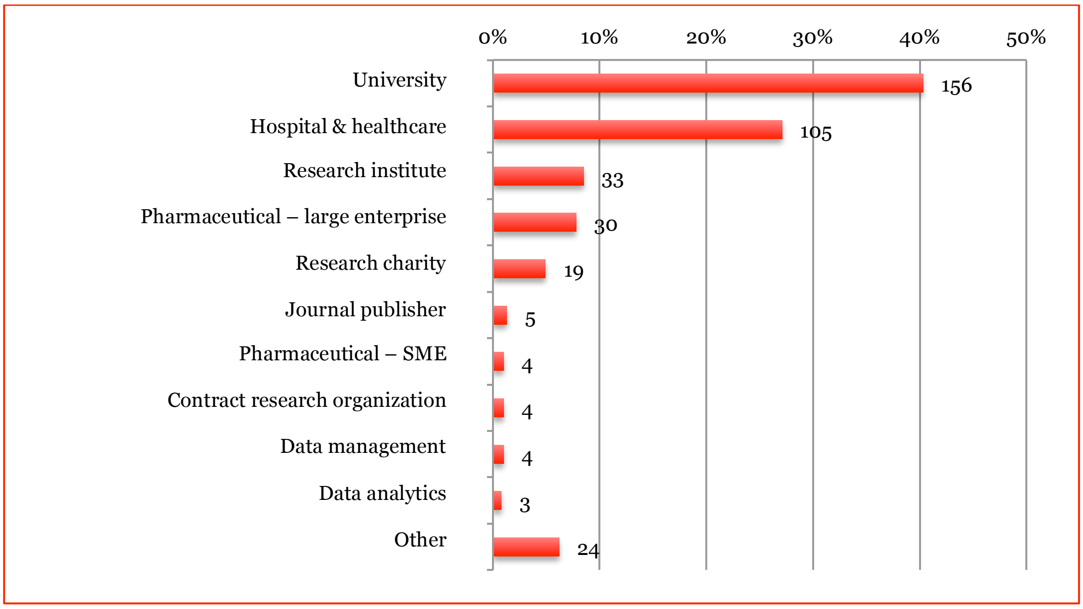
Both, UK and non-UK respondents indicated that "reviewed access through an independent data custodian" was the most suitable model, with 54% (75) and 65% (123) choosing this option, respectively.

In conclusion, the large proportion of UK survey respondents may have biased the overall findings to be:

a) less positive regarding enhanced access to data significantly influencing the direction of research and opening up new research approaches,
b) less favourable towards an open access model, and
c) more favourable towards access through the interface of trial sponsors.

Figure C 1 Respondent demographics and profile ("About me") n = 386, data labels within the chart indicate the number of respondents.

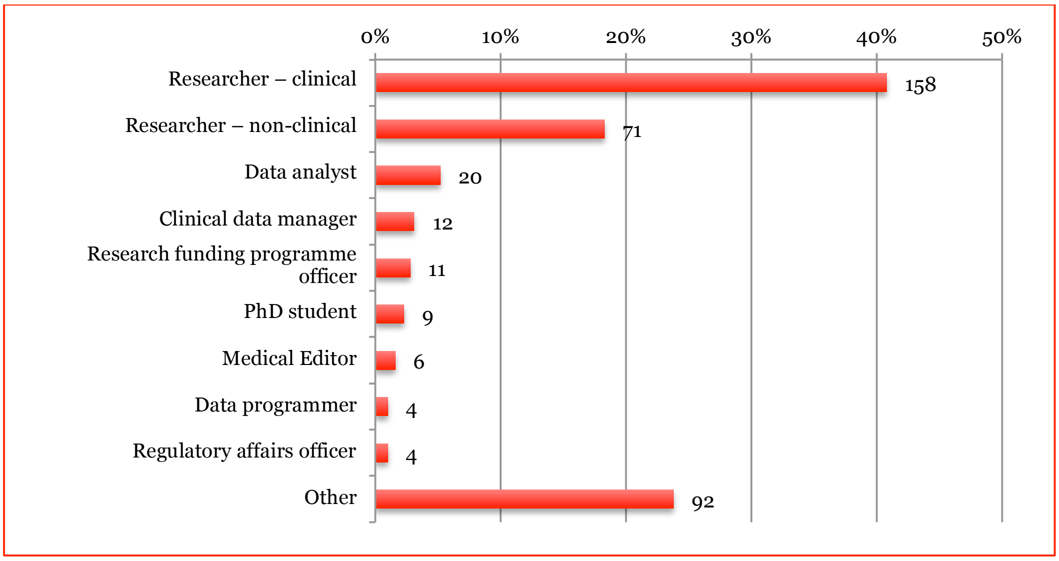
- Which of the below best describes your current employer?



- Which country are you based in?

Country	Number of respondents	Percentage
United Kingdom	219	57%
United States	93	24%
Other Europe:	48	12%
Germany	10	
Switzerland	8	
Italy	6	
Spain, Netherlands, France	4 each	
Other	12	
Rest of World:	26	7%
Japan	8	
Australia	5	
India	4	
Canada	2	
Israel, New Zealand, South Africa, Thailand, Mexico, Bangladesh, Argentina	1 each	

d) Which of the below best describes your current position?



e) Are you involved in / aware of research using individual participant data from clinical trials?

	Response count	Response percentage
I have been involved with research using individual participant data	310	80%
Colleagues within my organisation conduct research using individual participant data	128	33%
I know of research conducted by other organisations using individual participant data	100	30%

(Note that multiple responses were allowed; answers do not add up to 100%.)

C.3 Current uses of individual participant data from clinical trials (Survey section 2)

Survey respondents were asked to provide information on the research projects using IPD that they were either involved in, or aware of (Figure C 2).

Regarding the principal research objectives of these projects, most of the respondents chose comparison of the effects of different treatments (82%) and assessment of adverse events (61%). 49% indicated subgroup analysis as the principal research objective, 47% identification of new biomarkers, and 41% research to aid trial design. The order of objectives was the same for respondents from companies and from “other countries” respondents.

Projects addressed the diseases areas of cancer (54%), cardiovascular disease (36%), central nervous system or neuromuscular conditions (32%), mental health and behavioural conditions (23%), and digestive/endocrine, nutritional and metabolic diseases (23%). A higher proportion of the 42 respondents from companies indicated that they were involved in, or aware of, projects in all of these areas (e.g. cancer: 71%, cardiovascular disease: 52%, central nervous system or neuromuscular conditions: 48%). “Other countries” respondents (18) were mainly involved in or aware of projects in the area of infectious diseases (33%) and cardiovascular disease (33%).

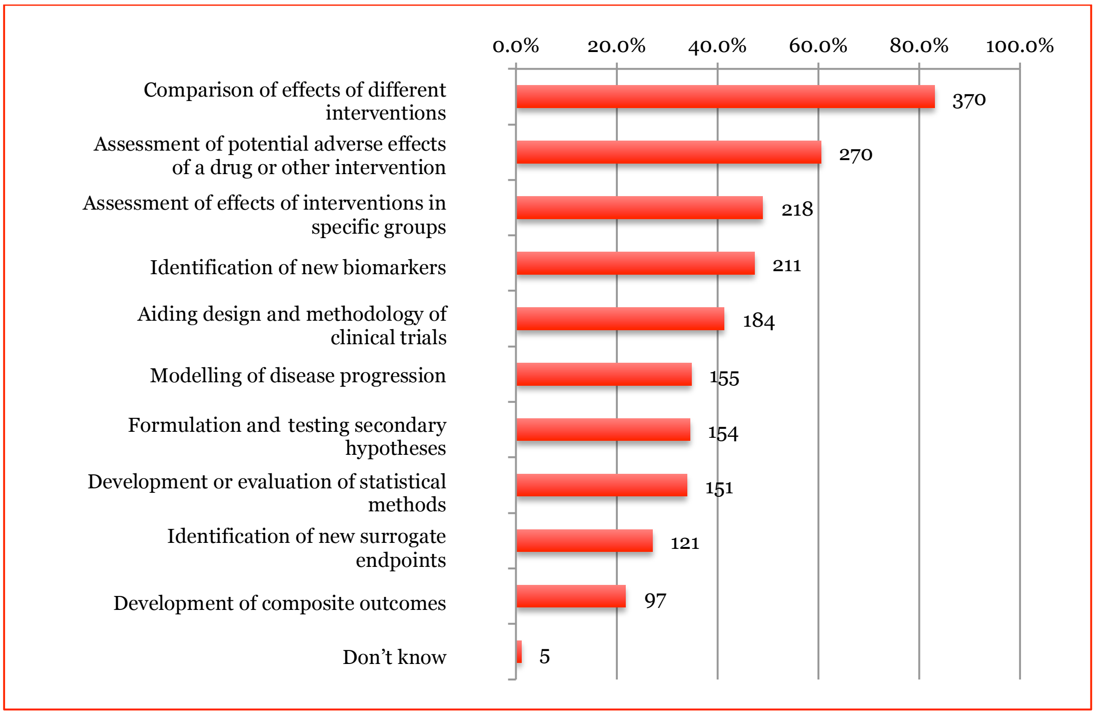
Most of the projects respondents were referring to made use of data on health outcomes (83%), demographics (78%), clinical laboratory test results (73%), medical history (71%), and adverse events (64%). 34% indicated that radiology reports and images had been used. A higher proportion of the 50 respondents from companies indicated that projects had involved data on adverse events (84%), while the remaining data types were used with comparable frequency. Responses from "other countries" were comparable to those of the overall responses.

Project researchers made use of a variety of methods and techniques to analyse individual participant data from clinical trials. Most projects carried out multivariate analysis (75%), logistic regression (51%) and univariate analysis (47%). 36% of projects involved one-stage meta-analysis, and 28% of projects two-step meta-analysis of data. Less traditional techniques were also noted, such as data mining (22%), machine learning (9%) and the use of genetic algorithms (6%), indicating a potential for the use and testing of a wide range of approaches if data can be accessed. Respondents from companies indicated similar use of methods and techniques, while respondents from "other countries" had not used machine learning and genetic algorithm techniques.

Figure C 2 Current practices in research using individual participant data

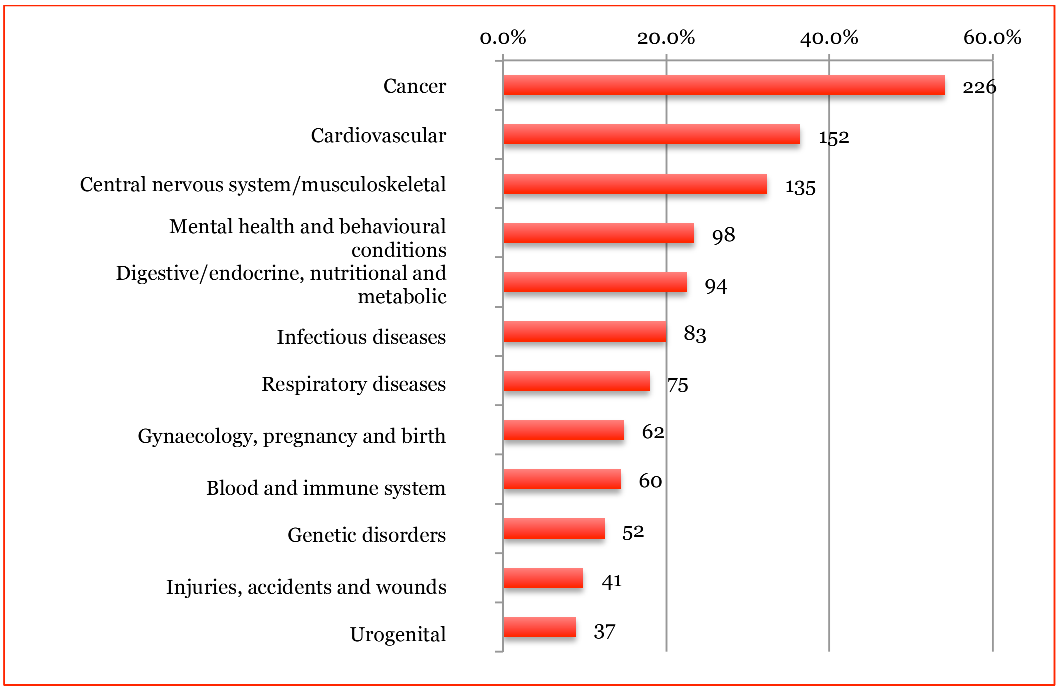
(Note that multiple responses were allowed; answers do not add up to 100%. Data labels within the chart indicate the number of respondents.)

a) Please indicate the principal objectives of the research using individual participant data you were involved in / aware of. (n=446)

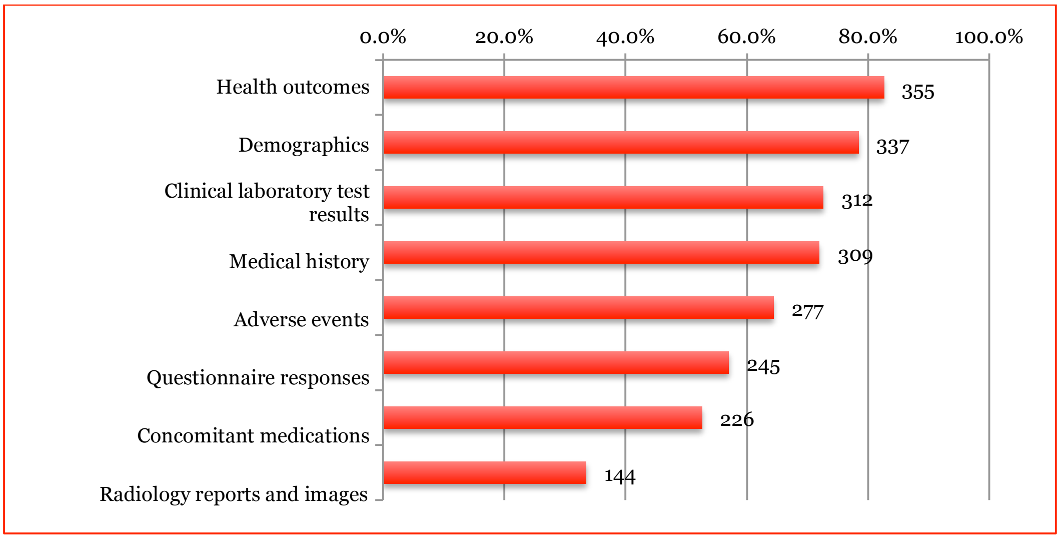


Please provide the following information related to the research you were referring to in the previous question:

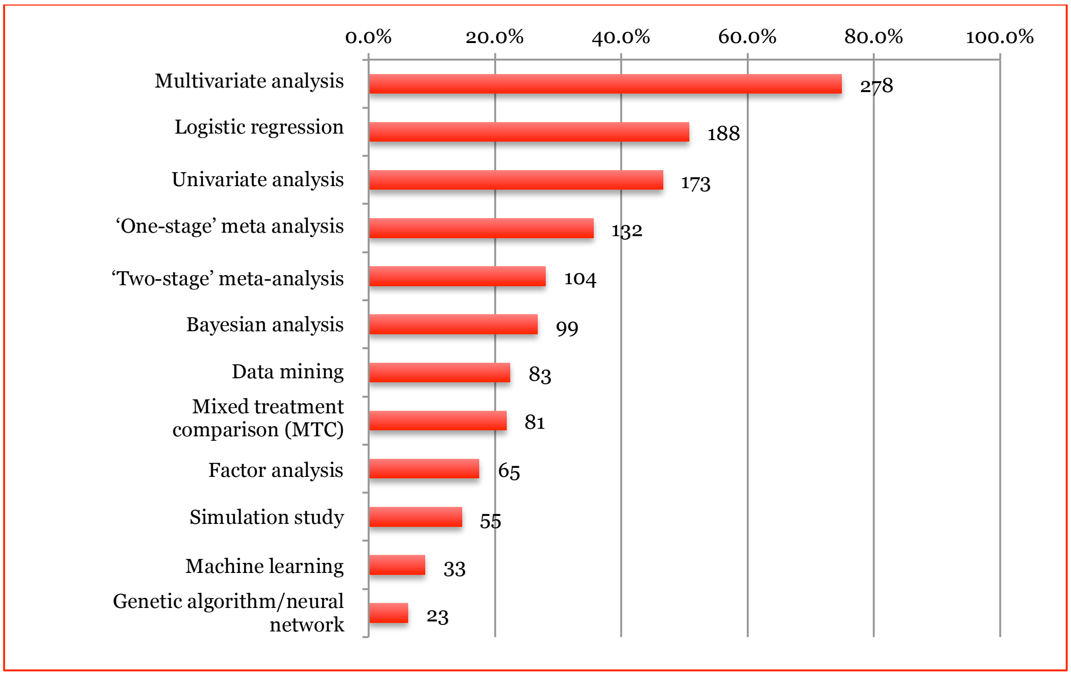
b) Disease area (n=418)



c) Type of individual participant data used (n=430)



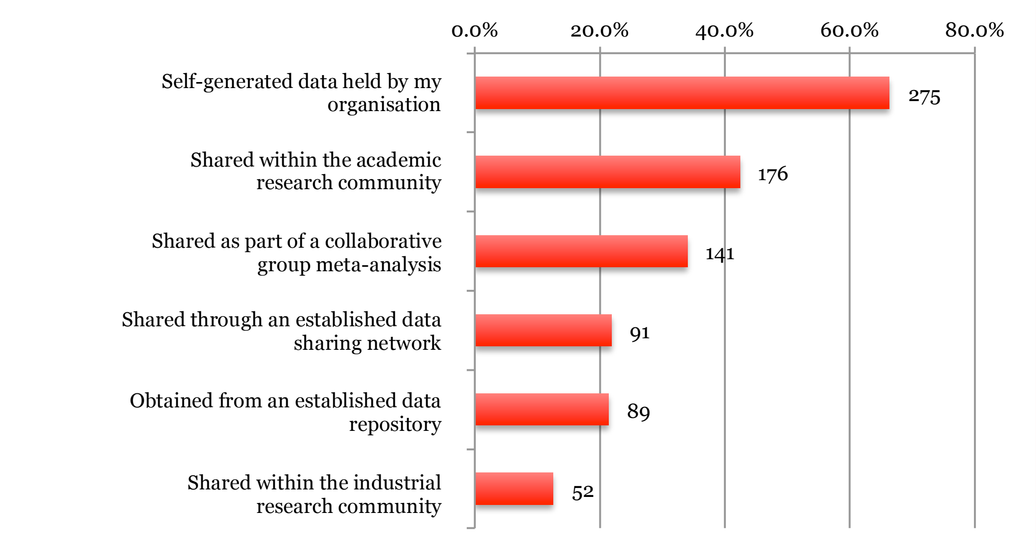
d) Analysis method used (n=371)



C.4 Source of individual participant data

The survey showed that data used for IPD analysis was predominantly generated within the organisation of the 415 respondent (66%), shared within the academic community (42%), as part of a collaborative group (34%), or shared through an established data sharing network (22%) (Figure C 3). Only 21% obtained data through an established repository, and 13% from within the industrial research community. Respondents employed by companies (48 responses for survey sections 2) indicated that data generated by their organisation was the primary source for IPD analysis (79%), followed by data shared by the academic research community (29%). 17% indicated they had obtained data through sharing within an established sharing network, and as part of a collaborative meta-analysis group, each. 15% had shared through an established repository, and 13% through networks within the industrial research community. "Other countries" respondents (19) showed a similar distribution as overall survey respondents, with data used for IPD analysis predominantly generated within their organisations (63%). A smaller proportion of projects had involved sharing within the academic community (21%).

Figure C 3 Source of individual participant data



n = 415; data labels within the chart indicate the number of respondents; multiple responses were allowed; and hence answers do not add up to 100%

C.5 Current barriers (Survey section 3)

The survey explored the potential challenges surrounding research projects involving IPD, asking respondents to indicate the extent to which a range of potential barriers impacted on researchers conducting projects involving individual participant data (Figure C 4). Answers were converted to numerical values and barriers ranked by average "impact score" (Table C 1).

On average, respondents gave the barrier of accessing relevant existing data and incomplete knowledge of what data currently exists the highest rating (2.8, and 2.4 respectively, between 'significantly' and 'moderately important'). Industry respondents were either less concerned than or had comparable views to the overall respondent population for most barriers, except for "concerns about providing competitive advantage", which was rated higher on average. The "Other countries" respondents tended to be more concerned about most barriers than the overall survey population, with the larges difference for "concerns about providing competitive advantage".

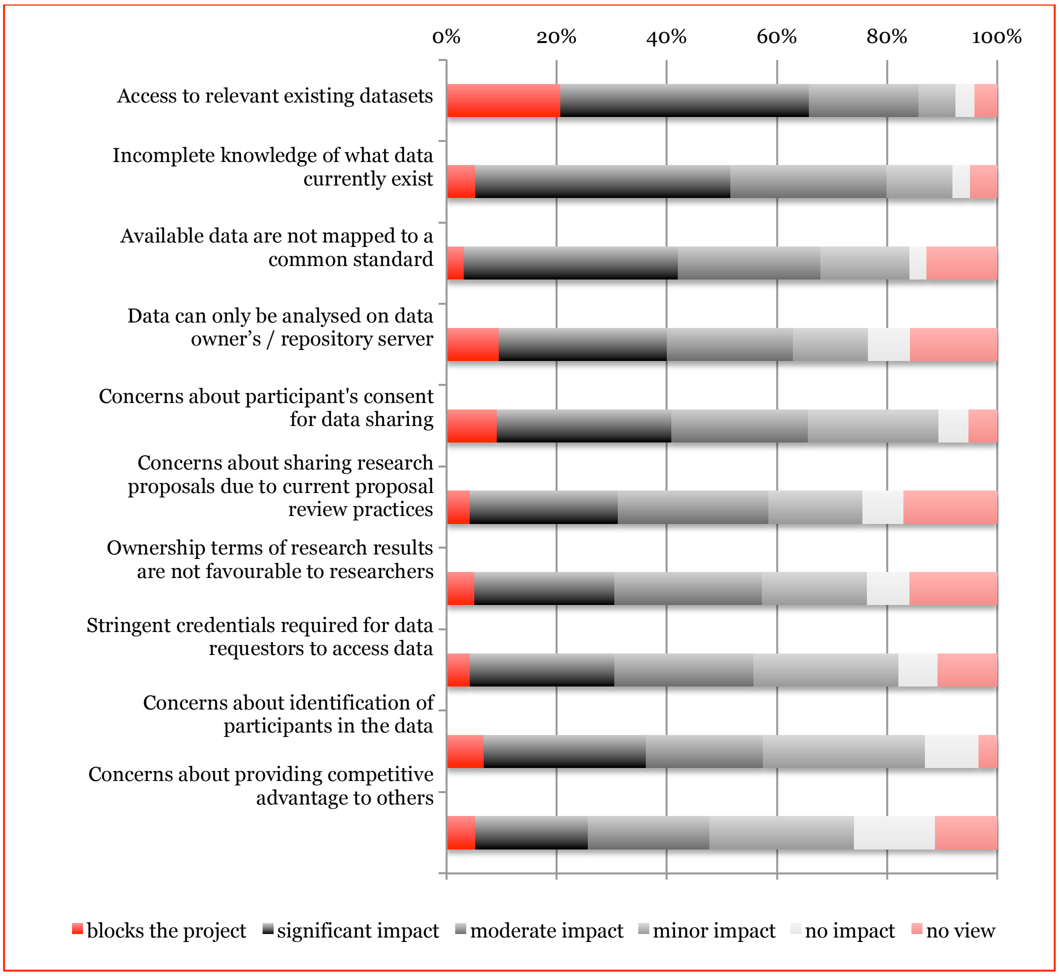
Table C 1 Potential current barriers to research involving individual participant data

Rank	**Answer option**	**Impact score:** all	**Impact score:** industry	**Difference** all -industry	**Impact score:** other countries	**Difference** all – other countries
1	Access to relevant existing datasets	2.8 (n=370)	2.6 (n=43)	0.2	2.5 (n=16)	0.3
2	Incomplete knowledge of what data currently exist	2.4 (n=365)	2.3 (n=42)	0.1	2.2 (n=16)	0.2
3	Available data are not mapped to a common standard	2.3 (n=334)	2.2 (n=43)	0.1	2.1 (n=16)	0.2
4 - 5	Data can only be analysed on data owner's / repository server	2.2 (n=320)	1.9 (n=40)	0.3	2.5 (n=14)	-0.3

Rank	Answer option	Impact score: all	Impact score: industry	Difference all -industry	Impact score: other countries	Difference all – other countries
	Concerns about participant's consent for data sharing	2.2 (n=364)	2.4 (n=43)	-0.2	2.1 (n=15)	0.1
6 - 7	Concerns about sharing research proposals due to current proposal review practices	2.0 (n=312)	1.6 (n=33)	0.4	2.3 (n=16)	-0.3
	Ownership terms of research results are not favourable to researchers	2.0 (n=317)	1.7 (n=35)	0.3	2.1 (n=14)	-0.1
8 - 9	Stringent credentials required for data requestors to access data	1.9 (n=339)	1.6 (n=40)	0.3	2.2 (n=16)	-0.3
	Concerns about identification of participants in the data	1.9 (n=368)	2.2 (n=42)	-0.3	2.1 (n=16)	-0.2
10	Concerns about providing competitive advantage to others	1.7 (n=338)	2.2 (n=40)	**-0.5**	2.2 (n=15)	**-0.5**

Survey question: “Based on your experience, please rate the extent to which the following current barriers have an impact on researchers conducting projects involving individual participant data” Answers were converted into numerical values, assigning the value zero to ‘no impact’, one to ‘minor impact’, two to ‘moderate impact’, three to ‘significant impact’, and four to ‘blocks project’. The values were multiplied by the number of responses, added up and divided by the total number of responses. ‘No view’ responses were not included. Resulting scores were subtracted from each other to obtain the difference in average rating.

Figure C 4 Current barriers to research involving individual participant data



n range: 375 – 385

Rank 1 and 2: The survey showed that respondents considered the impact of "incomplete knowledge of what data currently exists" and "access to relevant existing datasets" to be the largest barriers to current research using IPD. 53% of 383 respondents felt that incomplete knowledge had a 'significant impact' on research projects or completely 'blocked' them, compared to 15% of respondents who thought it had 'little' or 'no impact'. 66% of 385 respondents felt that current access to relevant data blocked or had 'significant impact' on research, compared to 11% who felt this issue was of 'little' or 'no importance'. The responses from industry representatives were comparable to these overall findings.

Rank 3: The lack of harmonisation ("available data are not mapped to a common standard") emerged as the barrier with the third highest impact score, but responses showed that views on this issue were divided. While the largest group of the 382 respondents (39%) felt that lack of a common data standard had a 'significant' effect, 26% indicated that it a had 'moderate impact', and 16% attributed 'minor impact'. Responses from industry representatives were comparable to these overall findings.

Rank 4 and 5: The survey investigated if the inability to download data ("data can only be analysed on data owner's/repository server") presented a barrier to current researchers. The 379 respondents had divergent views on this issue, with the largest group (31%) indicating that this inability had a 'significant impact' on current research. 23% felt it had a 'moderate impact', 14% and 8% that it had a 'minor' or 'no impact', respectively - but 10% felt it 'blocked' current research projects. Industry respondents tended to be less concerned about this issue: 25% of 44 respondents chose or 'minor impact', and 16% felt this issue had 'no impact'.

Most of the 383 survey respondents (32%) indicated that "concerns about participant's consent for data sharing" had a 'significant impact' on current research projects. However, 25% felt it had a 'moderate' or 'low impact', each, while 9% thought that it 'blocked projects'. Of the 44 respondents from companies, 34% indicated concerns about patient consent had a 'significant impact', 25% that it had a 'moderate impact', 16% each that it had a 'minor impact' or 'blocked projects', and 7% that it had 'no impact'.

Rank 6 and 7: Most survey respondents (342 responses) indicated that concerns about sharing research proposals due to current proposal review practices either had 'moderate' (28%) or 'significant impact' (27%) on researchers conducting projects involving IPD. 24% thought it had a 'minor' or 'no impact', whereas 4% indicated it 'blocked projects'. The 38 industry respondents tended to be less concerned this issue, with 34% indicating this had 'little' or 'no impact', whereas 37% considered review practices to have a 'moderate' or 'significant impact'.

Ownership terms of results derived from research using clinical trial data may be a potentially significant barrier to starting or completing a research project. Over half of the 376 survey respondents felt that ownership issues have a 'moderate' (27%) to 'significant' (26%) impact on the research project, with 20% indicating it had a 'minor impact'. Views of the 44 respondents from industry were comparable.

Rank 8 and 9: Survey respondents held divergent views on the impact of "stringent credentials required for data requestors to access data" on current research using IPD. Similar numbers of 379 respondents felt that this had 'minor impact' (27%), 'moderate impact' (25%) or 'significant impact' (27%). 7% thought it was of 'no importance', whereas 4% indicated it 'blocked projects'. The 43 respondents from industry tended to assign less impact to the current stringent requirements of credentials for access: half felt it had a 'minor' (40%) or 'no impact' (16%), while a total of 26% thought had a 'significant impact' or 'blocked' research projects.

Survey respondents expressed a wide spread of views on the impact of "concerns about identification of participants in the data" on current research using IPD. An equal number of the 380 respondents felt that this had a 'significant impact' (30%) or a 'minor impact' on research (30%). 10% thought it was of 'no importance', whereas 7% indicated it 'blocked projects'. A slightly higher proportion of the 42 respondents employed by companies were concerned about the impact of potential patient identification: 36% indicated this issue had a 'significant impact', and 14% felt it 'blocked' research projects; while 26% indicated that the impact was 'minor'.

Rank 10: Survey respondents expressed a spread of views on the impact of "concerns about providing competitive advantage to others" on current research using IPD. Similar numbers of the 380 respondents felt that this had 'minor impact' (27%), 'moderate impact' (22%) or 'significant impact' (21%). 15% thought it was of 'no importance', whereas 6% indicated it 'blocked projects'. Note: These views may refer to both, commercial competitive advantage, and academic competitive advantage (e.g. contributing to other research groups' successes without benefit to those who did the original research). The 44 respondents employed by companies were more concerned about the impact of the threat of competitive advantage: 36% indicated this issue had 'significant impact', and 7% felt it 'blocked' research projects. Still, 23% indicated that the impact was (currently) 'minor'.

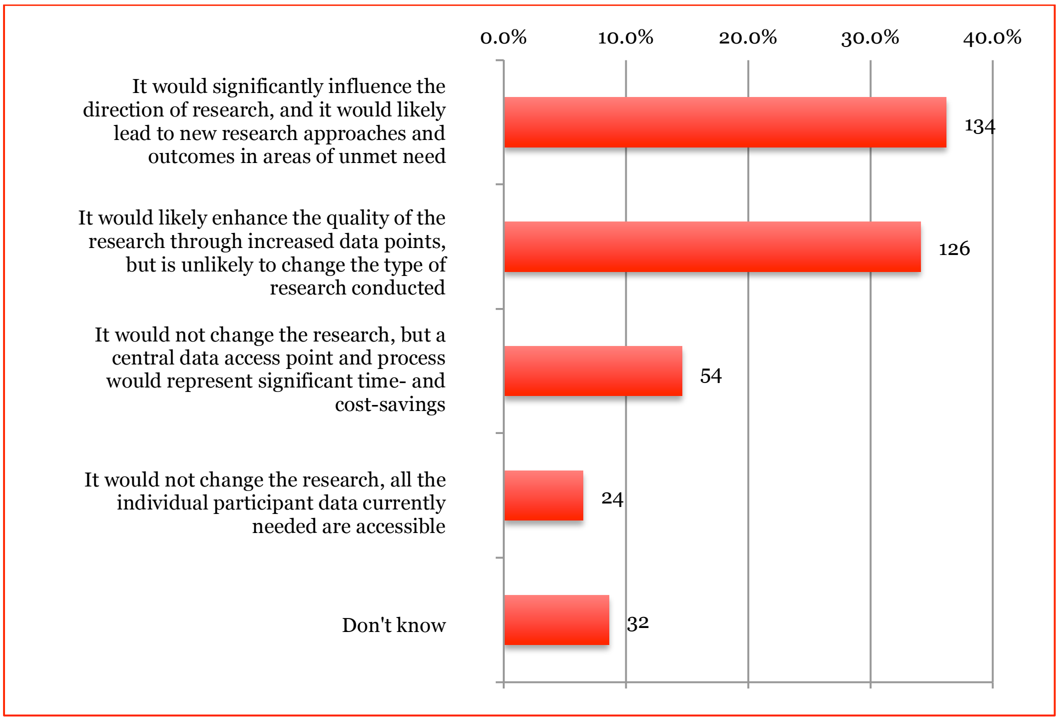
C.6 Future perspectives (Survey section 4)

The survey asked respondents to indicate how the ability to access a clinical trial data repository, containing IPD from industrial and academic trials, might change their, or their organisation's, current research (Figure C 5). Overall, most (69%) of the 375 respondents thought this would enhance the quality, or even influence the direction of research and views of the subset of respondents (42 responses) from companies provided a broadly similar picture (73%). Survey responses are explained in more detail, below with the proportion for the subset for responses from industry given in parentheses.

- 36% (29%) of respondents believed that the ability to access a clinical trial data repository, containing individual participant data from industrial and academic trials, would significantly influence the direction of research, and likely lead to new research approaches and outcomes in areas of unmet need.
- 34% (44%) thought it would likely enhance the quality of the research through increased data points, but was unlikely to change the type of research conducted.
- 14% (10%) believed that enhanced access would not change the research, but a central data access point and process would represent significant time- and cost-savings.
- 7% (15%) of respondent thought it would not change the research, as all the individual participant data currently needed were accessible.
- 9% (5%) of respondents indicated that they did not know how a repository might change research.

Of the 16 "other countries" respondents, 63% believed that the ability to access IPD would significantly influence the direction of research.

Figure C 5 Potential impact of a central individual participant data repository on research



Survey question: "How would the ability to access a clinical trial data repository, containing individual participant data from industrial and academic trials, change your / your organisation's current research? (n = 375)"

Respondents provided a range of views when explaining their thoughts on different access models (in answer to open-ended questions throughout the survey). A selection of these is presented in Box C 1 categorised by main message.

Box C 1 Survey respondents' views on access mechanisms ("Please briefly outline the main reasons for your responses above.")

1) Open access encourages use

- "Subject to necessary protection of individuals' data through robust anonymisation, the best access to data is open access. If I come up with a new idea but do not know whether sufficient data exist to tackle my question, I would like the ability to interrogate an aggregated data set quickly and without hindrance to decide whether the idea can be studied with existing data."
- "As long as the data are not identifiable, then open access is appropriate and most likely to generate the largest amount of research, including by students and researchers in their "spare" time."
- "Since the quality of the analyses carried on these dataset is hard to control anyway it does not make sense to be restrictive as long as data protection and ethical requirements are met."
- "Widely open data availability is the best way to make significant progress with the least delays."

2) Review introduces bias and delays

- "Only open access will help against bias."
- "History shows clearly that setting up rules and committees are bound to lead to arbitrary decisions, agony, abuse of power for monetary gains (by stopping unwelcome research that can threaten sales of drugs) and loads of unhappiness. Science thrives best without administrators and custodians and open access will benefit our patients most. Isn't this what it is all about? Or should be about? Isn't this why we became doctors?"
- "Reviewers or custodians can block or delay seriously the project. Data should be as easily available as possible."

3) Open access carries risk of rogue analysis

- "Totally open access is likely to lead to a flood of poor quality, sensational analyses, based on little understanding of statistics and probability."
- "Data on open access servers analysed inappropriately by researchers with ulterior motives or without adequate expertise in correct methodology can lead to the generation of misleading outcomes."
- "A review process is likely to be required to ensure appropriate use of data. However, it is important for this to be light touch and focused around risk."

4) Review ensures a sound scientific approach

- "Investigators should have a good reason, with a reasonable hypothesis and methodology, with good oversight before access is granted."
- "Open access leads to lack of hypotheses and likely Type II and I errors through multiple testing."
- "Reviewed access […] is most appropriate as it can be judged if the data is suitable for the project. This may be harder to do if data is completely open access and could result in irrelevant data being analysed."
- "I believe total open access will harm the level of the research and scare data owners away from making the data available."

5) Ethical considerations

- "[I] think it would be hard to justify to patients and ethics committees unregulated access to patient data."
- "An independent data custodian can provide appropriate safeguarding of data and ensure that any proposed research complies with ethical requirements and with relevant guidelines where necessary.“

C.7 Data access model

Our survey asked respondents to rank the suitability of different access models for their, or their organisation's, future research needs. We simplified the available models to three that combined data storage and approval process options (Figure C 6):

“open access”: data are downloadable from a central repository for any user with no or a minimum set of criteria assessed by an independent data custodian (n = 330 “all”; 37 “industry”; 14 “other countries”)

“reviewed access through independent data custodian”: data remain in a central repository secured by a trusted third party and access is granted by an independent scientific review board (n = 330 “all”; 41 “industry”; 9 “other countries”)

“reviewed access through interface of trial sponsors”: data remain with trial sponsors and access is via a specific interface granted by an independent scientific review board (n = 339 “all”; 42 “industry”; 13 “other countries”)

Respondents provided the following answers:

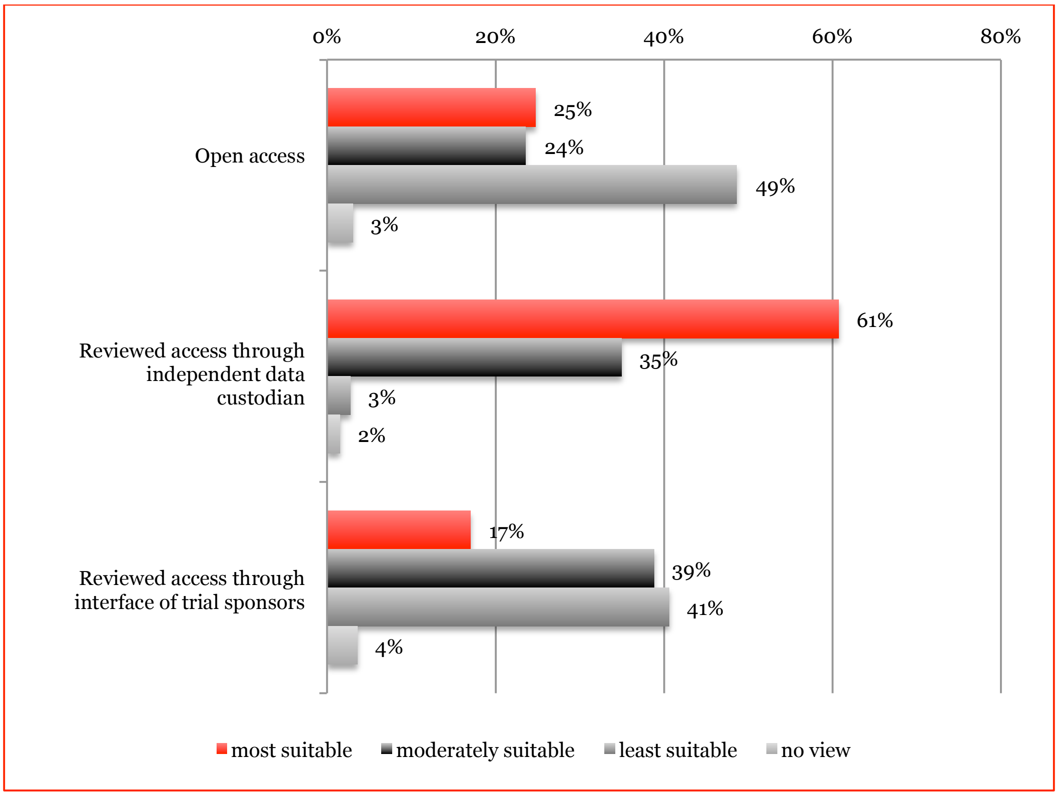
- 78% considered reviewed access ‘most suitable’,
- 61% thought it would be most suitable if data were stored in a central repository secured by a trusted, independent data custodian, and only 3% thought that this was the least suitable option,
- 25% considered “open access” to be ‘most suitable’, while 49% thought this access model ‘least suitable’.

Responses from individuals currently employed by companies, and “other countries” respondents presented a different picture:

- 91% of industry respondents chose reviewed access as the most suitable approach, while only 49% of “other countries” respondents shared this view.
- 68% of industry respondents indicated a central repository secured by a trusted, independent data custodian was ‘most suitable’, compared to only 33% of “other countries” respondents.
- 78% of industry respondents believed open access to be the ‘least suitable’, and 8% (3 respondents) thought this was the ‘most suitable’ access mechanism. Among “other countries” respondents, equal numbers felt that open access was ‘most suitable’, ‘moderately suitable’, and ‘least suitable’ (29% each).

17% of 339 survey respondents indicated that they considered access to data through the interface of the trial sponsor ‘most suitable’, while around 40% considered this approach ‘moderately suitable’ or ‘least suitable’, each. Respondents from companies were less concerned about this access model, with 23% considering it ‘most suitable’, 59% ‘moderately suitable’, and only 17% ‘least suitable’. The “other countries” respondents tended to be less in favour of an access model through the interface of trial sponsors: 58% considered it ‘least suitable’, 23% ‘moderately suitable’, and 15% ‘most suitable’.

Figure C 6 Access and data storage models



Survey question: "Please rank the suitability of the following data storage and access models for your / your organisation's future research needs." (n range 326-335)

Table C 2 provides a summary of arguments that survey respondents and interviewees made, responding to the request "Please briefly outline the main reasons for your responses above", in favour of and against three potential access mechanisms: open access (no review), access via review by an independent custodian, and access via review by the data owner.

Table C 2 Rationale for and against different data access models

Open access, no review	**Review by independent custodian**	**Review by data owner**
Easy exploration of data possible, likely to encourage use	Delays use of data, difficult to get to point of analysis – hence researcher may not attempt	Delays use of data, difficult to get to point of analysis – hence researcher may not attempt
Ensures no bias regarding access	Carries some risk of bias	May carry higher risk of bias (e.g. conflicts of interest with data provider)
Allows access to patients and 'new' researchers, including "spare time" analysts such as students	Conditions may be too stringent, limits researcher / patient access	Conditions may be too stringent, limits researcher / patient access

Open access, no review	Review by independent custodian	Review by data owner
High risk of rogue analysis due to malicious intent or incompetence	Controls risk of rogue analysis to some degree by monitoring qualifications of data requestors; ensures that the data are used to answer a scientific question, and that a properly formulated hypothesis is in place	Controls risk of rogue analysis to some degree by monitoring qualifications of data requestors; ensures that the data are used to answer a scientific question, and that a properly formulated hypothesis is in place
High risk of rogue analysis due to failure to understand data, unless direct contact with original researcher established	High risk of rogue analysis due to failure to understand data, unless direct contact with original researcher established	Controls risk of rogue analysis to higher degree as direct interaction with original researcher is required
Data may be used for research for which the appropriate patient consent is not in place	Ensures that data are used in a manner that is covered by patient consent	Ensures that data are used in a manner that is covered by patient consent

C.8 Current and future demand

When survey respondent were asked about the number of data request they made over the past year, 43% of the 228 respondents indicated that they did not make any data requests; 19% had submitted just one data request, 25% had made between 2 and 5 requests, and 13% had made more than 5 requests (Table C 3). Survey responses from companies (26 responses) showed that 65% made no data requests in the past year (possibly because they were using internal data, i.e. did not have to submit requests), 20% had made between 1 and 5 requests for data, and 16% had submitted more than 5 requests. Survey responses from the "other countries" group (12 responses) indicated that 25% had not made any data requests, and 75% had made between 1 and 5 requests.

Respondents indicated an increase in demand for IPD if a suitable access model were made available. While 43% had indicated that they had not made any requests in the last year, only 14% thought they would not make any data requests over the next year should a new repository become available. Similarly, while 57% of respondents indicated that they had requested data in the past year, this figure increased to 81% for the coming year (57% 1-5 requests, 24% 6 or more requests). Respondents from industry (26 responses) signalled a similar shift in the number of requests: 65% indicated they had not requested data over the last year, with this figure dropping to 23% for the next year if a repository become available. The proportion of those requesting data one or more times increased from 35% last year to 77% for the next year. All respondents in the "other countries" group (13 responses) were planning to make requests.

Table C 3 Current number of requests for individual participant data, and potential future demand

Estimated number of data requests per year	0	1	2-5	6-10	10<	Response count (n)
Last year with current access model	**43% (97)**	19% (44)	25% (57)	3% (6)	11% (24)	228
Next year with a potential new repository	14% (32)	17% (39)	**45% (102)**	10% (23)	14% (31)	227

Survey question: "How many data requests do you think you would make in the next year to conduct new research projects if individual participant data from commercial and academic trials were made available through the most suitable data access model? Please also indicate the estimated number of requests you made in the past year to conduct research using IPD."

C.9 Preferred characteristics of a future repository

The survey explored the importance of a number of characteristics of a potential future data repository (Figure C 7). Answers were converted to numerical values and barriers ranked by average "impact score" (Table C 4).

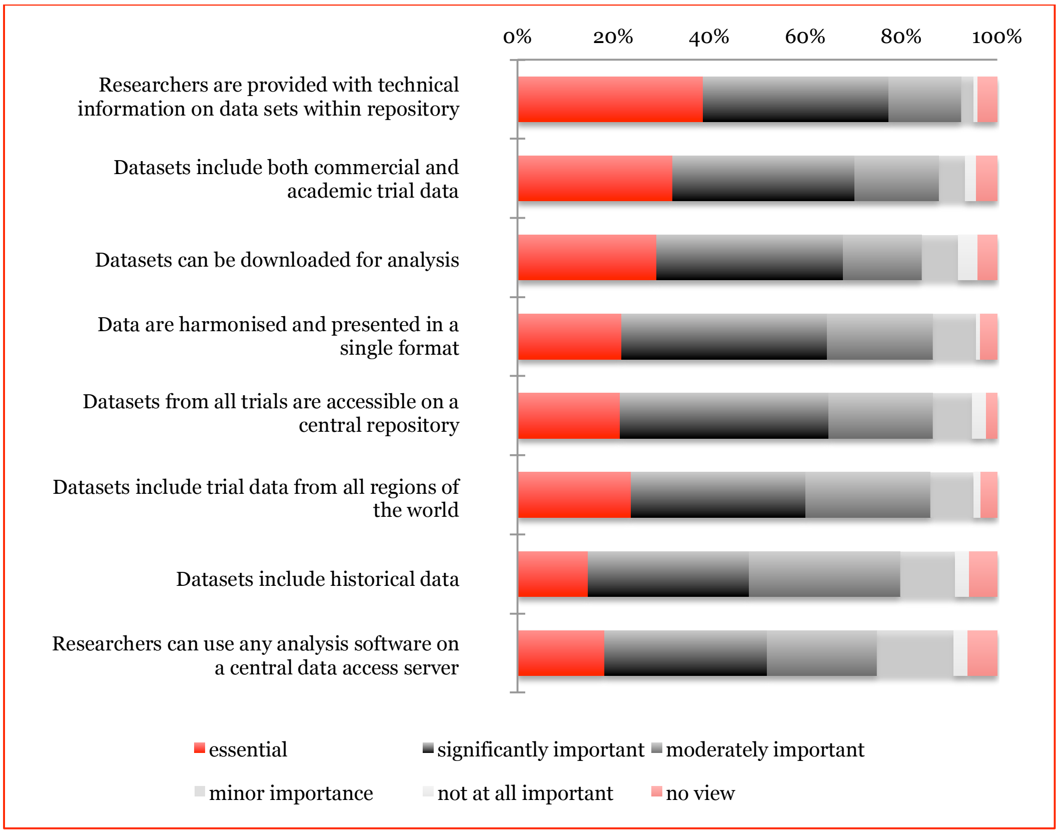
On average, respondents felt it was most important a future data access model would provide the researcher with technical information in relation to trials / data sets within the repository (score of 3.2, above 'significantly important'). Respondents also rated highly that a future repository include both commercial and academic trial data, that datasets could be downloaded for analysis, and that data was harmonised and presented in a single format (scores of 3.0 and 2.8). Respondents were least concerned about the inclusion of historical data in the repository, and the ability to analyse data with any software – but with a score of 2.5 (between 'moderately' and 'significantly important'), even these achieved a high rating. Industry respondents assigned less importance to all characteristics, with the largest difference in the ranking of the ability to download data for analysis (1.0 difference in "importance score"). Lower ranked were also the inclusion of both academic and commercial datasets, and historical data, and access of data via a central server, with the ability to use any software (0.4 difference, each). "Other countries" respondents tended to rate most characteristics more highly (compared to all responses), with all averages achieving a score of 2.9 or higher.

Table C 4 Preferred characteristics of a future data access model

Rank	**Answer option**	**Imp. score** all	**Imp. score** industry	**Difference** all - industry	**Imp. score** other countries	**Difference** all – other countries
1	Researchers are provided with technical information in relation to trials / data sets within the repository	3.2 (n=329)	2.9 (n=37)	0.3	3.2 (n=16)	0.0
2	Datasets include both commercial and academic trial data	3.0 (n=328)	2.6 (n=38)	0.4	2.8 (n=16)	0.2
3 - 4	Datasets can be downloaded for analysis	2.8 (n=331)	1.8 (n=37)	**1.0**	3.3 (n=16)	**-0.5**
	Data are harmonised and presented in a single format	2.8 (n=320)	2.6 (n=37)	0.2	3.3 (n=16)	**-0.5**
5 - 6	Datasets from all trials are accessible on a central repository	2.7 (n=333)	2.3 (n=38)	0.4	3.2 (n=16)	**-0.5**
	Datasets include trial data from all regions of the world	2.7 (n=321)	2.7 (n=37)	0.0	2.9 (n=16)	-0.2
7 - 8	Datasets include historical data	2.5 (n=330)	2.1 (n=38)	0.4	2.9 (n=15)	-0.4
	Researchers can use any analysis software on a central data access server	2.5 (n=328)	2.1 (n=39)	0.4	2.9 (n=16)	-0.4

Survey question: "Please rate the importance of the following statements relating to the characteristics of a future data repository for the type of research you / your colleagues may want to conduct. "Answers were converted into numerical values, assigning the value zero to 'not at all important', one to 'minor importance', two to 'moderately important', three to 'significantly important', and four to 'essential'. The values were multiplied by the number of responses, added up and divided by the total number of responses. 'No view' responses were not included. Resulting scores were subtracted from each other to obtain the difference in average rating. Imp. = Importance

Figure C 7 Key characteristics of a future data access model



n range: 344 – 347

Rank 1: When asked to rate how important it was that a future repository provide researchers with "technical information in relation to trials / data sets within the repository", 39% of the 347 respondents chose 'essential' or 'significantly important' each. Only a total of 4% felt this was of 'little' or 'no importance'.

Rank 2: The survey results indicate that most of the 346 respondents felt a future repository should hold both academic and commercial trial datasets. 71% of respondents considered this to be 'significantly important' (38%) or 'essential' (33%). 17% thought it was 'moderately important' while a total of 8% gave it a 'minor' or 'no importance' rating. The 39 respondents from companies attributed slightly less importance to combining these data in a central repository, with 54% giving a 'significantly important' or 'essential' rating. 31% thought it 'moderately important', whereas a total of 13% felt it was of 'minor' or 'no importance'.

Ranks 3 and 4:

Two-thirds of 342 survey respondents felt it was 'significantly important' (39%) or 'essential' (29%) to be able to download datasets for analysis. 17% thought this ability was 'moderately important', and 8% indicated 'little' and 4% 'no importance'. As was the case for current perceived barriers ("data can only be analysed on data owner's / repository server"), a smaller proportion of the 40 respondents from companies were concerned about this aspect of a future repository: only one third thought it was of 'significant importance' (20%) or 'essential' (13%), whereas 25% felt this was of 'no importance,' and 15% attributed 'little importance'. This was the question for which answers from industry differed most from those of the entire survey population.

The survey also asked about the importance of making data available after harmonisation to a common format. 65% of 344 respondents indicated that harmonisation was of 'significant importance' or 'essential', while a total of 10% considered it to be of 'little' or 'no importance'. The breakdown was similar in 39 responses from companies (60% and 18%, respectively).

Ranks 5 and 6:

When asked to rate the importance of being able to access datasets from all trials on a central repository, the majority of respondents indicated that this was 'significantly important' (43%), followed by 'moderately important' (22%) and 'essential' (21%). Of the respondents from companies, a much smaller proportion (5%) felt a central access point was 'essential' (compared to 21% for all respondents). 44% indicated it was 'significantly important', 36% considered central access 'moderately important' and 10% felt it was of 'minor importance', and 5% it was 'not at all important'.

The survey results indicate that most of the 346 respondents considered the inclusion of data from all regions in a future repository to be 'significantly important' (36%). 26% felt it was 'moderately important' to include these data, 24% deemed it 'essential', and 10% thought it of 'minor importance' or 'no importance'. The results were broadly similar for survey respondents from companies only.

Ranks 7 and 8:

The survey results indicate that most of the 343 respondents considered the inclusion of historical data in a future repository to be 'significantly' or 'moderately important' (66%, 33% each). 15% felt it was 'essential' to include these data, while 15% thought it of 'minor importance' or 'no importance'. Respondents from companies attributed slightly less importance to the inclusion of historical data compared to the entire population of survey respondents, with 26% indicating that it was of 'minor' or 'no importance', and only 8% indicating that this was 'essential'.

When asked to rate the importance for researchers to be able to use any analysis software on a central data access server, the majority of the 342 respondents indicated that this was 'significantly important' (34%), with 23% each choosing 'moderately important' and 18% 'essential'. 16% indicated that this was of 'minor importance' and 3% that it was 'not at all important'. The 39 respondents from companies assigned less importance to this characteristic, with a smaller proportion (10%) indicating a central access point was 'essential' (compared to 18% for all respondents), 23% indicating it was of 'minor importance' (compared to 16%), and 10% it was 'not at all important' (compared to 3%). 33% indicated it was 'significantly important', and 23% considered central access 'moderately important'.

C.10 Main concerns about data deposition

The survey asked respondents to describe "the one thing that you believe would **impede researchers' willingness to deposit data in a clinical trial data repository**". Of the 220 respondents, 8 cited more than one issue. The resulting 228 responses were analysed by grouping them into overarching categories (Table C 5). 27 responses were from companies. There were only 7 responses from "other countries", all touching on similar issues. Due to the low number, these were not analysed separately.

40% of survey responses (92 of 228) related to researchers' fear of losing control over how the data would be used. Within this group, 30% specifically mentioned risks to data protection and patient privacy (27 of 92), 16% the risk of misinterpretation or deliberate misuse of data (15 of 92), and 9% for each, potential lack of appropriate patient consent for secondary analysis (8 of 92), and fear of criticism of the original analysis (8 of 92). 50% of respondents from commercial companies (14 of 27) described "loss of control over data" as the main barrier.

The second most cited barrier was the risk that the data would be exploited without any benefit for the original researcher or study sponsor. Overall, 34% (78 of 228) of responses listed this issue as their main concern. 63% of these responses specifically mentioned a fear of lack of recognition of the trialist's contribution, e.g. co-authorship (49 of 78), 27% cited potential competitive advantage to others (21 of 78), and 10% were concerned with loss of IP.

21% respondents from companies (6 of 27) considered exploitation of data by "others" without benefit to the study sponsor as the main barrier.

11% of all responses addressed the effort and cost associated with depositing data in a database (26 of 228). 7% responses from companies listed this issue as the main barrier (2 of 27).

Table C 5 Barriers to researchers' willingness to deposit data in a repository

Main concern cited by respondent	All respondents	Industry respondents
Losing control over how the data would be used	40%	50%
Data exploited without any benefit for the original researcher or study sponsor	34%	21%
Effort and cost associated with depositing data	11%	7%

Survey question: "Please describe the one thing that you believe would impede researchers' willingness to deposit data in a clinical trial data repository". n = 228; n = 27 (industry)

C.11 Main concerns about researchers accessing individual participant data from a repository

The survey asked respondents to describe "the one thing that you believe would **stop researchers from using a clinical trial data repository**". Of the 201 responses, 14 cited two issues. 38 responses were irrelevant to the question and were subtracted, resulting in 177 responses in total (Table C 6). 26 respondents from companies provided answers for this question, one of which cited two issues, and 5 responses were irrelevant and hence subtracted, resulting in 22 responses in total. There were only 7 responses from "other countries", all touching on similar issues. Due to the low number, these were not analysed separately.

34% of the responses (60 of 177) were most concerned about issues with the quality of the deposited data, with 9% of all respondents expressing concern about a lack of data harmonisation or poor data structure. 20% felt that a cumbersome administrative approval process would be the main barrier, and 12% cited technical issues such as prescribed use of software or inability to download data. 11% thought researchers would be put off by the cost and effort involved in using the data, including potential access fees, and 7% listed a lack of understanding of the data, or not knowing what data was available, as the main barrier.

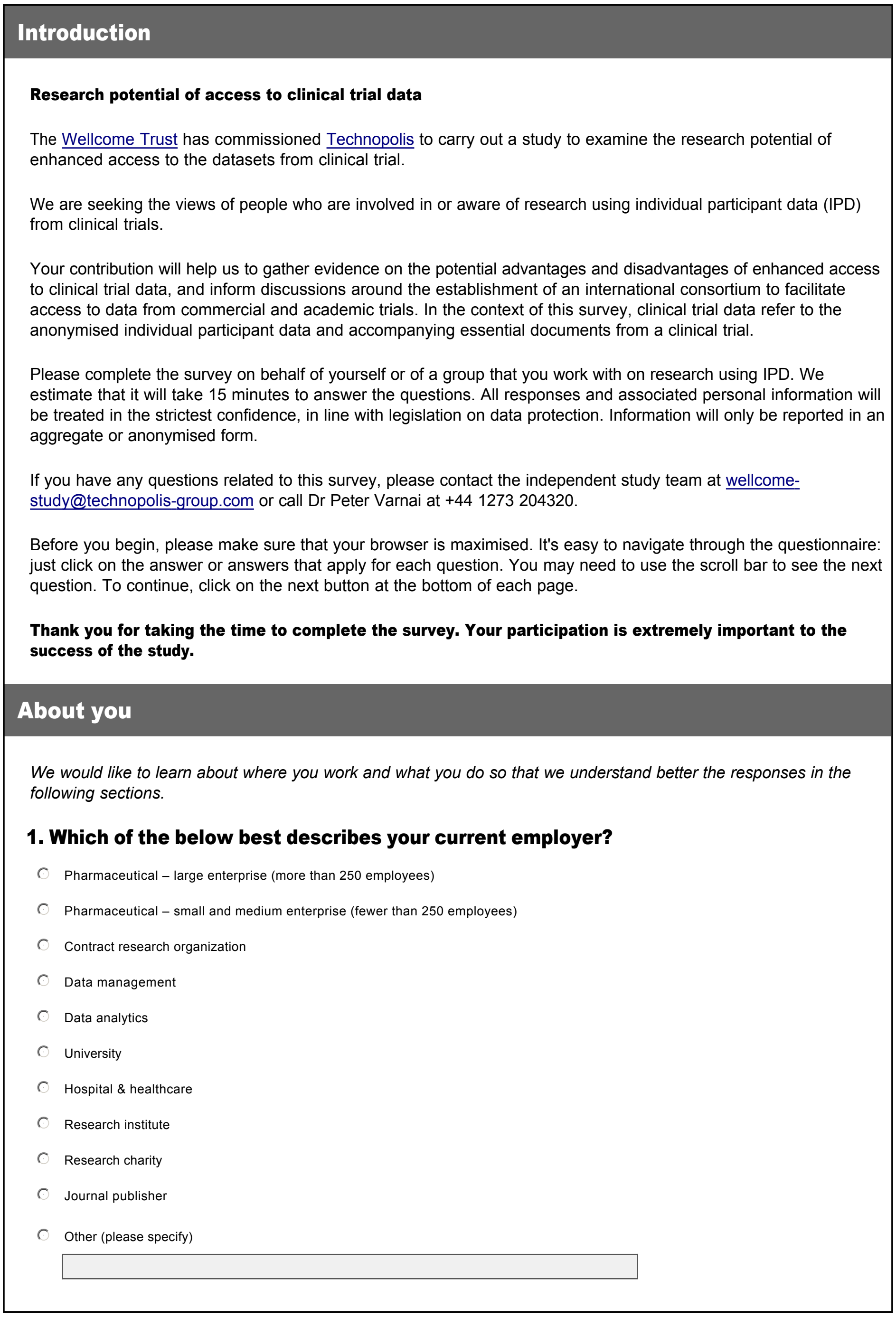
The responses from companies only were most commonly concerned with a lack of quality of deposited data, specifically with a lack of harmonised data or poor data structure (18%, 4 of 22), followed by cost and resources required for analysis (12%). Two responders (9% for each item) listed: burdensome access approval mechanism, and a lack of knowledge if the data was suitable for the planned analysis, insufficient data deposited in the database, and conflicts of interest if the data owner controls access to the data. Two (9%) respondents said that there were no barriers at all. Hence, while numbers were low, representatives from companies appeared to be concerned in particular by the level of harmonisation of datasets (18%, as compared to 9% of all responses), but were less concerned about burdensome approval processes (9%, as compared to 20% of all responses). The remaining responses covered a broad range of issues, with one respondent highlighting each.

Table C 6 Barriers to researchers' interest in using an individual participant data repository

Main concern cited by respondent	All respondents	Industry respondents
Quality of deposited data • Specifically mentioned lack of harmonisation or poor data structure:	34% 9%	27% 18%
Cumbersome administrative approval process	20%	9%
Technical issues	12%	0%

Survey question: "Please describe the one thing that you believe would stop researchers from using a clinical trial data repository". n = 177; n = 22 (industry)

Appendix D Survey questionnaire



Introduction

Research potential of access to clinical trial data

The Wellcome Trust has commissioned Technopolis to carry out a study to examine the research potential of enhanced access to the datasets from clinical trial.

We are seeking the views of people who are involved in or aware of research using individual participant data (IPD) from clinical trials.

Your contribution will help us to gather evidence on the potential advantages and disadvantages of enhanced access to clinical trial data, and inform discussions around the establishment of an international consortium to facilitate access to data from commercial and academic trials. In the context of this survey, clinical trial data refer to the anonymised individual participant data and accompanying essential documents from a clinical trial.

Please complete the survey on behalf of yourself or of a group that you work with on research using IPD. We estimate that it will take 15 minutes to answer the questions. All responses and associated personal information will be treated in the strictest confidence, in line with legislation on data protection. Information will only be reported in an aggregate or anonymised form.

If you have any questions related to this survey, please contact the independent study team at wellcome-study@technopolis-group.com or call Dr Peter Varnai at +44 1273 204320.

Before you begin, please make sure that your browser is maximised. It's easy to navigate through the questionnaire: just click on the answer or answers that apply for each question. You may need to use the scroll bar to see the next question. To continue, click on the next button at the bottom of each page.

Thank you for taking the time to complete the survey. Your participation is extremely important to the success of the study.

About you

We would like to learn about where you work and what you do so that we understand better the responses in the following sections.

1. Which of the below best describes your current employer?

○ Pharmaceutical – large enterprise (more than 250 employees)

○ Pharmaceutical – small and medium enterprise (fewer than 250 employees)

○ Contract research organization

○ Data management

○ Data analytics

○ University

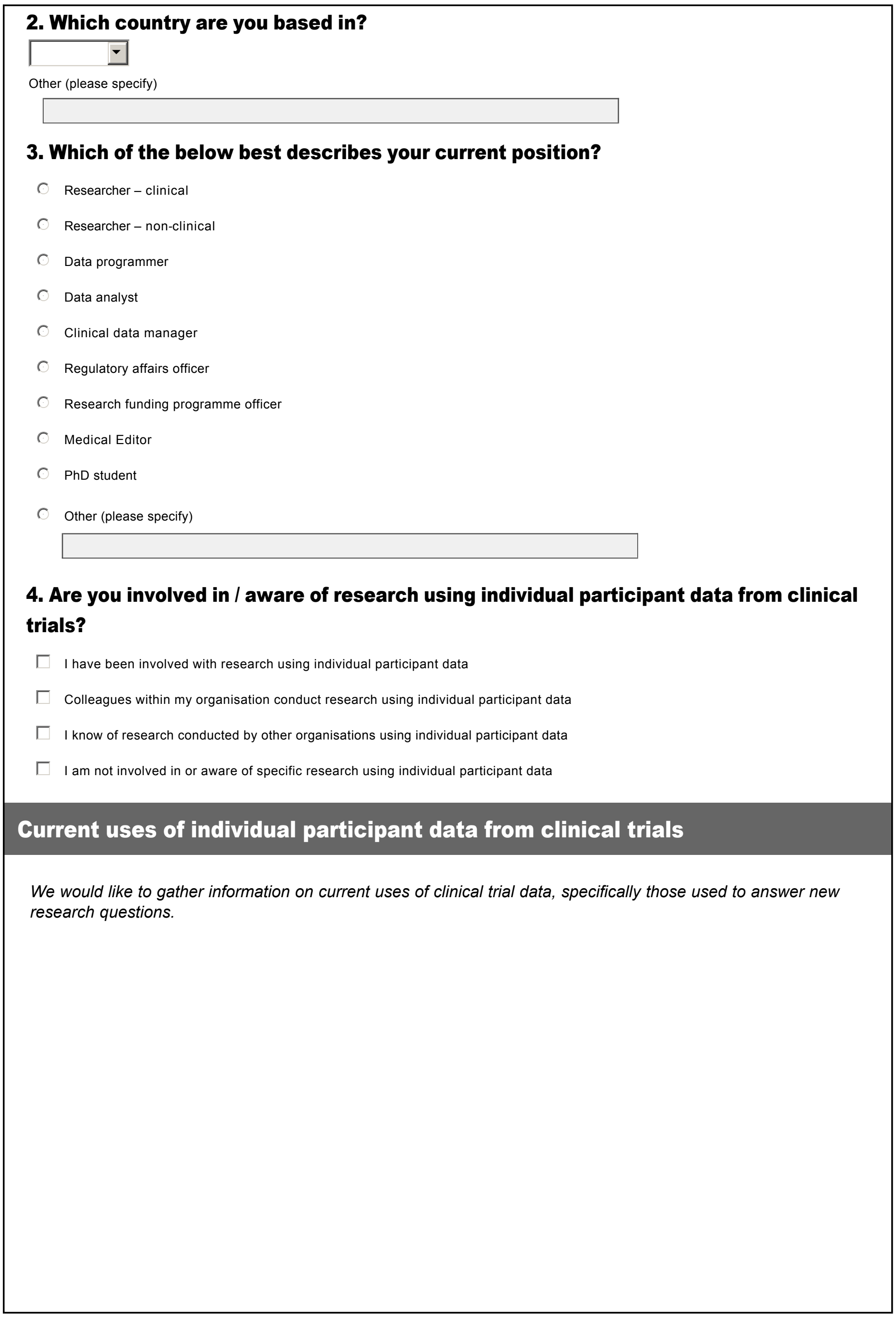
○ Hospital & healthcare

○ Research institute

○ Research charity

○ Journal publisher

○ Other (please specify)



2. Which country are you based in?

Other (please specify)

3. Which of the below best describes your current position?

○ Researcher – clinical

○ Researcher – non-clinical

○ Data programmer

○ Data analyst

○ Clinical data manager

○ Regulatory affairs officer

○ Research funding programme officer

○ Medical Editor

○ PhD student

○ Other (please specify)

4. Are you involved in / aware of research using individual participant data from clinical trials?

□ I have been involved with research using individual participant data

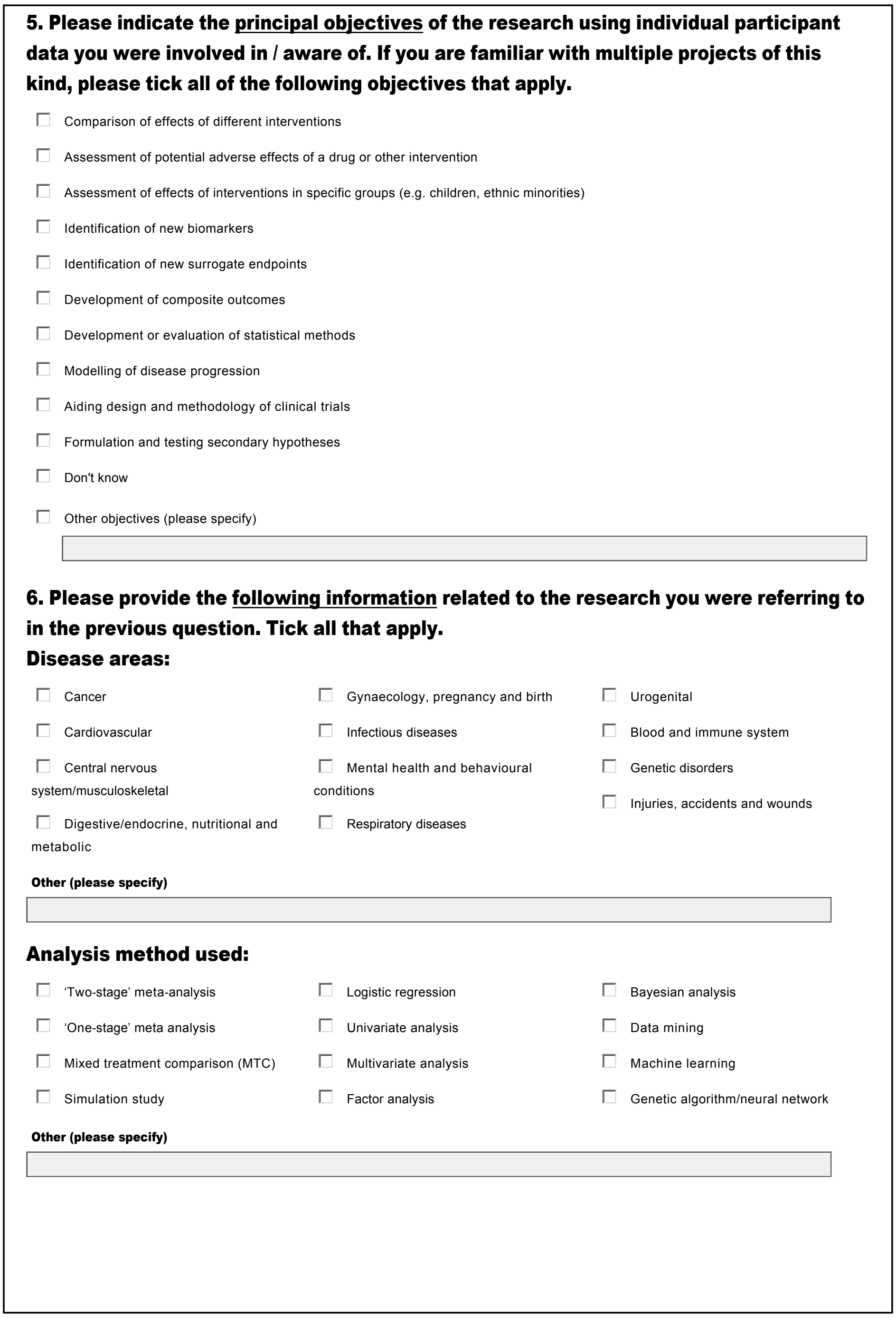
□ Colleagues within my organisation conduct research using individual participant data

□ I know of research conducted by other organisations using individual participant data

□ I am not involved in or aware of specific research using individual participant data

Current uses of individual participant data from clinical trials

We would like to gather information on current uses of clinical trial data, specifically those used to answer new research questions.



5. Please indicate the principal objectives of the research using individual participant data you were involved in / aware of. If you are familiar with multiple projects of this kind, please tick all of the following objectives that apply.

□ Comparison of effects of different interventions

□ Assessment of potential adverse effects of a drug or other intervention

□ Assessment of effects of interventions in specific groups (e.g. children, ethnic minorities)

□ Identification of new biomarkers

□ Identification of new surrogate endpoints

□ Development of composite outcomes

□ Development or evaluation of statistical methods

□ Modelling of disease progression

□ Aiding design and methodology of clinical trials

□ Formulation and testing secondary hypotheses

□ Don't know

□ Other objectives (please specify)

6. Please provide the following information related to the research you were referring to in the previous question. Tick all that apply.

Disease areas:

□ Cancer

□ Cardiovascular

□ Central nervous system/musculoskeletal

□ Digestive/endocrine, nutritional and metabolic

□ Gynaecology, pregnancy and birth

□ Infectious diseases

□ Mental health and behavioural conditions

□ Respiratory diseases

□ Urogenital

□ Blood and immune system

□ Genetic disorders

□ Injuries, accidents and wounds

Other (please specify)

Analysis method used:

□ 'Two-stage' meta-analysis

□ 'One-stage' meta analysis

□ Mixed treatment comparison (MTC)

□ Simulation study

□ Logistic regression

□ Univariate analysis

□ Multivariate analysis

□ Factor analysis

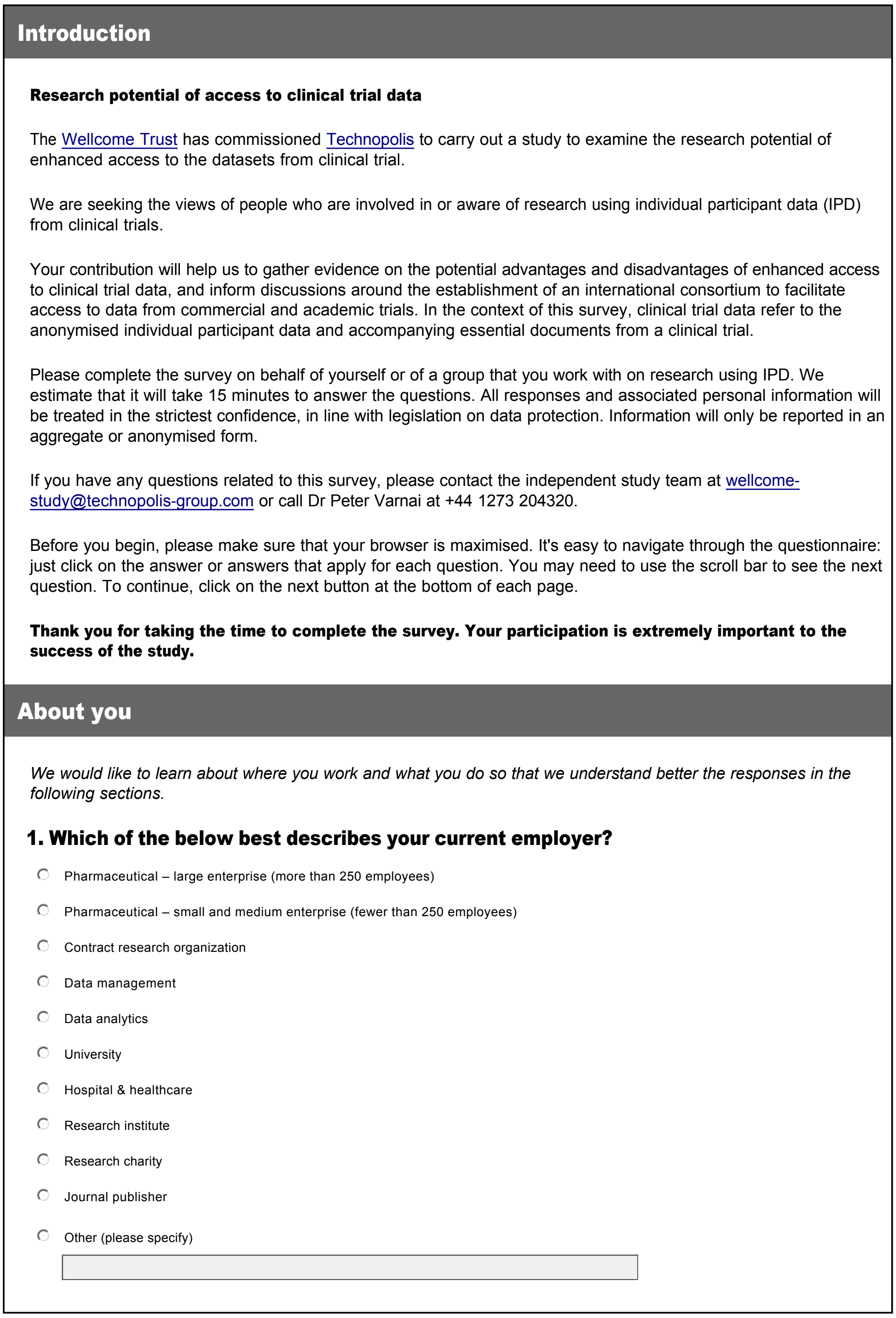
□ Bayesian analysis

□ Data mining

□ Machine learning

□ Genetic algorithm/neural network

Other (please specify)



Introduction

Research potential of access to clinical trial data

The Wellcome Trust has commissioned Technopolis to carry out a study to examine the research potential of enhanced access to the datasets from clinical trial.

We are seeking the views of people who are involved in or aware of research using individual participant data (IPD) from clinical trials.

Your contribution will help us to gather evidence on the potential advantages and disadvantages of enhanced access to clinical trial data, and inform discussions around the establishment of an international consortium to facilitate access to data from commercial and academic trials. In the context of this survey, clinical trial data refer to the anonymised individual participant data and accompanying essential documents from a clinical trial.

Please complete the survey on behalf of yourself or of a group that you work with on research using IPD. We estimate that it will take 15 minutes to answer the questions. All responses and associated personal information will be treated in the strictest confidence, in line with legislation on data protection. Information will only be reported in an aggregate or anonymised form.

If you have any questions related to this survey, please contact the independent study team at wellcome-study@technopolis-group.com or call Dr Peter Varnai at +44 1273 204320.

Before you begin, please make sure that your browser is maximised. It's easy to navigate through the questionnaire: just click on the answer or answers that apply for each question. You may need to use the scroll bar to see the next question. To continue, click on the next button at the bottom of each page.

Thank you for taking the time to complete the survey. Your participation is extremely important to the success of the study.

About you

We would like to learn about where you work and what you do so that we understand better the responses in the following sections.

1. Which of the below best describes your current employer?

○ Pharmaceutical – large enterprise (more than 250 employees)

○ Pharmaceutical – small and medium enterprise (fewer than 250 employees)

○ Contract research organization

○ Data management

○ Data analytics

○ University

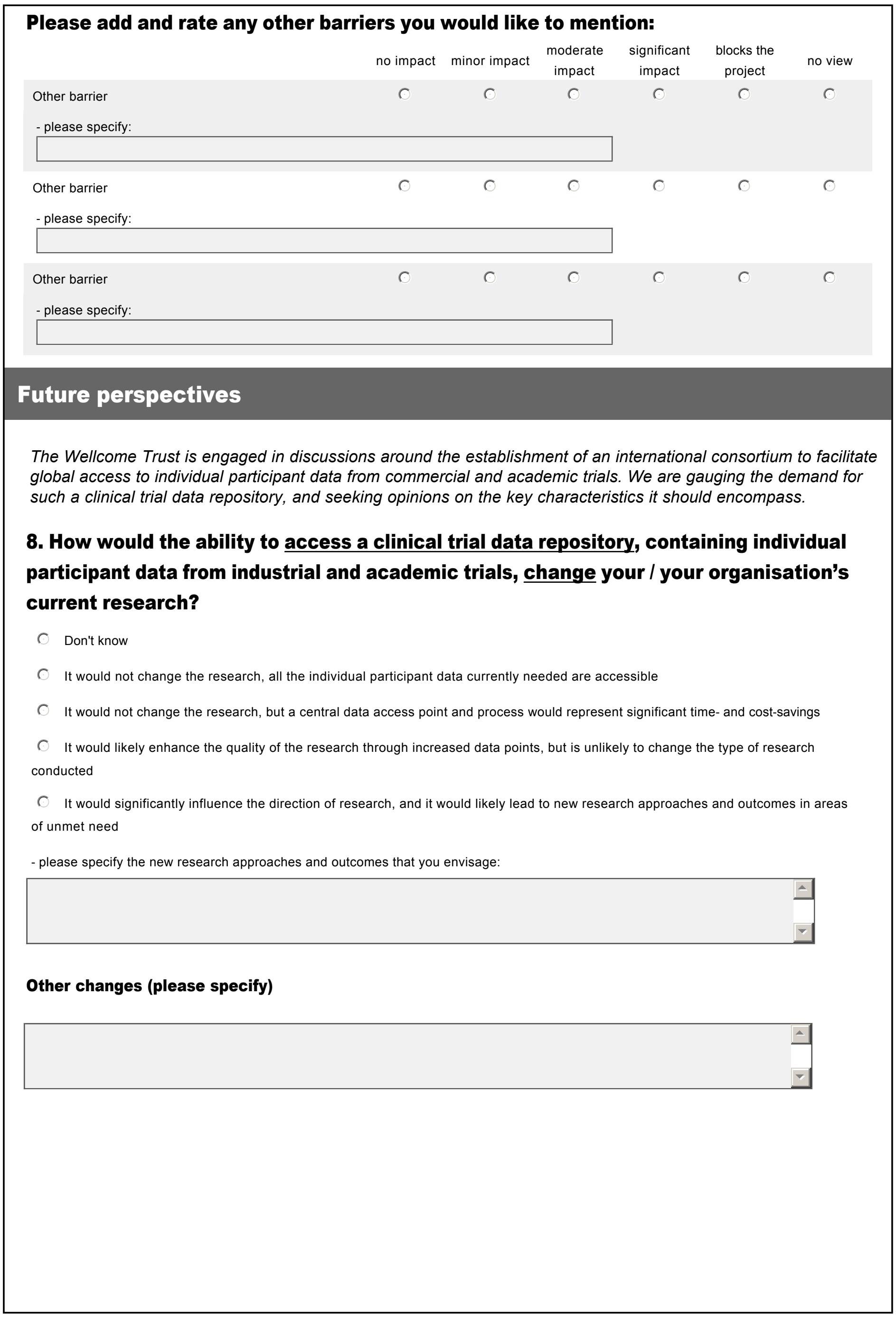
○ Hospital & healthcare

○ Research institute

○ Research charity

○ Journal publisher

○ Other (please specify)



Please add and rate any other barriers you would like to mention:

	no impact	minor impact	moderate impact	significant impact	blocks the project	no view
Other barrier	○	○	○	○	○	○
- please specify:						
Other barrier	○	○	○	○	○	○
- please specify:						
Other barrier	○	○	○	○	○	○
- please specify:						

Future perspectives

The Wellcome Trust is engaged in discussions around the establishment of an international consortium to facilitate global access to individual participant data from commercial and academic trials. We are gauging the demand for such a clinical trial data repository, and seeking opinions on the key characteristics it should encompass.

8. How would the ability to access a clinical trial data repository, containing individual participant data from industrial and academic trials, change your / your organisation's current research?

○ Don't know

○ It would not change the research, all the individual participant data currently needed are accessible

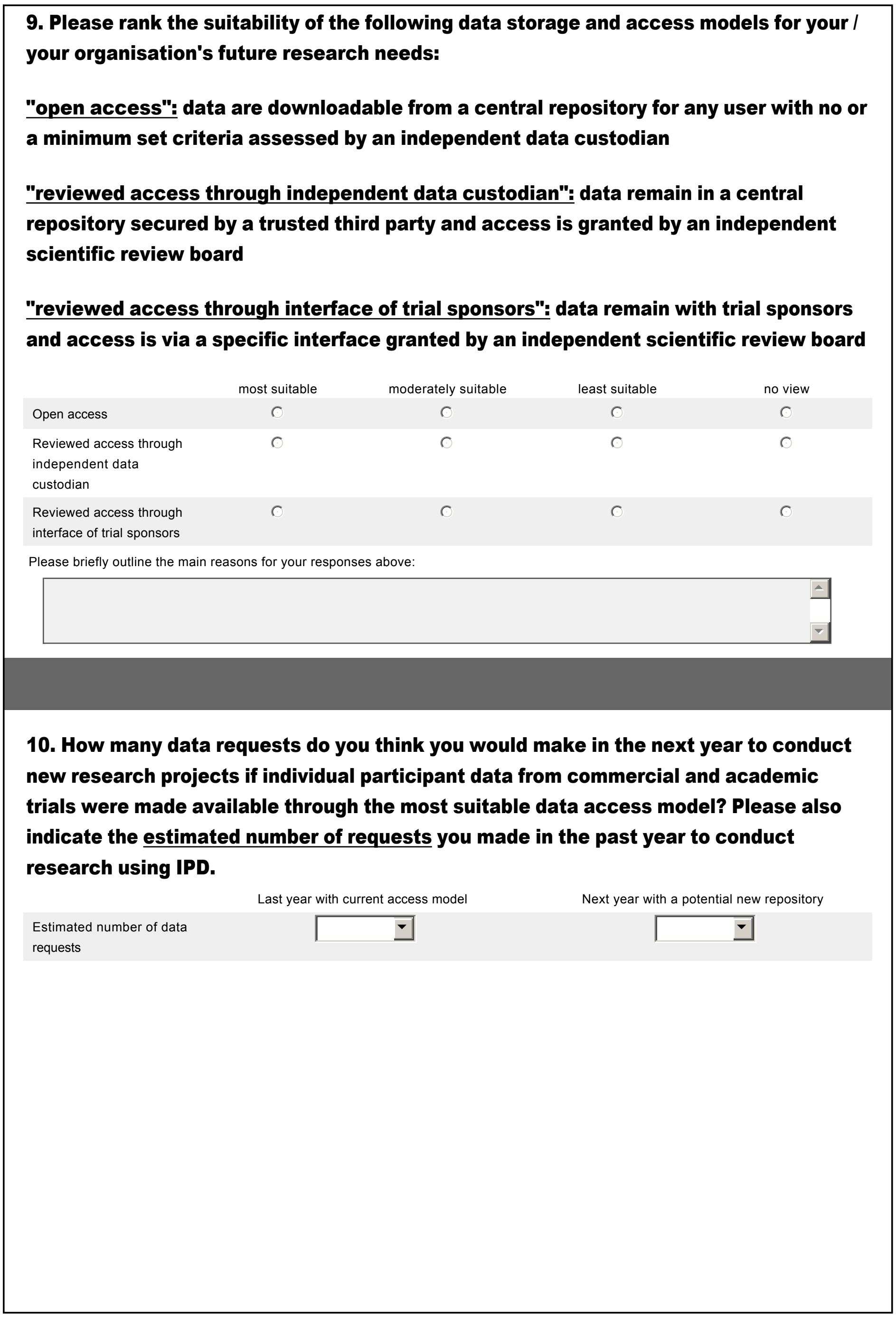
○ It would not change the research, but a central data access point and process would represent significant time- and cost-savings

○ It would likely enhance the quality of the research through increased data points, but is unlikely to change the type of research conducted

○ It would significantly influence the direction of research, and it would likely lead to new research approaches and outcomes in areas of unmet need

- please specify the new research approaches and outcomes that you envisage:

Other changes (please specify)



9. Please rank the suitability of the following data storage and access models for your / your organisation's future research needs:

"open access": data are downloadable from a central repository for any user with no or a minimum set criteria assessed by an independent data custodian

"reviewed access through independent data custodian": data remain in a central repository secured by a trusted third party and access is granted by an independent scientific review board

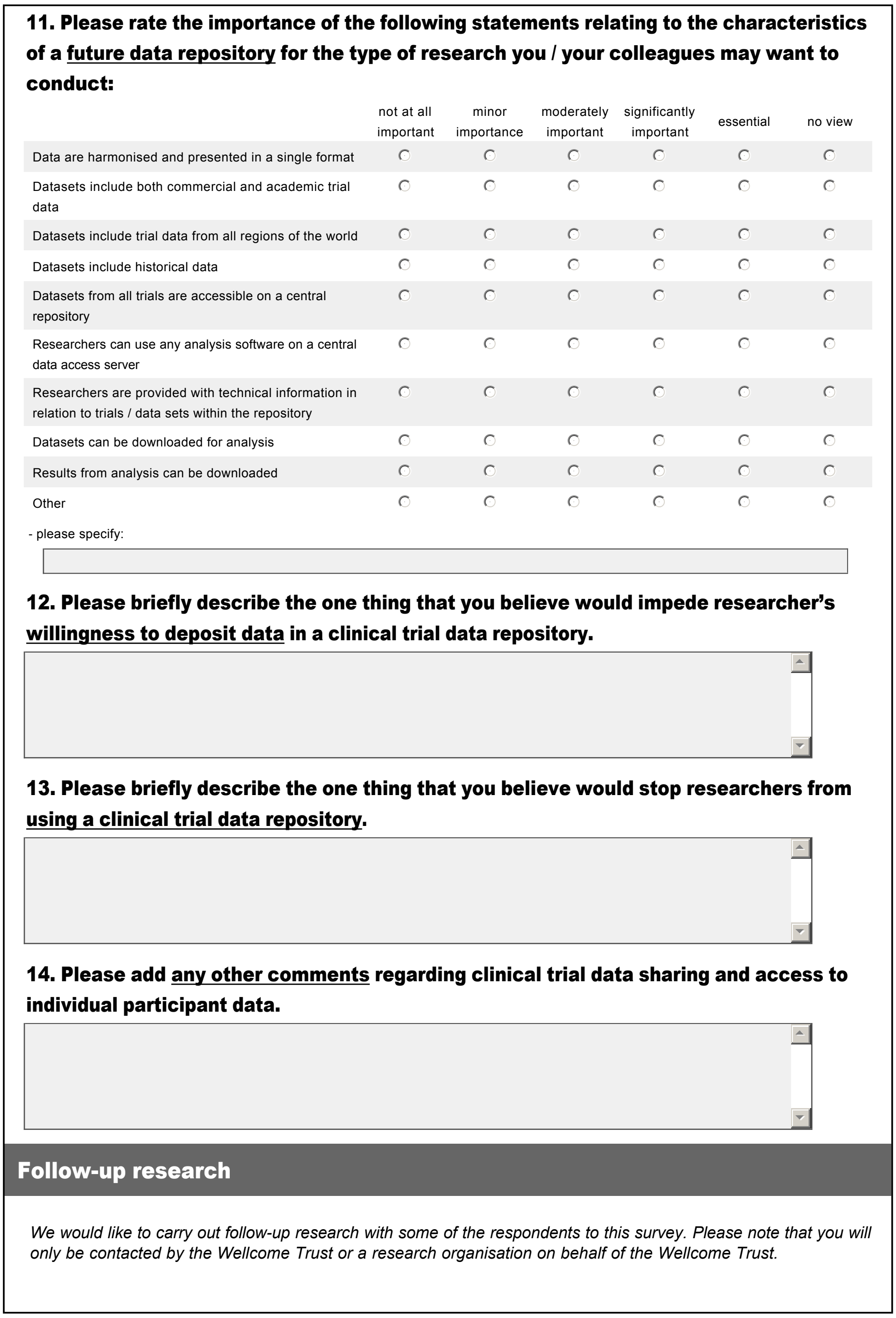
"reviewed access through interface of trial sponsors": data remain with trial sponsors and access is via a specific interface granted by an independent scientific review board

	most suitable	moderately suitable	least suitable	no view
Open access	○	○	○	○
Reviewed access through independent data custodian	○	○	○	○
Reviewed access through interface of trial sponsors	○	○	○	○

Please briefly outline the main reasons for your responses above:

10. How many data requests do you think you would make in the next year to conduct new research projects if individual participant data from commercial and academic trials were made available through the most suitable data access model? Please also indicate the estimated number of requests you made in the past year to conduct research using IPD.

	Last year with current access model	Next year with a potential new repository
Estimated number of data requests	▾	▾



11. Please rate the importance of the following statements relating to the characteristics of a future data repository for the type of research you / your colleagues may want to conduct:

	not at all important	minor importance	moderately important	significantly important	essential	no view
Data are harmonised and presented in a single format	○	○	○	○	○	○
Datasets include both commercial and academic trial data	○	○	○	○	○	○
Datasets include trial data from all regions of the world	○	○	○	○	○	○
Datasets include historical data	○	○	○	○	○	○
Datasets from all trials are accessible on a central repository	○	○	○	○	○	○
Researchers can use any analysis software on a central data access server	○	○	○	○	○	○
Researchers are provided with technical information in relation to trials / data sets within the repository	○	○	○	○	○	○
Datasets can be downloaded for analysis	○	○	○	○	○	○
Results from analysis can be downloaded	○	○	○	○	○	○
Other	○	○	○	○	○	○

- please specify:

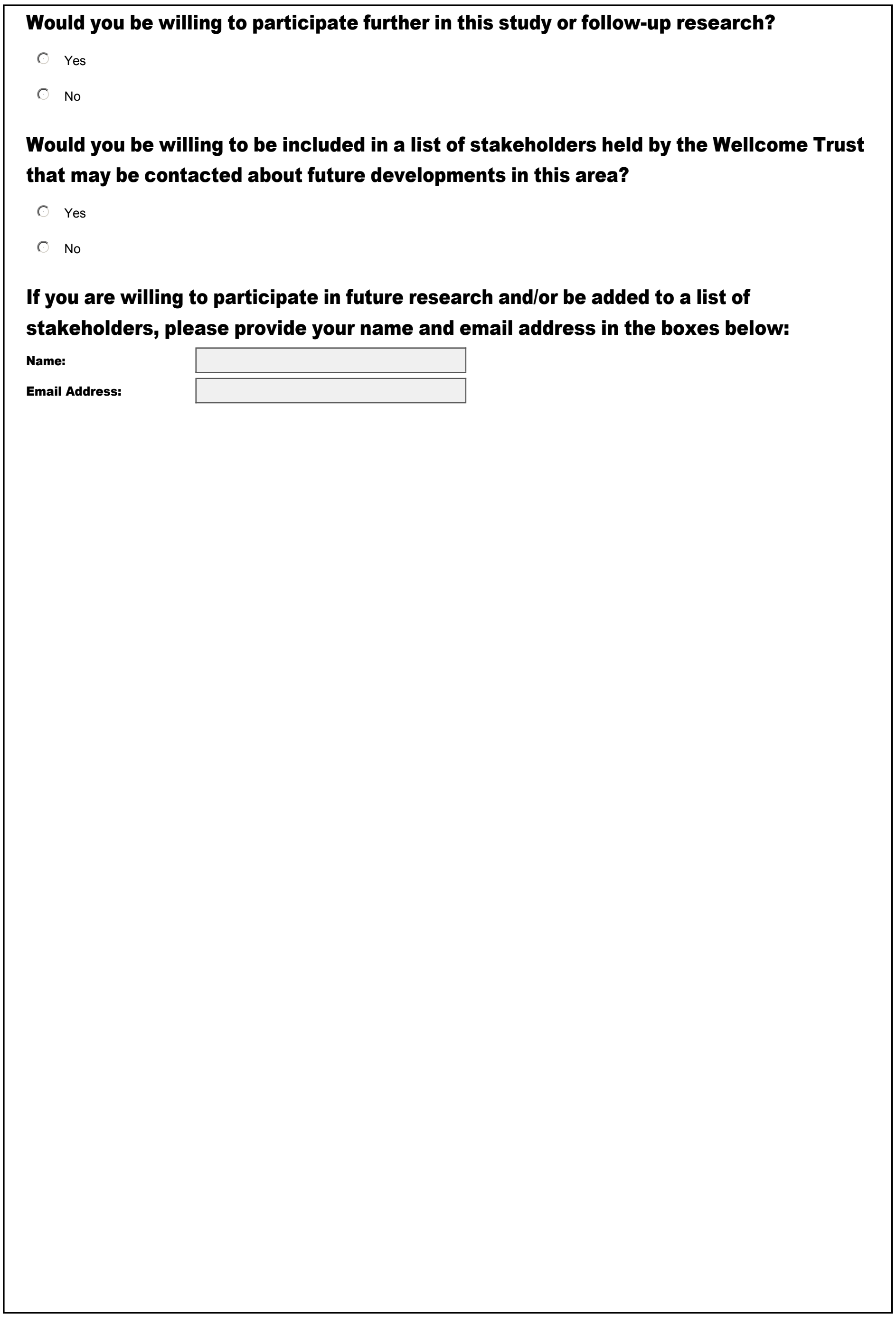
12. Please briefly describe the one thing that you believe would impede researcher's willingness to deposit data in a clinical trial data repository.

13. Please briefly describe the one thing that you believe would stop researchers from using a clinical trial data repository.

14. Please add any other comments regarding clinical trial data sharing and access to individual participant data.

Follow-up research

We would like to carry out follow-up research with some of the respondents to this survey. Please note that you will only be contacted by the Wellcome Trust or a research organisation on behalf of the Wellcome Trust.



Would you be willing to participate further in this study or follow-up research?

○ Yes

○ No

Would you be willing to be included in a list of stakeholders held by the Wellcome Trust that may be contacted about future developments in this area?

○ Yes

○ No

If you are willing to participate in future research and/or be added to a list of stakeholders, please provide your name and email address in the boxes below:

Name:

Email Address:

Appendix E List of Interviewees

Name		Organisation
Doug	Altman	University of Oxford
Enrique	Aviles	C-PATH Institute
Jesse	Berlin	Johnson & Johnson
Jan	Bogaerts	European Organisation for Research and Treatment of Cancer (EORTC)
Marc	Buyse	IDDI / Cluepoints
Mike	Clarke	Queen's University Belfast / EBCTCG
Sean	Coady	US National Institutes of Health - NHLBI
Robert	Cuffe	ViiV Healthcare
Martin	Daumer	Sylvia Lawry Centre for MS Research
Ruxandra	Draghia-Akli	Research DG of the European Commission
Christine	Fletcher	Amgen
Susan	Forda	Lilly / EFPIA
Andrew	Freeman	GSK
Robert	Frost	GSK
Ben	Goldacre	London School of Hygiene and Tropical Medicine
Phillipe	Guerin	WWARN
Beth	Hodshon	Yale University / YODA
Sally	Hollis	AstraZeneca
Torsten	Hothorn	University of Zurich
François	Houÿez	EURORDIS
Dipak	Kalra	University College London / IMI - EHR4CR
Steven	Kern	Bill & Melinda Gates Foundation
Ronald	Krall	formerly GSK
Kate	Law	Cancer Research UK
Andrew	Maas	University Hospital Antwerp
David	Madigan	Columbia University
Sarah	Meredith	University College London, MRC CTU
Sarah	Nolan	University of Liverpool
Nicola	Perrin	Wellcome Trust
Liz	Philpots	Association of Medical Research Charities
Subha	Rajanaidu	University of Nottingham, formerly GSK
Rebekah	Rasooly	US National Institutes of Health, NIDDK
Bina	Rawal	Association of the British Pharmaceutical Industry
Fiona	Reddington	Cancer Research UK
Haleema	Shakur	London School of Hygiene and Tropical Medicine
Rebecca	Sudlow	Roche
Matt	Sydes	University College London, MRC CTU
Catrin	Tudur-Smith	University of Liverpool
Bart	Vannieuwenhuyse	Janssen Pharmaceutica NV / IMI - EMIF
Paul	Wicks	Patientslikeme
Neta	Zach	Prize4Life

Appendix F Topic Guide for Interviews

Interviewer	
Name of the interviewee	
Date	

Aims and objectives

- Elaborate on the survey responses to answer the evaluation questions in more detail
- Complement existing stakeholder views
- Develop case study material that illustrate findings of the study

Background

Confirm current position and affiliation

History of involvement with IPD

General interview guide

Current uses of individual participant data

1) Your project/expertise, past projects:

- Disease area
- Aim (biomarker, compare treatments, etc)
- Impact – potential and achieved
- Source of data – if external, describe the access process, experiences

- Current barriers, difficult steps, issues
- How could this have been enhanced?

2) Projects of "others" (for case studies):

- What do you consider the study / area of work using existing IPD with the most impact?
- What are one or two very exciting, promising, **novel** uses of IPD you are aware of?
- Are you aware of other databases?

Future data sharing:

1) Access: Please share your views and concerns on different data access models: open, independent review, held by study sponsor.

2) How would such a model provide access to data from different trials from different time periods? Is this important?

3) What are the benefits or otherwise of providing access to data from academic and non-commercial trials alongside or in combination with those from commercial trials?

4) What are the appropriate safeguards needed to ensure scientific robustness and to protect against inappropriate use or disclosure?

5) Where do you see exciting new uses of IPD? Do you think enhanced access will draw in new research communities?

6) What do you think is future demand?
What could block interest?
What are suitable incentives, for depositing data and for using data?

7) What types of clinical trial data might be used for such research? What types of data have the greatest research potential? What about images etc?

Are you happy for us to:

- send you the write-up of this interview for you to check
- incorporate your comments into the report (statements will not be attributed to you)
- send you the sections relating to your repository that will be included in the final report, for comment

Specific questions for Interviews with Database coordinators

1) Database / repository basics:

- Date repository was created
- Organisations that initiated/financed the repository
- Organisation/country the repository is hosted by
- Disease area(s) repository includes data for
- Type of data deposited (treatment arm, comparator/placebo arm, others e.g., biobank samples)
- Size of the repository (e.g., number of distinct datasets)
- Organisations that contribute to the repository (names of academic/ commercial organisations with number of data sets if available)
- What types of data are included?

2) Access to data in repository:

- access model: open, independent review, held by study sponsor (views, pros and cons)
- Describe process: steps, average duration from application to data access
- Eligible requests definition (e.g., qualified researcher, reasonable request)
- Data types provided to researchers
- Access modality and practicalities of data analysis
- Statistical software – 'R' and 'SAS' are provided
- Data analyses types supported
- Retrieval format of results

3) Outputs / impact:

- Demand: number of requests over years, trends, academic/industry
- Research outputs of the repository

- Accessibility of the resulting research output (e.g., peer-reviewed publications in open access journals, conference presentations, etc)
- If you had to profile what has been achieved through data sharing, what would be the study to highlight?

4) Experience:

- What are the main obstacles and/or bottlenecks:
  at the point of data deposition (data provider / database staff)
  at the point of data access (data user / database staff)
  regarding data preparation (database staff)

5) Other:

- Are you aware of other data repositories?
- What are future plans for your repository?

Are you happy for us to:

- send you the write-up of this interview for you to check
- incorporate your general comments into the report
- send you the sections relating to your repository that will be included in the final report, for comment

Appendix G Expert workshop summary report

This report is a summary of the workshop that took place at the Wellcome Trust, London (UK) held on June 27, 2014. This one-day event was organised by the Wellcome Trust and Technopolis Group to explore current and future research opportunities using individual participant clinical trial data (IPD) through presentations, discussions and teamwork. The workshop was intended to further inform investigations conducted as part of the study "Assessing the research potential of access to clinical trial data", commissioned by the Wellcome Trust in April 2014.

Participants included clinical researchers from academic and industrial settings, data scientists, and managers with policy, funding and database management backgrounds. A list of the participants is available in Annex I.

The aims of the workshop was to bring together a range of perspectives and generate new insights into the wider research opportunities of clinical trials data sets:

- Identify the types of clinical research questions that are currently possible by rapid stock taking and cataloguing current activities.
- Explore potential future research directions that would be opened up through enhanced access to IPD from pooled academic and commercial sources.
- Explore how different access models impact on the potential future research directions

The organisers of the workshop recognise that clinical trial data sharing is a complex subject and requires a careful analysis of many aspects including the appropriate incentives for sharing, patient consent, patient privacy, data ownership, data harmonisation, regulatory issues, commercial sensitivity, financial costs, governance, etc. The focus of this workshop was deliberately on the potential for new research areas, and thus the detailed discussion of the other issues was beyond the scope of this workshop.

Agenda

9:00-9:30	Registration, coffee and tea
9:30-9:50	Welcome - Alison Cave (Wellcome Trust)
	Introduction - Jeff Rodriguez (facilitator)
9:50-10:20	Scene setting
	Consensus study on responsible sharing of clinical trial data by the Institute of Medicine - Trudie Lang (Oxford University)
	Overview of current IPD sharing initiatives - Maike Rentel (Technopolis)
10:20 -10:40	Stock-taking exercise on current clinical research directions
10:40-11:00	Break, coffee and tea
11:00-11:10	Report back to plenary on categories of current research directions
11:10-11:40	What can be achieved with IPD-level data? The PRO-ACT database Neta Zach (Prize4Life) and Robert Küffner (Helmholtz Centre Munich)
11:40-12:10	Introduction to the 'Utopia database' - Peter Varnai (Technopolis) Individual time to develop mini research proposals
12:10-13:00	Lunch

13:00-14:20	Research exercise in teams to develop research projects
14:20-15:20	Plenary discussions: new research ideas and future action plan
15:20-15:30	Conclusions and next steps

1. Welcome and general introduction

Alison Cave welcomed workshop participants and emphasised the importance of discussing with diverse stakeholders the topic of responsible sharing of participant level clinical trial data (IPD) in order to advance science and improve therapies.

2. Setting the Scene

Trudie Lang provided an overview of current clinical research practice and data sharing, Maike Rentel presented an overview of the current IPD sharing landscape, and provided examples of sharing initiatives holding data related to specific diseases, publicly and commercially funded trials, and their respective access mechanisms. In the ensuing discussions participants noted that researchers needed access to pooled clinical trial data to enable development in targeted medicine.

3. Stock-taking exercise - current activities

This individual exercise aimed at identifying the types of clinical research questions that are currently possible using IPD. Participants were asked to write down (i) the top projects from their own work where they used IPD; (ii) other major/innovative work they were aware of. The study team subsequently sorted these into categories, available in Annex II.

4. What can be achieved with individual participant data?

Neta Zach provided a historical overview of the PRO-ACT database by Prize4Life, a patient-led project to approach pharmaceutical companies and collect all available ALS clinical trial data, funded by the ALS Therapy Alliance's. Their goal is to understand disease heterogeneity and make clinical trials more effective. The database contains cleaned, harmonised, and aggregated data that have been accessed by researchers over 350 times through submitting a research proposal since December 2012. Scientific findings derived from the database include comparative survival benefits of site of onset, BMI and age; and prediction of slow and fast progressors. Robert Küffner presented the experience of the ALS Prediction Prize Challenge that was launched with the aim to develop an algorithm to predict 3-month disease progression. Over 1000 participants were involved in the challenge, which was crowd-sourced via Innocentive's global network. The resulting algorithms predicted disease progression much more accurately than 12 top clinicians based on data alone. The algorithms may help clinicians in the future to stratify patients and reduce the number of patients needed for clinical trials. After the presentations, the participants discussed the higher quality and coverage of data from controlled clinical trial in contrast with clinical data outside of trials. Despite the fact that PRO-ACT contains data from clinical trials only, the datasets had missing information. Neta Zach explained that it was not possible to know whether the data was missing or never collected as part of the trial. In addition, data was anonymised in order to protect patient privacy, by removing the link between individual patients and the trials they were part of. Participants of the workshop discussed the limitations that this brought for any re-analyses of these datasets; and many felt that this strategy should be avoided because it would limit the type of meta-analysis that could be conducted.

4. Research exercise using the Utopia database

Peter Varnai described the idea of an imaginary database that included IPD from academic and industrial clinical trials with a global geographic coverage and multiple disease areas. This database, tentatively named Utopia, would make anonymised datasets from both treatment and placebo arms available with all recorded parameters harmonised to a common data standard. Users would access data on a centralised server where any analysis software could be used and analysis results were downloadable. The repository would go far beyond what was currently possible globally. Participants had been given the description of the Utopia database before the workshop and were asked to reflect on what research they would propose based on such data. The criteria for research projects were that they made maximum use of the database (no existing database can fit the requirements of the project), preferably used innovative methodological approaches and the results would ultimately benefit the patient and the research community.

The purpose of the session was to encourage participants to generate ideas about the potential research uses of having enhanced access to IPD. Participants first developed mini-proposals independently and then pitched these ideas in four break-out groups. Participants were seated in pre-arranged groups so that to maximise diversity within each group; a balance was accomplished with academic and industrial clinical statisticians, 'outsider' data scientists, other relevant people with policy, funding and database management expertise, and a moderator. Each project idea was presented and discussed in turn, and finally the group selected a proposal (or created a combined proposal) that best matched the criteria based on consensus. The selected proposal was further developed in a collaborative fashion that resulted in a 'Utopia project' for each group.

Next, several constraints were introduced by the facilitator to gauge what effect these may have on the 'Utopia project'. The groups discussed and prioritised the impact of relevant data access restrictions for the project from a list including

- No simultaneous access to multiple datasets
- Data not harmonised
- Only restricted statistical software allowed
- Limited geographical spread of data
- No access to a set of parameters to privacy issues
- Missing data
- No exploratory data analysis allowed

Participants were encouraged to find work-around solutions to the constraints imposed without shutting down the entire project. Participants found the idea sharing exercise stimulating and contributed to collective learning. Finally, each group presented its project in a plenary session followed by a group discussion.

Summary of Team proposals

The following proposals were put forward by the participating teams to utilise the Utopia database.

Finding off-target effects of drugs

The area of study is the investigation of drug side effects in terms of safety and unexpected benefits. This type of investigation is currently not possible because of the limited size and scale of the data currently accessible (small data set from single clinical trial). The objectives are to determine the association between drug characteristics and (1) adverse effects and (2) disease modifying effects. This will help to inform (1) safe use of existing medicines, (2) generation of new targets for further research and (3) early development of research into compounds. The proposed approach involve classification of drugs according to mechanism of action, short-term biological effects & physico-chemical properties; exploration of the association between drug characteristics and events; collation of incidences from various; validation of classes of side-effects. The rationale for using Utopia are: (1) its size (individual adverse side effects don't turn up frequently); (2) its heterogeneity (the ability to explore drug); (3) its data quality.

Cohort analysis of unexpected severe adverse events

The aim of the project is the prediction of severe adverse events (i.e., myocardial infarction, hemorrhagic stroke, ischemic stroke, severe infection, sudden death) in the short to medium term. This can be achieved through the identification of new biomarkers and the development of a risk model for unrelated major health issues. Since severe adverse events represent rare events, a large dataset (both placebo and treatment arms) with detailed clinical and biochemical data prior to event is required. The project will perform a survival analysis with traditional statistics first and then extend it to machine learning to capture more complex interaction between variables in an exploratory phase, followed by validation and cross-validation with other datasets.

Co-existing disease conditions: Dementia

The objective of this research project is to differentiate disease subtypes and classify clinical features in order to understand rate of progression. A public health challenge and a significant unmet need where cause and effect remain unclear. The expected drug response differs by disease type and phenotype of patient characteristics. Therefore treatment will be targeted by disease subtypes. In a pooled data source specific issues associated with under-represented minorities and age groups (young, elderly, minorities) will be alleviated. The proposal includes a hypothesis-generating phase and then testing in a clinical trial. It is expected that such an analysis will save cost in clinical trials. Additional data beyond those measured in clinical trials may be helpful: imaging data, and life-style data on nutrition, previous employment history etc

Identification of Surrogate End-Points

The aim of this project is to look at a particular intervention across various conditions and identify putative surrogate end-points, validate and optimise those to reduce burden on patients and improve clinical trial design potentially with a view to changing medical and regulatory practice and speed up the clinical trial process. This type of project can benefit from developing and applying novel techniques, such as machine learning, to identify potential endpoints in a test dataset for validation in a larger dataset, made possible by a dataset as large as Utopia. The development of prediction tools such as machine learning algorithms.

Stratifying patient populations

The aim of this project is to stratify patient populations according to comorbidities, medicines taken and lab results to identify sub-groups of patients to look for shared pathways and understand disease mechanism based on clinical data and lab data. This will bring together information from lab data and clinical profiles from co-morbidities to identify new pathways associated with disease. This will also contribute to identifying rare adverse events and improving understanding and predicting patient outcomes. Also provides opportunity for repurposing drugs for new indications. The approach proposed the use of Network Analysis, Cluster Analysis and Hierarchical Analysis of demographic, clinical data, lab data, comorbidity data.

Effects of restrictions on data access

Participants discussed the different possible constraints and how those may affect their projects. Certain projects proposed exploratory data analysis which, if not allowed, would block a project from the outset; this team highlighted the need for the opportunity to let the data lead researchers. Although a comprehensive dataset may not be needed for meaningful analyses, genetic variability from diverse geographical areas would be very important. If key data measured in clinical trials were not made available to researchers due to privacy issues, this would again block the project to succeed. However, anonymisation of data and public perception were considered very important, and thus we heard the advice to tread carefully.

On the technical side, researchers felt that if no simultaneous access is provided for multiple datasets, this would seriously hinder the applicability of certain key techniques. The need for harmonised (good quality) datasets and straightforward access to those were considered important points for research efficiency and data usability.

It was noted that linking different databases, including data from observational and (epi)genomic studies, at the participant level will be very important to enhance the potential for future research projects. Sequencing data will be routinely collected in future clinical trials, but currently little is known how to maximise the use of this information. As a case in point, it was remarked that the tumour genomic profile changes over time and need regular sampling in order to predict drug resistance.

5. Conclusions

In the final plenary discussions, participants discussed the characteristics and benefits of enhanced access to a pooled database of IPD based on the previous research exercise. Key features of such a database were the size, diversity, and quality of the underlying data. Linking clinical trial data to other types of data (genomic, ‘real world’, trauma, lifestyle, registries, etc) was considered an essential next step to increase the information content of the data for analysis. Anonymisation of data was an issue in itself, particularly in relation to comparing results in trials and accounting for differences between trials. In addition, linking anonymised data to other data sources might not be possible. The risk to privacy of trial participants was acknowledged to be a significant issue, and public awareness about the use of participant data for research should be boosted.

In terms of research use of pooled data, patient stratification was repeatedly highlighted by participants. Combined datasets will allow unprecedented insights into ‘difficult’ patient groups, e.g. children or patients with rare diseases. While some of these groups exhibit homogeneous characteristics, patients with common diseases (e.g., inflammation, pain) may show a large number of symptoms, very heterogeneous in nature. This requires extensive data to achieve statistical power in analysis. Predicting adverse effects and investigating timing of treatment in chronic diseases provide further examples in the long list of research opportunities enabled by enhanced access to IPD workshop participants uncovered.

Participants were asked what needs to happen to make these research ideas possible within the next 3-5 years; the following points were collected:

- We are not starting from zero as the stock-take showed, we have already started out on this trajectory.
- We have to take account of public anxiety about access to IPD, based on a wider set of worries and trust in professionals and others with human tissue, data etc. This may point to the need of launching a programme of public education, debate and interaction around the benefits and risks of access to IPD.
- The research community itself needs to be a subject of further awareness and development. The positive approach seen today may not be shared everywhere. Real examples used as part of a campaign will be persuasive.

- There should be caution about too much de-identification: there may be positive benefits to some linking back to an individual patient, especially of subsequent re-examination of data reveals a condition on which remedial action needs to be taken.
- Entirely 'open access' will be problematic as it will make it more difficult to control for the robustness of the data and rigorousness of the data user.
- Standardisation of data elements is a top priority.
- Academic research tends to focus on clinical research and not on data release. Incentives and contractual obligations will help focus the academy on both.
- A lot of preparatory and persuasive work needs to be done to knit systems together – a trusted and independent party may well take on the to assemble a coalition of interested parties
- Tackling industry anxiety -fear of reputational risk- will be one of the first tasks.
- The catalogue of current restrictions and constraints needs to become part of the agenda for change.
- A single point of access needs to be considered, perhaps as a longer-term goal. However, in the initial phases there may be a need to progress competing initiatives as no generally accepted model exists.
- Philosophically, we should remain impatient in pursuing this agenda because we are still at the 'crawling' stage.

Final thoughts on the workshop from participants

'Good news'	**'Bad news'**
Diversity of participants working together	We can't do it right now
Researchers are trying to maximise what they can learn	Even with Utopia, solving major medical problems will still be a challenge
Many exciting research ideas	There are still reasons to do nothing
A more aware research community	
Good to see different perspectives	
We talked about it!	

Annex I List of Participants

Name		Organisation	Position
Judith	Bliss	Institute of Cancer Research, London	Professor of Clinical Trials, Director of the Cancer Research UK funded Clinical Trials & Statistics Unit (ICR-CTSU) and Deputy Head of the Division of Clinical Studies
Robert	Cuffe	ViiV Healthcare	Head of Statistics
Maria	Dilleen	Pfizer	Senior Director Statistician at Pfizer
Robert	Frost	GSK	Policy Director, Medical Policy
Christopher	Hart	AstraZeneca	Information Practice Leader
Robert	Küffner	Helmholtz Centre Munich	Group leader of the Practical Informatics and Bioinformatics Group
Trudie	Lang	U Oxford	Principal Investigator, Global Health Network
Marcia	Levenstein	Pfizer	VP Statistics
Stephen	Pyke	GSK	SVP Quantitative Sciences
Fiona	Reddington	Cancer Research UK	Head of Clinical and Population Research Funding
Peter	Sasieni	QMUL	Professor of Biostatistics and Cancer Epidemiology, Director of Cancer Prevention Trials Unit
Haleema	Shakur	LSHTM	Senior Lecturer (Clinical Trials) & CTU Co-Director
John	Shaw-Taylor	UCL	Director of the Centre for Computational Statistics and Machine Learning (CSML)
Ricardo	Silva	UCL	Lecturer at the Department of Statistical Science and Adjunct Faculty of the Gatsby Computational Neuroscience Unit
Mark	Simmonds	U York	Research Fellow, Centre for Reviews and Dissemination
Lesley	Stewart	U York	Centre for Reviews and Dissemination
Katherine	Tucker	Roche	Patient Level Data Co-ordinator
Chris	Watkins	Royal Holloway	Reader, Department of Computer Science
Neta	Zach	Prize4Life	Scientific Director
Will	Greenacre	Wellcome Trust	Policy Officer
Alison	Cave	Wellcome Trust	Head of Cellular, Developmental and Physiological Sciences
Peter	Varnai	Technopolis	Principal, Health & Life Sciences
Maike	Rentel	Technopolis	Senior Research Associate
Paul	Simmonds	Technopolis	Managing Director
Tammy	Sharp	Technopolis	Consultant
Bastian	Mostert	Technopolis	Consultant

Annex II Stock taking exercise (agenda item 3)

Participants were asked to write down (i) the top projects from their own work where they used IPD; (ii) other major/innovative work they were aware of. The study team subsequently sorted these into categories, which are presented in this Annex.

Trial methodology, Design, Methods Applied
Tools to simulate clinical trials, reasons for dropout in clinical trials
Disease progression prediction, disease stratification
Accessing free text part of GP records to detect onset of ovarian cancer and coronary heart disease
Databases
Freebird, Project DataSphere, EORTC, RECIST, Transcelerate, PRO-ACT, internal oncology database, research collaborator
Improved model of disease course
Using correlations between questionnaire data and proteomic patterns to refine categorisation accuracy for different stages of malaria and TB
Exome sequencing of schizophrenia trios: disease causing (accompanying) sequence variations
Effectiveness of interventions
Reasons for drug discontinuation, drug switching, choice of drug cost-effectiveness
Systematic review of IPD meta-analysis of rhBMP2 for spinal fusion surgery
Linking colposcopy treatment to obstetric outcome (PaCT)
Identify genetic characteristics of drug responders
Assessing consistency of response to treatment across subgroups
Long term effects of aromatase inhibitors
Risk prediction of adverse outcomes arising from febrile neutropenia in children with cancer (PICNICC)

Standards, Tools, Methods, Platform
Automating the analysis of multiple clinical trial data through advanced Artificial Intelligence methods
Intergrowth-21 project, on new-born health aimed at collecting outcome data and create an agreed set of standards
Sharing oncology datasets for RECIST criteria exercise
Observational clinical data: observational medical outcomes partnership (OMOP)
CART analysis to identify characteristics of responders and non-responders to medicines
Deriving PK/PD model from clinical trial data
Predictors of clinical progression
Health (NHS) datasets available via NCIN, HSCIC, CPRD, for informing likely recruitment rates
Diseases
Statins, long term safety choice of drug interactions
Multiple omics data generation and phenotypisation to characterise rare disease causing factors (KORA cohort)
Melanoma tumour size to overall survival
Trial Set-Up Methodologies
New clinical trials informed by analysis of previously conducted trials in relevant population
Translation of early response to late response to design proof of concept studies for new compound development
Use of CAMD placebo data and model to inform proposed trials in Alzheimer’s disease

Data Sharing, Collaborations
IPD systematic reviews, meta analysis, database of projects held by Cochrane Collaboration clinicalstudydatarequest.com SAGE bionetwork Early Breast Cancer Trialists’ Collaborative Group, CTSU Oxford Worldwide antimalarial resistance network, WWARN. Data retrospectively standardised for meta-analysis from academic and industry studies. P. Data Sphere, IMI, IDEAPOINT, CAMD, IMS, NHS Exec, real world data
IPD meta-analysis
Meta-analysis explaining bone density/scan data in patients with osteoporosis Centre for Reviews and Dissemination, University of York, Keele University, DARE database which has systematic reviews of health interventions Cochrane IPD MA Methods group
Objectives
Methodological, patient stratification, treatment regimes, biomarkers, co-morbidities, lifestyle factors

technopolis |group| United Kingdom
3 Pavilion Buildings
Brighton BN1 1EE
United Kingdom
T +44 1273 204320
E info@technopolis-group.com
www.technopolis-group.com

The Wellcome Trust

The Wellcome Trust is a global charitable foundation dedicated to improving health. We support bright minds in science, the humanities and the social sciences, as well as education, public engagement and the application of research to medicine.

Our investment portfolio gives us the independence to support such transformative work as the sequencing and understanding of the human genome, research that established front-line drugs for malaria, and Wellcome Collection, our free venue for the incurably curious that explores medicine, life and art.

Wellcome Trust
Gibbs Building
215 Euston Road
London NW1 2BE, UK
T +44 (0)20 7611 8888
F +44 (0)20 7611 8545
E contact@wellcome.ac.uk
wellcome.ac.uk

The Wellcome Trust is a charity registered in England and Wales, no. 210183. Its sole trustee is The Wellcome Trust Limited, a company registered in England and Wales, no. 2711000 (whose registered office is at 215 Euston Road, London NW1 2BE, UK). SP-6177/XK/03-2015/LP