Towards Open Research

Practices, experiences, barriers and opportunities

October 2016

Veerle Van den Eynden, Gareth Knight, Anca Vlad, Barry Radler, Carol Tenopir, David Leon, Frank Manista, Jimmy Whitworth and Louise Corti

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

UK Data Service

Table of Contents

Suggested citation: Van den Eynden, Veerle et al. (2016) Towards Open Research: practices, experiences, barriers and opportunities. Wellcome Trust. https://dx.doi.org/10.6084/m9.figshare.4055448

EXECUTIVE SUMMARY

This study, commissioned by the Wellcome Trust, investigates researchers' attitudes and behaviour towards open research, examining the sharing and reuse of research data, code, and open access publications, in order to identify practical actions the Wellcome Trust can take to remove or mitigate barriers and maximise the opportunities for practising open science.

More specifically, the study gathered evidence on:

- researchers' views on various aspects of open research
- current practices in open access publishing
- views on future developments of publishing
- current practices and experiences in data sharing and reuse
- barriers to sharing and motivations for making data available
- current practices and experiences in code sharing and reuse
- barriers and motivations for code sharing

Evidence was gathered via an online survey with 583 respondents (25.6% of invited Wellcome Trust grant holders) in July-August 2016, and focus group discussions with 22 participants in early September 2016. Respondents formed an excellent representation of the different categories of grant holders.

Results were contrasted against results from a parallel survey with researchers funded by the Economic and Social Research Council (ESRC), as a group of social science researchers who carry out research within the context of a funder with a mandatory data sharing policy and data infrastructure to support it. Both funders have different open access publishing requirements, with Wellcome having a defined open access repository, Europe PubMed Central and grant holders expected to make their publications available within 6 months; ESRC-funded researchers are required to only make publications open access within 12 months. Also the funding models for open access, via institutional block grants and individual grants, differ. This parallel survey was carried out in August-September 2016 and received 259 responses.

Both surveys provided very detailed and granular quantifiable information to test existing perceptions and knowledge about these topics for a specific research community. The detailed findings can serve as baseline evidence to develop very specific actions targeted at different groups of researchers. Survey data and focus group transcripts have been published via the UK Data service.

General findings

Wellcome Trust-funded researchers are already practicing open research in many ways, as is illustrated throughout this report, either by publishing their writings in open access publications, or by making their research data and code available to the academic community for reuse. Different drivers and barriers are at play. Some aspects of open research apply to all researchers in general, whilst other characteristics are very much determined by research discipline, career stage, the location where a researcher is based or carries out research, and the kind of research methods used and data generated. Open research practices are on the increase, with focus group participants indicating many recent developments in their open research practices, usually with positive experiences, such as open peer review, publishing preprints of papers and sharing code.

Publishing

Where the publishing of peer-reviewed papers is concerned, the key deciding factors that matter to researchers are journal reputation, journal audience, high quality peer review and journal impact factor. The ability to publish papers as open access is less important to researchers in comparison. Still, many Wellcome Trust-funded researchers do publish their papers as open access, thanks to funding provided by the Wellcome Trust, with over 70% of all papers published as open access and a third of researchers publishing all their papers as open access.

The ratio of open access publishing is independent of career stage. Early-career researchers are, however, less likely to use Wellcome Trust funding to cover article processing charges; the same applies to social science researchers, whereas biomedical and clinical researchers are more likely to do so.

In selecting literature to use in research, papers being openly accessible or supplementary data being available that underpin the findings, bear little importance. Instead, content quality, journal and author reputation and institutional subscription are the main deciding factors when choosing papers to consult in research.

With Wellcome having recently launched Wellcome Open Research, a new platform that facilitates publishing a range of outputs, researchers indicate that the principal features of this platform should be open and transparent peer review, all outputs being available as open access, and the cost of publishing to be covered by Wellcome. In addition, researchers would like future publishing systems to have visible reviewer comments, to have a forum for commentary and discussion of papers, to enable data visualisation in papers, and the rapid publishing of preprints papers that can afterwards be submitted to established journals. The priority of desired features is determined by career stage and research discipline.

Data

Half of researchers make research data available so they can be used by other researchers, either as full datasets or as subsets, with each researcher having made available on average four datasets over the last five years. Data are mostly released via institutional and community repositories as open access.

The main reasons for making research data available are funder and journal requirements, it being considered good research practice, to facilitate collaborations, and to enable validation and replication of research. The prime benefits researchers have experienced from sharing their data are new collaborations and higher citation rates. However, most researchers have not experienced any direct benefits from making their data available. Neither have they experienced many bad effects from sharing their data.

The main perceived barriers to sharing data are the fear that data will be misused or misinterpreted, the fear that sharing data can jeopardise future publication opportunities, and the time and effort that are required to prepare and deposit data. The fact that very few people have actually had bad experiences from data sharing shows that these fears are largely unfounded. Positive is the fact that benefits outweigh barriers for most researchers.

Researchers indicate that they would be motivated to make more data available in future if they received extra funding to cover the cost of data preparation, if making their data available would enhance their academic reputation, if they knew how other researchers were using their data, and if data sharing was taken into account in future funding and career promotion decisions.

Reasons for sharing data, experienced benefits, barriers and motivations are determined by a researcher's career stage and discipline.

Researchers reuse existing data, principally to provide background information and context to their research, for research validation, to help develop their methodologies and for new analyses. Levels of data reuse again depend on career stage, discipline and the research methods used. Still, a quarter of researchers have never reused existing data. Data for reuse are mainly obtained from colleagues, repositories or directly from the creator. Important is for data to be from a reputable source, of high quality and well documented. Data being openly and immediately available is less important to researchers.

Code

Code sharing is more in its infancy compared to data sharing. It is being less practised and results in fewer benefits, but is also less problematic for researchers. With two-fifth of researchers generating code in their research, less than half of them also make it available for access and use by others.

Code sharing activities are primarily motivated by a desire to comply with good research practice and enabling other researchers to collaborate and contribute to the work. Direct personal benefits are new collaborations, and publications with accompanying code receiving higher citation rates.

No significant barriers to code sharing exist, other than the time, funding and skills needed to prepare code for sharing, especially due to rapid software changes that makes long-term validity challenging.

Incentives that would motivate researchers to make more code available are recognising code sharing activities in funding and promotion decisions, evidence of code citation by other researchers, additional funding and assistance for preparing code for sharing from institutional or funder staff.

Reasons to share code, experienced benefits and barriers are largely independent of career stage and discipline. Motivations are determined by career stage with early-career researchers being motivated more by career-enhancing incentives.

Code reuse practices are currently limited, with just over a third of researchers having used existing code in their research. The availability of good documentation and the code being available from a reputable source are of prime importance when evaluating code to reuse.

Career stage, research discipline and location matter

In data sharing, the reasons why researchers make their data available, the benefits they may experience, the barriers they perceive to exist and possible incentives to make more data available, depend at times strongly on a researcher's career stage, research discipline and location. In code sharing only career stage has an influence. These differences, together with recommendations to address them, are as follows.

Career stage

Early-career researchers' big priority is clearly their career development and maximising opportunities to advance to a permanent position. This influences their open research views and practices. They show a positive attitude to open access publishing and indeed their ratio of open access publications is not different from that of senior researchers. They also show a positive attitude towards data sharing, with ethical motivations to share data for public health benefits, to respond to health emergencies and to the benefit of research participants as key drivers; but they have in practice made less data available. The fear to lose

future publication opportunities by making their data available plays strongly. For these researchers it is important that open research practices do not jeopardise their career prospects, when high impact papers remain an immensely important criterion in career assessments and funding decisions. Career-enhancing incentives such as co-authorship on data reuse papers, increased citation rates resulting from reuse of their data and the publishing of data leading to the publication of a data paper, can aid encouraging more open research. They also need positive affirmations about open research practices from their supervisors, as supervisors tend to point out that openness may be too risky for their career.

For **established researchers** funding is of primary concern in open research decisions. They do share their data and publish open access. They readily use Wellcome funding to cover article processing charges and are motivated to share more data and code if extra funding were provided to cover the costs.

Research discipline

Biomedical researchers actively practice open research: they share data, reuse data, frequently experience the benefits of data sharing, and report no significant barriers to sharing. Neither do they indicate that extra motivations for sharing are needed. Their research mainly produces quantitative, omics and imaging data.

Clinical researchers follow closely in practising open research. Data are readily shared, with a high likelihood of personal benefits such as increased publication opportunities and new collaborations. Only the fear of misuse of data poses a barrier to data sharing. Their research mainly produces quantitative, omics, imaging and longitudinal data. Data sharing can be increased for this group by having support from funder or institution for data preparation, having the ability to control access to data, and more rewards such as co-authorship on reuse papers.

Population and public health researchers produce mainly quantitative and longitudinal data, perform much secondary analysis in their research and benefit from data sharing: more funding opportunities, more publications and new collaborations. They do experience challenges in data sharing that need addressing, as data preparation takes time and effort, data may be disclosive, contain confidential information, have third-party rights and participant permission for data sharing may be a challenge. Data sharing can be increased for this group by having support from funder or institution for data preparation, having the ability to control access to data and more rewards such as co-authorship on reuse papers. The ESRC data infrastructure and support model may provide solutions to facilitate such support and access controls, as well as guidance on how to facilitate sharing confidential data.

Humanities researchers have very little experience of data sharing and seemingly not much could motivate them to share their data. At the same time they report no significant barriers to data sharing. More often they indicate that their research does not generate or rely on data, being manuscripts, narratives and observations. They may benefit from guidance that the information sources used in their research do constitute research data.

Social science researchers have little experience of data sharing and reuse and perceive minimal benefits from data sharing. Data are frequently qualitative or longitudinal. They do experience challenges in data sharing that need addressing, as data may be disclosive, contain personal and confidential information, have third-party rights and participant permission for data sharing may be a challenge. They also fear misuse or misinterpretation of their data. Data sharing could be increased for this group by the ability to control access to data and support for data preparation and more rewards. The ESRC data infrastructure and support model (via the UK Data Service) may provide solutions to facilitate such support and access controls, as well as guidance on how to facilitate sharing confidential data.

Wellcome funding area

(an alternative indicator for discipline)

Researchers in **Cellular, developmental and physiological sciences** do not share much data, but can benefit from Wellcome showcasing successful examples of data sharing and reuse as well as facilitating researchers knowing how their data may be used to increase data sharing. Also additional funding to cover the cost of data preparation and rewards such as co-authorship and data sharing being considered seriously in funding and promotion decisions can motivate more sharing. They frequently generate imaging data, which due to their large size, lack of sharing standards and practices, and lack of repository solutions are difficult to share.

Researchers in **Genetic and molecular sciences** practice open research very well, making much of their data available as open access and readily sharing code.

Researchers in **Infection and immunobiology** practice open research very well, sharing much data and code and frequently reusing existing data.

Researchers in **Neuroscience and mental health**, who frequently produce imaging data, find data sharing problematic, whilst frequently producing and sharing code. Keen to practice open research, the sheer size of their data files and absence of data sharing infrastructure solutions and sharing practices or standards is problematic. They can benefit from developments of infrastructure, community standards and extra funding to cover the cost of data preparation.

Researchers in **Population health** have challenges in data sharing as data frequently contain confidential information, have third-party rights and participant permission for data sharing may be a challenge. Data sharing can be increased for this group by extra funding to cover the cost of data preparation and promotional activities such as showcasing shared data and how data are being reused. Also the ESRC data support model (via the UK Data Service) may provide some solutions to facilitate the sharing of confidential data.

Researchers in **Humanities and social sciences** have very little experience of data sharing, but can benefit from Wellcome showcasing successful examples of data sharing and reuse as well as facilitating researchers knowing how their data may be used to increase data sharing.

Location

Researchers based in or doing research in low and middle income countries may lack the funding to prepare data for sharing and face challenges because data frequently contain confidential information, have third-party rights. Also lack of participant permission for data sharing and country-specific regulations that prohibit sharing can be problematic. Data sharing can be increased by showcasing shared data and how data may be used by other researchers, and through rewards such as data deposit leading to the publication of a data paper. The ESRC data support model (via the UK Data Service) may provide some solutions such as access controls and guidance to facilitate the sharing of confidential data.

UK-based researchers can be motivated to make more data available in future if this was looked upon more favourably in funding and career promotion decisions.

Comparison with ESRC-funded researchers

Comparing survey findings with those for ESRC-funded social science researchers shows that, whilst in general Wellcome Trust-funded researchers publish more open access papers, this is not the case for Humanities and Social Sciences (HSS) researchers. And whilst in general Wellcome Trust-funded researchers use more Wellcome funding to cover article processing charges for open access publishing, this is not the case for HSS researchers.

Wellcome Trust-funded HSS researchers do not make significantly less data available than ESRC-funded researchers and neither are there differences in the reasons for sharing data or the benefits researchers experience from sharing data. Lack of skills to prepare data for sharing is, however, a more important barrier for Wellcome Trust-funded HSS researchers. The lack of suitable data repositories and the fear to lose publication opportunities are more important barriers for Wellcome Trust-funded researchers in general. Motivations for data sharing are not different.

Wellcome Trust-funded HSS researchers are more likely to never have reused existing research data, compared to ESRC-funded researchers, and are less likely to reuse data for new analysis and replication.

With regards code, Wellcome Trust-funded HSS researchers reuse less code than ESRC-funded researchers, but show no other significant differences in code sharing.

Recommendations

Open access publishing recommendations

Overall, open access publishing seems to work very well across the different categories of researchers. The findings indicate as practical actions the Wellcome Trust can take to promote more open access publishing:

- early-career researchers and social science researchers are groups that can be targeted to encourage them to use Wellcome funding to increase their open access publishing;
- the suggested priority features for Wellcome Open Research (Fig 8) can be used by Wellcome to direct the scope and features of this new platform, in particular an open and transparent peer review process, open access to all published outputs and free to use for Wellcome-funded researchers;
- early-career researchers give high importance to many features of Wellcome Open Research, and could therefore be targeted as pilot audience for initial submissions;
- biomedical and clinical researchers could be targeted to invite submission of a wide range of research outputs and submission of research data that underpin published papers, since they value these features highly;
- additional features that could be important for the platform, as well as for other publishing systems, are publishing reviewer comments in a visible manner, providing a forum for commentary and discussion of papers, data visualisation options and the publishing of preprints.

Data sharing recommendations

Currently only half of Wellcome Trust-funded researchers actually make data available, so making data available in repositories can be promoted more in general. In addition, practical actions the Wellcome Trust can take to encourage more data sharing are:

- training and guidance on data preparation, data management and data sharing, especially to address challenges in data sharing such as for disclosive and confidential data; the action points listed in section 6.7 can be used as the basis for a guidance and training wish list;
- repository infrastructure for Wellcome trust-funded researchers that offers simple methods for deposit, and provides repository solutions for diverse and challenging resource types such as imaging data, confidential data and qualitative data; the best option would be for Wellcome to work with existing well-established repositories to possibly develop bespoke data deposit routes or tools that facilitate deposit;

repositories to work with could be Dryad, figshare or Zenodo for a wide range of open access datasets; the UK Data Service or Dataverse for disclosive and confidential data related to human subjects that requires access controls; for imaging data, community initiatives can be developed or existing initiatives such as euro-bioimaging and Omero supported

- an up-to-date list of suitable existing data repositories for a range of data types, as is for example provided by the Scientific Data journal[1];
- funding and/or support to assist researchers with data preparation and deposit; this can be via support staff that can assist researchers with data sharing activities, or extra funding to individual researchers or labs;
- in low and middle income countries, continued support for community networks to ensure that data skills capacity is built and maintained;
- value and recognise open research practices directly in Wellcome funding decisions, and through awards or by recognising data sharing champions; researchers practising open research can gain career progression benefits from such initiatives;
- show researchers why their data are important to other researchers through networking events for data creators and data users, by showcasing examples of data sharing and success stories of data reuse, and by aiding the development and uptake of resources that allow researchers to easily promote their data outputs within the wider research community and monitor usage of their shared data; ResearchGate was indicated as a good system for such showcasing and tracking;
- clarify requirements on the access level and type of sharing that is expected for specific resource types in the Wellcome Trust data sharing policy.

Code sharing recommendations

The findings indicate that practical actions the Wellcome Trust can take to promote more code sharing are:

- skills training and expert guidance or support on best practice for code development and sharing, tailored to the needs of researchers working in specific domains, e.g. via Software Carpentry events;
- additional funding to help with the process of preparing code for sharing and providing ongoing maintenance over time;
- greater clarity on Wellcome's expectation for code development and sharing;
- considering the setup of a repository to host code produced by and of use to the Wellcome community, e.g. a Wellcome GitLab;
- target early-career researchers via career-enhancing incentives.

[1] http://www.nature.com/sdata/policies/repositories

1. Introduction

This study was commissioned and funded by the Wellcome Trust, and carried out by the UK Data Service and the London School of Hygiene and Tropical Medicine, with input from a wider advisory group. Its aim is to investigate researchers' attitudes and behaviour towards open research, examining the sharing and reuse of research data and code, and open access publishing of papers.

1.1. The Wellcome Trust and open research

The Wellcome Trust, a biomedical research charity, is one of the world's largest funders of scientific research, supporting major initiatives in the areas of biomedical science, population health, product development and applied research, humanities and social science, as well as public engagement. The Wellcome Trust seeks to foster an open research culture, underpinned by policies and funding that aim to make research outputs widely and easily accessible. Its Open Access Policy[2] establishes a mandate for the sharing of research papers, monographs and book chapters; and its Data management and Sharing policy[3] seeks to develop community-led practices for the sharing and citation of research data. Wellcome also contributes to national and international initiatives such as Research Council UK's Common Principles on Data Policy[4] and Concordat on Open Research Data[5].

In practice this means that research papers that have been accepted for publication in a peer-reviewed journal, and are supported in whole or in part by Wellcome Trust funding, are to be made freely available through the PubMed Central (PMC) and Europe PubMed Central (Europe PMC) repositories within six months. Monographs and book chapters are to be made available through PMC Bookshelf and Europe PMC. Wellcome provides supplementary funding to grant holders and block grants to key institutions to cover open access publication charges. Grant applicants are required to develop a data management and sharing plant that outlines plans to make research data openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.

The policies are underpinned by major initiatives to establish and enhance the research infrastructure that supports open research practice. Wellcome Trust worked with the Howard Hughes Medical Institute and Max Planck Society to establish eLife[6], a peer-reviewed open-access scientific journal for the biomedical and life sciences in 2012 and, more recently, it has announced the launch of Wellcome Open Research[7], a common platform for publishing a range of outputs produced by Wellcome-funded researchers, including those that may not otherwise be made available, such as research resulting in null or negative results. Wellcome also contributes significant funding to data sharing infrastructure such as EMBL-EBI.

1.2. ESRC and open research

The Economic and Social Research council (ESRC), the main public funder for social sciences research in the UK, is guided in its approach to open research by the Research Council UK Policy on Open Access[8], the Research Council UK's Common Principles on Data Policy and the Concordat on Open Research Data. These policies are complemented by investments in data infrastructure and support services.

---

[2] http://wellcome.ac.uk/funding/managing-grant/open-access-policy
[3] http://wellcome.ac.uk/funding/managing-grant/policy-data-management-and-sharing
[4] http://www.rcuk.ac.uk/research/datapolicy/
[5] http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/
[6] https://elifesciences.org/
[7] http://wellcomeopenresearch.org/
[8] http://www.rcuk.ac.uk/research/openaccess/policy/

The Policy on Open Access expects peer-reviewed research and review articles to be made open accessible immediately by the publishers, or via a repository within six months for STEM disciplines and within twelve months for papers in the arts, humanities and social sciences. The Research Councils provide block grant funding to universities and eligible research organisations to cover the cost of article processing charges (APCs).

In the mid-1990s already the ESRC adopted a research data sharing policy[9] mandating data sharing as a condition of research funding. Research grant holders are expected to deposit the data that result from their research project with the UK Data Service, to enable their future reuse for research and learning.

ESRC funds its own dedicated data infrastructure for the curation, preservation and access to data through the UK Data Service, whereby different access levels to facilitate controlled access to disclosive and sensitive data are crucial. ESRC also funds data support services to provide guidance, advice and training to data creators, data depositors and data users, with emphasis on the handling consent and anonymization to share confidential and sensitive data resulting from research with human subjects. ESRC also have specific funding streams to promote secondary use of large data assets in the social sciences.

1.3. Data and code sharing

Various studies done in the last decade across different research disciplines and in different geographical areas provide us with much insight into the attitudes of researchers towards data sharing, and the real and perceived barriers towards the sharing of data. Overall researchers are in favour of data sharing, but this willingness to share does not necessarily translate into actual sharing (Wallis et al 2013). Barriers to data sharing are largely social in nature, rather than technical (Federer et al 2015) and deeply rooted in research practices and culture (Tenopir et al 2011), whereby researchers carrying out research with human participants are less likely to share data (Tenopir et al 2015). Common barriers that multiple studies report are fear of competition, the lack of time and funding to prepare data and documentation for sharing, the absence of professional rewards for data sharing, the lack of standards and data infrastructure, ethical and legal constraints, and the fear of misuse or misinterpretation of data (Piwowar 2011, Savage and Vickers 2009, Tenopir et al 2011, Tenopir et al 2015, Youngseek and Stanton 2012).

Barriers to data sharing can be contrasted against Baker's (2016) findings that reproducibility is a major concern to researchers. A survey with 1576 responses found that 70% of respondents have tried and failed to reproduce other researchers' experiments and half failed to reproduce their own research, with selective reporting, pressure to publish and poor statistics or analysis as major factors.

Various studies have also identified determinants and drivers for data sharing. These include perceived career benefits, infrastructure development, the provision of data support services and training, data management skills, funder and journal mandates for data sharing, peer pressure (practices), reuse citation and metrics (EAGDA 2014, Sayogo and Pardo 2013, Tenopir et al 2011, Youngseek and Adler 2015, Youngseek and Stanton 2012). Disciplinary and research group differences in data sharing practices and actual sharing of data are often reported. Van den Eynden and Bishop (2014) found that important motivations for researchers to share their data are when science directly drives the need for data sharing; when data sharing increases the visibility of the researcher; the data sharing cultural norms that exist within a research group, community or discipline; and a framework of policies, infrastructure and data services as external drivers.

[9] http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy/

Specific studies have looked at data sharing practices and barriers in biomedical and health research. Piwowar (2011) reports low rates of sharing biological gene expression microarray data (a type of data generally commonly shared) in human and cancer studies. Federer et al (2015) report informal sharing of biomedical data may be common, but little data are deposited in a repository and barriers are mainly social (Federer et al 2015). Rathi et al (2012) found for clinical trial data a strong willingness to share, but concerns of appropriate reuse, and data mainly being shared upon request. For disease surveillance and public research respectively, Ross (2014) and van Panhuis et al (2014) confirm barriers in sharing data, the first through interviews with various stakeholders, the latter via systematic literature review. Solutions proposed are rewards, data standardisation, capacity-building and tools for sharing; the development of data policies, legal frameworks and data governance agreements; and the building of trust (Barbui 2016, Sane and Edelstein 2015).

The Expert Advisory Group on Data Access in the UK (2014), based on interviews with key stakeholders and a survey with researchers and data managers, recommends as essential incentives for data sharing that research funders should strengthen and finance data management and sharing planning requirements, continue funding and development of infrastructure and support services, recognise high quality datasets as valued research outputs in the Research Excellence Framework, and establish career paths and progression for data managers as members of research teams. They recommend that research institutions should develop clear policies on data sharing and preservation and provide training and support for researchers to manage data effectively. And they recommend that journals establish clear policies on data sharing and processes, with datasets underlying published papers readily accessible, with appropriate data citation and acknowledgement.

We are not aware of similar studies that have investigated researchers' attitudes and practices for code sharing.

2. Objectives and conceptual framework

This study, funded by the Wellcome Trust, investigates researchers' attitudes and behaviour towards open research, examining separately the sharing and reuse of research data, code, and papers. The specific objectives are to study:

- attitudes of researchers to the idea of open research, in particular sharing of data, sharing of code, and sharing of papers;
- current open research practices applied by researchers;
- barriers that inhibit or prevent researchers from practising open research;
- researcher-focused incentives and motivators for practising open research;
- practical actions the Wellcome Trust can take to remove or mitigate barriers and maximise the opportunities for practising open science.

The study focused on researchers funded by the Wellcome Trust and by the Economic and Social Research Council (ESRC), providing in-depth and baseline information on which policy and practice can be developed. Whilst the primary purpose is for the Wellcome Trust to develop its open research strategy, the parallel survey with ESRC-funded social science researchers in the UK provides interesting comparable information, as the latter community has had a mandatory data sharing policy for over a decade and has set up research data sharing infrastructure and data sharing support services to support this policy. This allows comparing the attitudes and practices towards open research of two groups of researchers, both carrying out research with

human participants, but within different contexts of data sharing policy, data infrastructure provision and support services.

3. Methodology

The study consists of two components: an online questionnaire survey performed via Qualtrics, using a combination of structured, coded questions and open-ended exploratory questions; and five semi-structured focus-group discussions with Wellcome-funded researchers (N=22), with attendees selected from the survey and by separate invitation. Three focus groups were held face-to-face in London and two via video conference to accommodate all researchers. Discussions were audio recorded and transcribed. Data have been published via the UK Data Service (Van den Eynden, Knight and Vlad 2016).

3.1. Sample selection

The Wellcome Trust provided us with a list of 2281 current grant holders, representing the various grant types and disciplines they fund. These are researchers holding pre-doctoral studentships, early-career fellowships (postdoctoral), intermediate fellowships, senior and principal fellowships, investigator awards, as well as those receiving project funding, strategic funding, centres and infrastructure funding, equipment and resources funding as well as other personal support. ESRC provided us with a list of 927 current holders of research grants and fellowships. No holders of studentships were included in the list. For both populations, all grant holders in the lists were invited to participate in the online survey, with two reminders sent.

3.2. Survey

The survey instrument (Van den Eynden, Knight and Vlad 2016) was designed based on extensive literature review and input from experts on the project's advisory committee. This means that much targeted questions were asked, based on data and code sharing practices, barriers and motivations well reported in literature, in order to obtain quantifiable agreement or disagreement with this existing knowledge across our population of researchers. The questionnaire (Van den Eynden, Knight and Vlad 2016) contains single-response and multiple-response multiple-choice questions and Likert scales with a 5-scale response mode (ranging from 'not at all important' to 'extremely important'). Each question also provided the ability for respondents to add other options or choices with free text descriptions. The instrument also has four open questions, asking researchers to give their views on the future of publishing and actions Wellcome (or other funders) could take to advance open research:

- In just a few words, what single thing would encourage you to publish more of your work in fully Open Access journals?
- In just a few words, what would you change about the scholarly publishing system if you were able to?
- Overall, what one or two key things could Wellcome do to help you make more data available in a repository or other online form?
- Overall, what one or two key things could Wellcome do to help you make more code available in a repository or other online form?

These four topics were explored in more detail during the focus group discussions.

The survey was available online for a month: the survey for Wellcome Trust grant holders from 14 July until 15 August 2016; the survey for ESRC grant holders from 8 August until 12 September 2016. After the initial invitation, two reminders were sent: one two weeks after the launch of the survey, and the second reminder 4 days before the closure of the survey. For the Wellcome survey we received 583 responses (25.6% response rate). This provides a representative sample at confidence level 95% with 3.5% margin of error. For the ESRC survey we received 259 responses (27.9% response rate). This provides a representative sample at confidence level 95% with 5.2% margin of error.

3.3. Focus groups

Wellcome Trust-funded researchers were invited via the survey to participate in a series of focus group. Twenty-two researchers joined a focus group, providing a good representation of the total population across funding areas, research disciplines, junior versus senior researchers and UK-based versus non-UK (Table 1). Details of participants all available in Van den Eynden, Knight and Vlad (2016).

Topics for discussion during the focus groups (Van den Eynden, Knight and Vlad 2016) were developed from the themes emerging from the survey results, and centred around barriers, incentives and motivations for open access publishing, data sharing and code sharing, as well the what researchers thought the Wellcome Trust could do in practice to make researchers change practices and to advance open research.

3.4. Methods of analysis

Survey results were exported from Qualtrics, with cleaning and coding carried out in MS Excel and SPSS. Analyses were carried out in SPSS. Significant associations between categorical variables from multiple-choice questions were tested through cross-tabulation of variables, with a Pearson's chi-square test done to test dependence of variables ($P < 0.05$). For continuous scale variables such as numbers of papers published and numbers of datasets and code packages shared, ANOVA tests were tests. For Likert scale questions, differences of averages (indicating overall level of importance) were analysed using ANOVA tests.

Three main parameters were used to test for significant differences across respondents: career stage, location and research discipline. For Wellcome-funded respondents, each parameter could be identified in two different ways, based on information provided by Wellcome in the grants database, and information provided by the respondents in the survey. ESRC-funded respondents could be typified via information provided in the survey responses.

Career stage, distinguishing early-career researchers from senior investigators, could be determined via:

- the grant type (grants database), separating early-career researchers (researchers holding pre-doctoral studentships or early-career fellowships) and established researchers (researchers holding any other grant or fellowship);
- the number of years a researcher has been working in research (survey question Q1.4), whereby researchers were grouped in blocks of 0-5 years, 6-10 years, 11-15 years, 16-20 years, 21-25 years, 26-30 years, 30+ years in research.

Location, distinguishing UK-based researchers from those based and/or working in overseas programmes, could be determined via:

- country where the grant holder is based (grants database), distinguishing researchers based in the UK, in low and middle income countries (LMIC) and in non-UK high income countries (HIC)[10];
- country where the grant holder mainly carries out research (survey question Q1.5), distinguishing researchers doing research in UK, in LMIC and in non-UK HIC.

Research discipline could be determined via:

- the funding area (grants database), distinguishing humanities and social science; neuroscience and mental health; population health; cellular, developmental and physiological sciences; genetic and molecular science; and infection and immunobiology;
- the research discipline (survey question Q1.3), distinguishing biomedical scientists, clinical researchers, population and public health researchers, and humanities and social science researchers.

In addition, the two survey groups (Wellcome versus ESRC) were compared, both as entire groups and as groups of humanities and social sciences researchers, to test for differences determined by the data policy and funding model of both funders.

The four questions eliciting free text responses were coded in NVivo. Responses were often so detailed that these provided as much detail as focus group discussion. Focus groups did provide more context to understand the realities of research practices.

4. Characterising grant holders and respondents

4.1. Wellcome Trust-funded respondents (N=583)

Amongst the 583 respondents, 23% are graduate students (MSc or PhD level), 23% are researchers and 49% have a position of lecturer, reader or professor. The survey sample has a good balance between early-career and senior researchers that reflects the ratio amongst grant holders (Table 1). A third are early-career researchers, holding a studentship or early-career grant; and 47% have up to ten years of research experience. The remainder are established researchers with more than ten years research experience (Fig 1).

The research discipline also reflects the population of grant holders, whereby 52% of respondents describe themselves primarily as biomedical researcher, 17% as a clinical researcher, 9% as a population or public health researcher and 16% as a humanities or social science researcher.

The Wellcome Trust funds researchers both in the UK and internationally. Ten percent of grant holders are based abroad, 6% in low and middle income countries (LMIC). Research is carried out in the UK and abroad. A slightly higher ratio of non-UK based researchers responded to the survey: 12.5% of respondents are based abroad and 9% in LMIC. Twenty-four percent of all respondents carry out research abroad, and 16% in LMIC, with as principal countries Ireland, India, Kenya, South Africa and the USA. International researchers are therefore represented well in the survey sample (Table 1, Fig 2).

[10] https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

TABLE 1. CHARACTERISTICS OF WELLCOME TRUST-FUNDED RESEARCHERS AND SURVEY AND FOCUS GROUP SAMPLES

		Wellcome Trust grant holders		Survey respondents		Focus group participants	
		N	%	N	%	N	%
Population		2281	100	583	100	22	100
Location	UK	2060[3]	90.3	445[11] 510[12]	76.3 87.5	17	77.3
	Non-UK HIC	91[5]	4.0	46[13] 21[14]	7.9 3.6	3	13.6
	LMIC	130	5.7	92[15] 52[16]	15.8 8.9	2	9.1
Career stage	Early-career	901	39.5	173[17] 275[18]	29.7 47.2	6	27.3
	Established	1380	60.5	410 308[19]	70.3 52.8	16	72.7
Funding area	Cellular, developmental and physiological sciences	394	17.3	86	14.6	2	9.1
	Genetic and molecular science	340	14.9	70	12.0	4	18.2
	Infection and immunobiology	454	19.9	121	20.6	4	18.2
	Neuroscience and mental health	348	15.3	103	17.7	3	13.6
	Population health	222	9.7	61	10.5	2	9.1
	Humanities and social sciences	300	13.2	87	14.9	4	18.2
	Other	223	9.8	55	9.4	1	4.5
Research discipline	Biomedical scientist	-	-	301	51.6	11	50.0
	Clinician or clinical researcher	-	-	100	17.2	2	9.1
	Population or public health researcher	-	-	53	9.1	3	13.6
	Humanities researcher	-	-	53	9.1	2	9.1
	Social science researcher	-	-	42	7.2	3	13.6
	Other	-	-	34	5.8	1	4.5

Researchers use a range of methodologies, which are strongly determined by the research discipline (Table 2, Fig 3b). Chi-square tests show significance ($p<0.001$) for certain methods being used more in certain disciplines (see Table 2). Across the respondents and with the ability to select multiple responses, 74% use experiments, 33% observations, 33% secondary or meta-analysis, 21% simulations, 21% qualitative methods and 17% surveys. Other methods used include archival research, philosophy, theorising and text mining.

Besides receiving research funding from the Wellcome Trust, responding researchers have in the last five years also been funded by public research funders (56%), charities and not-for-profit organisations (43%), their own institution (41%), industry (15%) and private funders (15%).

[11] Research in UK (question 1.5)
[12] Based in UK
[13] Research outside UK (question 1.5)
[14] Based outside UK
[15] Research in LMIC (question 1.5)
[16] Based in LMIC
[17] Studentship or early-career grant
[18] <=10 years in research
[19] > 10 years in research

FIGURE 1. RESPONDENTS' LENGTH OF RESEARCH EXPERIENCE (N=583)

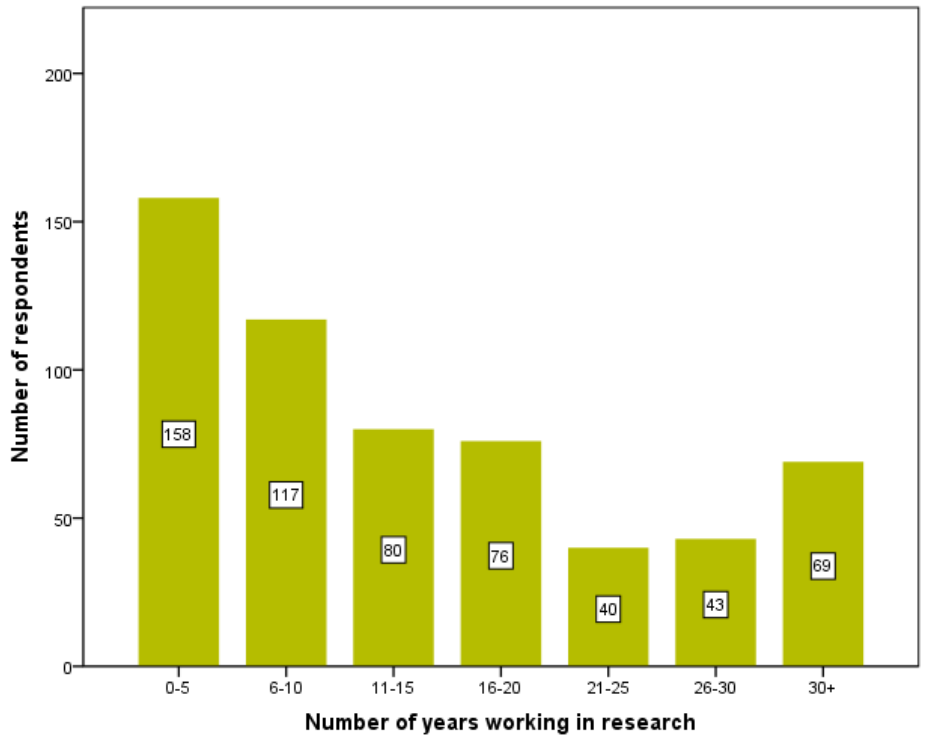


FIGURE 2. INTERNATIONAL REPRESENTATION OF RESPONDENTS (N=583)

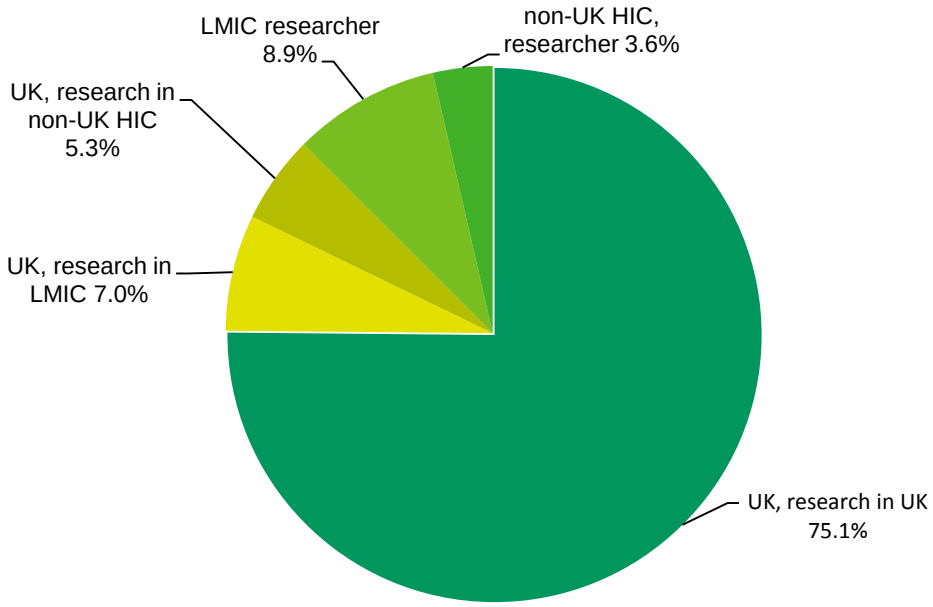
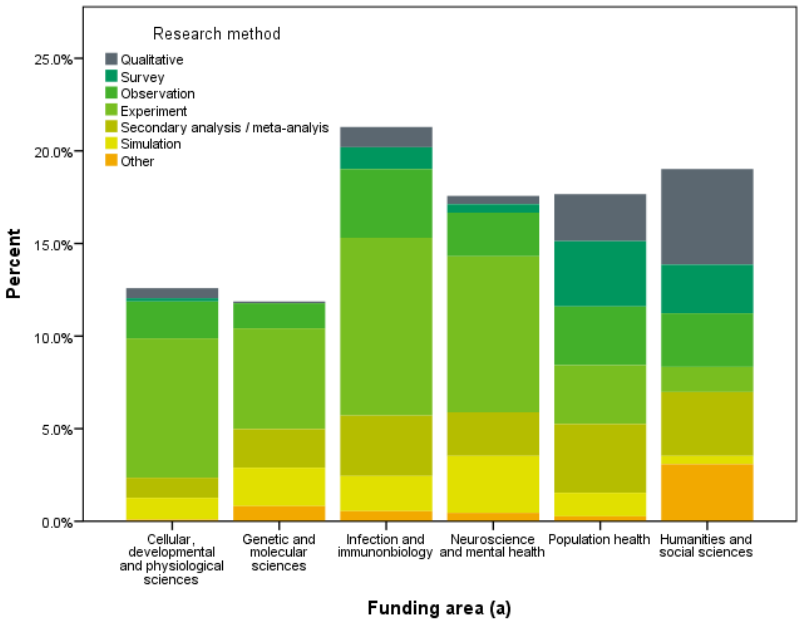
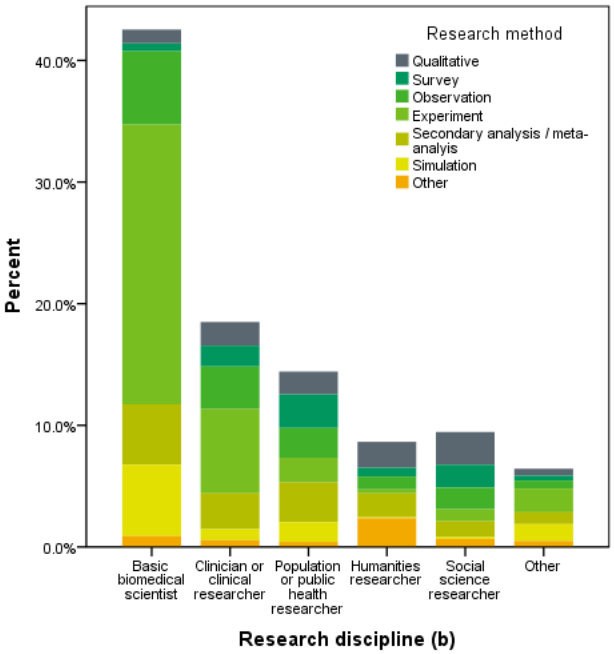


TABLE 2. SIGNIFICANT DEPENDENCIES BETWEEN RESEARCH DISCIPLINES AND RESEARCH METHODS USED BY RESPONDENTS (N=583)

Basic biomedical scientist	Clinician or clinical researcher	Population or public health researcher	Humanities researcher	Social science researcher
experiments ~ simulations	observations ~ experiments	qualitative methods ~ surveys ~ observations ~ simulations ~ secondary and meta-analysis	qualitative methods ~ secondary and meta-analysis ~ other methods	qualitative methods ~ surveys ~ observations

FIGURE 3 RESEARCH METHODS USED BY FUNDING AREA (A) AND RESEARCH DISCIPLINE (B) AS PERCENTAGE OF RESPONDENTS (N=583)





4.2. ESRC-funded respondents (N=259)

The list of current ESRC grant holders did not include studentships, therefore no students responded to the survey. On average ESRC grant holders have been working in research for 19 years, which is higher than amongst Wellcome respondents (Fig 4).

With regards location, 86.5% of respondents carry out research in the UK, 12% in LMIC and 1.5% in non-UK HIC.

Researchers use a range of methodologies, which are strongly determined by the research discipline (Fig 5). Chi-square tests show significance ($p<0.001$) for certain methods being used more in certain disciplines. Across the respondents and with the ability to select multiple responses, 65% use qualitative research methods, 53% carry out secondary analysis, 50% use surveys, 43% use observations, 33% experiments and 11% simulations.

Besides receiving research funding from ESRC, responding researchers have in the last five years also been funded by public research funders (68%), their own institution (57%), charities and not-for-profit organisations (45%), private funders (18%) and industry (10%).

FIGURE 4. ESRC-FUNDED RESPONDENTS' LENGTH OF RESEARCH EXPERIENCE (N=259)

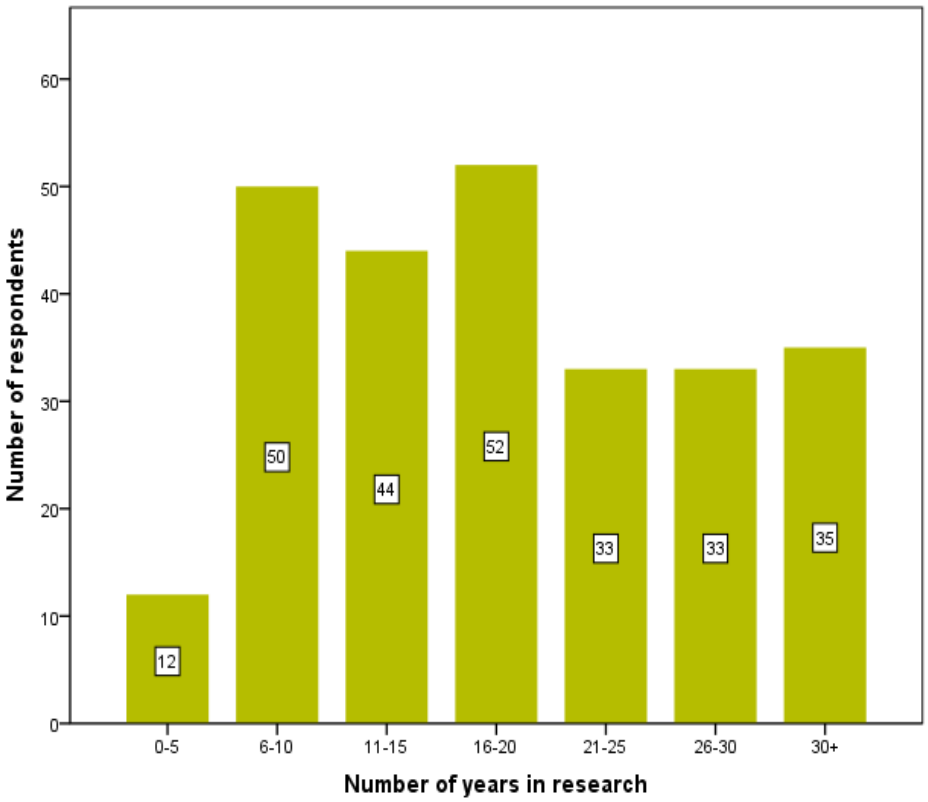
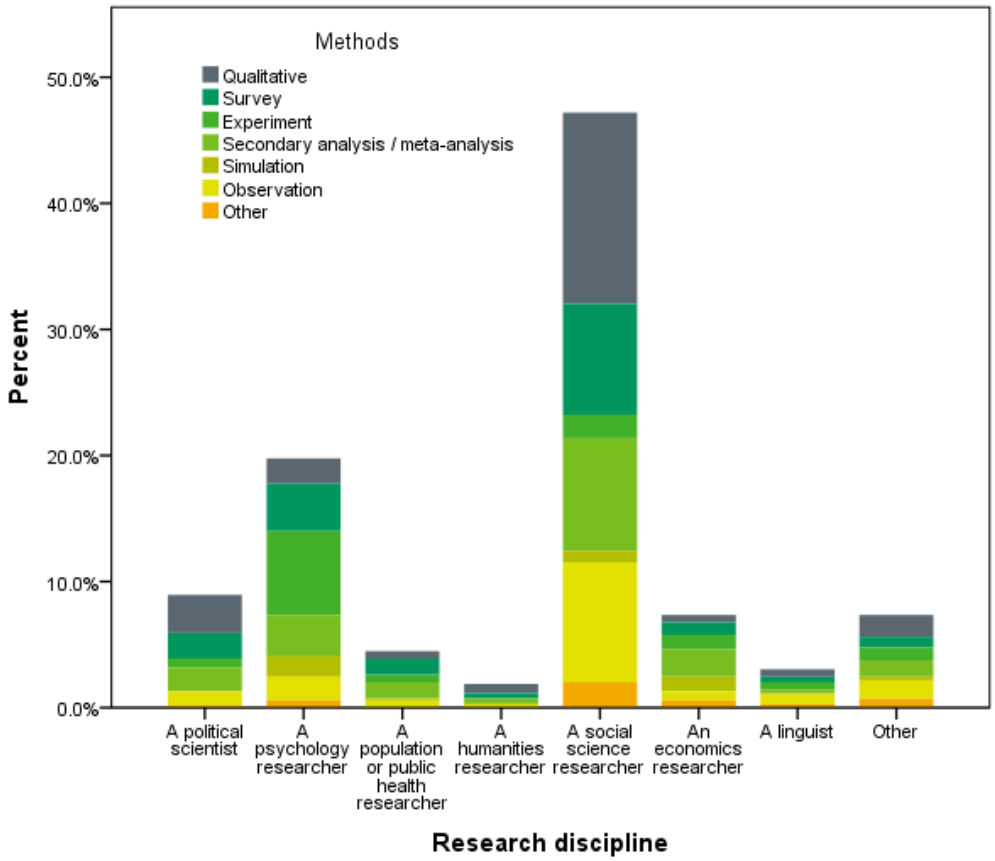


FIGURE 5. RESEARCH METHODS BY RESEARCH DISCIPLINE FOR ESRC-FUNDED RESPONDENTS (N=259)



5. Open Access publishing

5.1. Current publishing practices

Journal reputation, journal audience, high quality peer review process and journal impact factor are key factors when researchers decide in which journal to publish their work (Fig 6). Respondents quote these factors as being very to extremely important on a Likert scale of five degrees of importance. Being able to publish papers for free or at low cost have minimal importance. Cross-tabulations and chi-square tests of the Likert scales scores against the career stage and numbers of years that researchers have being doing research, show as significant dependencies that senior researchers give higher importance to journal reputation and peers publishing in a journal when choosing a journal in which to publish; whereas early-career researchers give more importance to the journal making the article immediately available through open access.

The content quality of a paper is extremely important for researchers when selecting papers to read and use in their research (Fig 7). Also journal reputation, author reputation, and the paper being available through an institutional subscription are very important factors. A paper being available through open access and having access to the data that underpin the paper are only slightly to moderately important.

Over the last 5 years, a researcher has published on average 18 peer-reviewed papers (ranging from 0 to 200), of which 73% are published as open access. This obviously represents papers resulting from all research the researcher has carried out, so not just Wellcome-funded projects. Testing whether open access publishing may depend on career stage shows no significant difference in the average percentage of papers published open access versus the number of years in research (ANOVA test).

Thirty percent of respondents have over the last five years published all their papers as open access papers. When papers have not been published as open access papers, this is due to the journal not having an open access option (31%), lack of funding to cover the article processing charges (30%), because papers are being uploaded to social network platforms such as ResearchGate, Academia, Mendeley (8%) or since the lead author decided against open access (4%).

Fifty percent of respondents have over the last five years used Wellcome Trust funding to cover article processing charges (APC) to publish papers resulting from Wellcome grants as open access. We tested for association with research discipline, career stage and location, and found significant association for career stage and research discipline (chi-square test, $p<0.001$) but not for location. Early-career researchers are less likely to use Wellcome funding to cover APC charges (36%). Biomedical and clinical scientists are more likely to use Wellcome funding to cover APC charges (55% and 57% resp.) and social scientist are less likely to do so (33%). Within funding areas, researchers funded within genetic and molecular sciences and infection and immunobiology streams are more likely to use Wellcome funding to cover APC charges (59% and 57% resp.); researchers funded in humanities and social sciences streams are less likely to do so (46%).

Focus group discussions highlighted that open access papers are crucial for researchers in low and middle income countries (LMIC) to ensure that researchers, policy makers and practitioners have access to up-to-date and relevant information. On the other hand, LMIC researchers may follow different publishing routines than their UK-based collaborators as in the UK publication pressure tends to be higher, demanding more rapid publishing, due to Research Excellence Framework and career requirements. This requires clear agreements in partnerships with LMIC researchers over the timeline of publishing research findings by different researchers.

> *"As much as I love the idea, my long term career prospects currently depend on obtaining high impact papers, so fully Open Access journals have to be of comparable merit."*

FIGURE 6. IMPORTANCE OF FACTORS IN CHOOSING A JOURNAL TO PUBLISH WORK

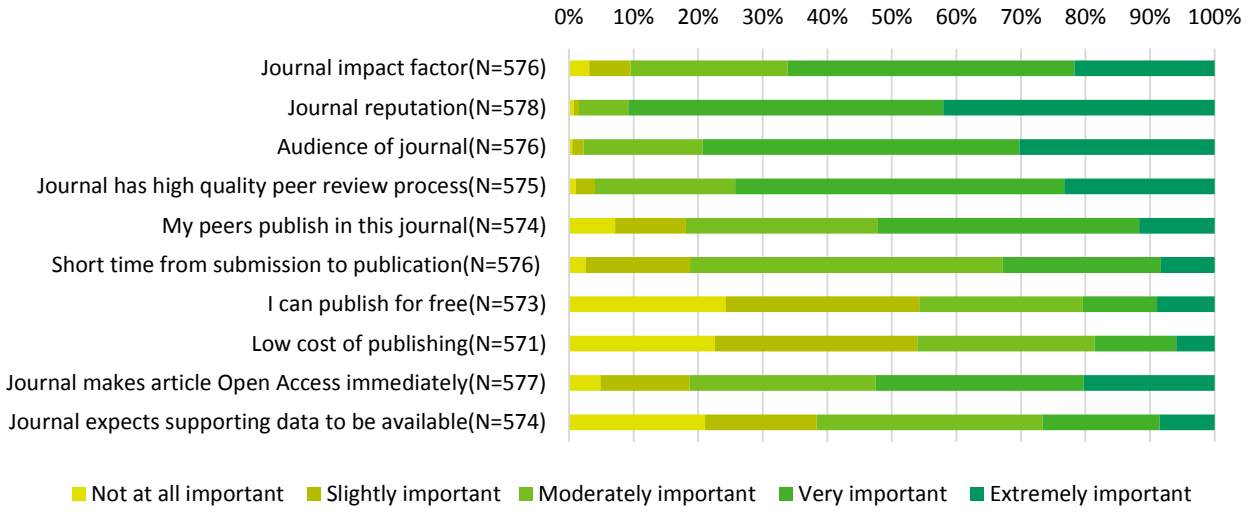
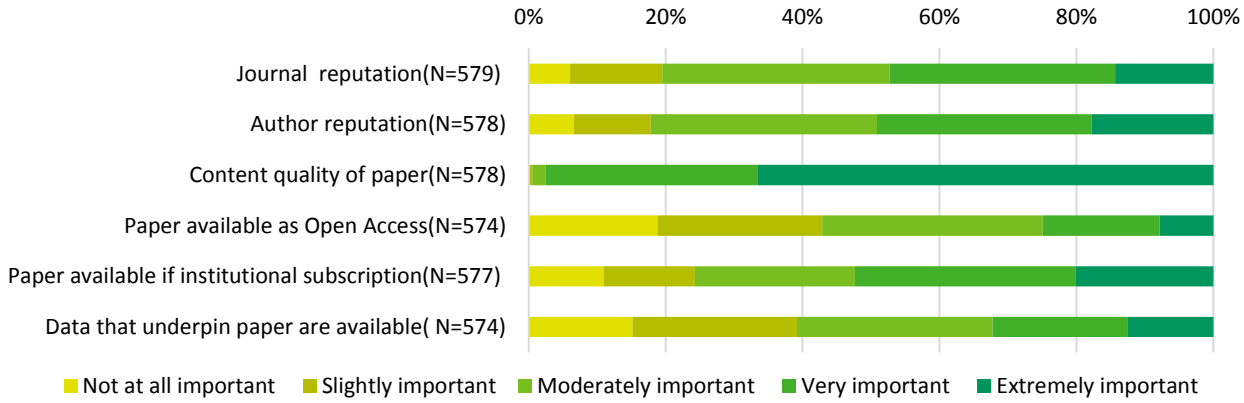


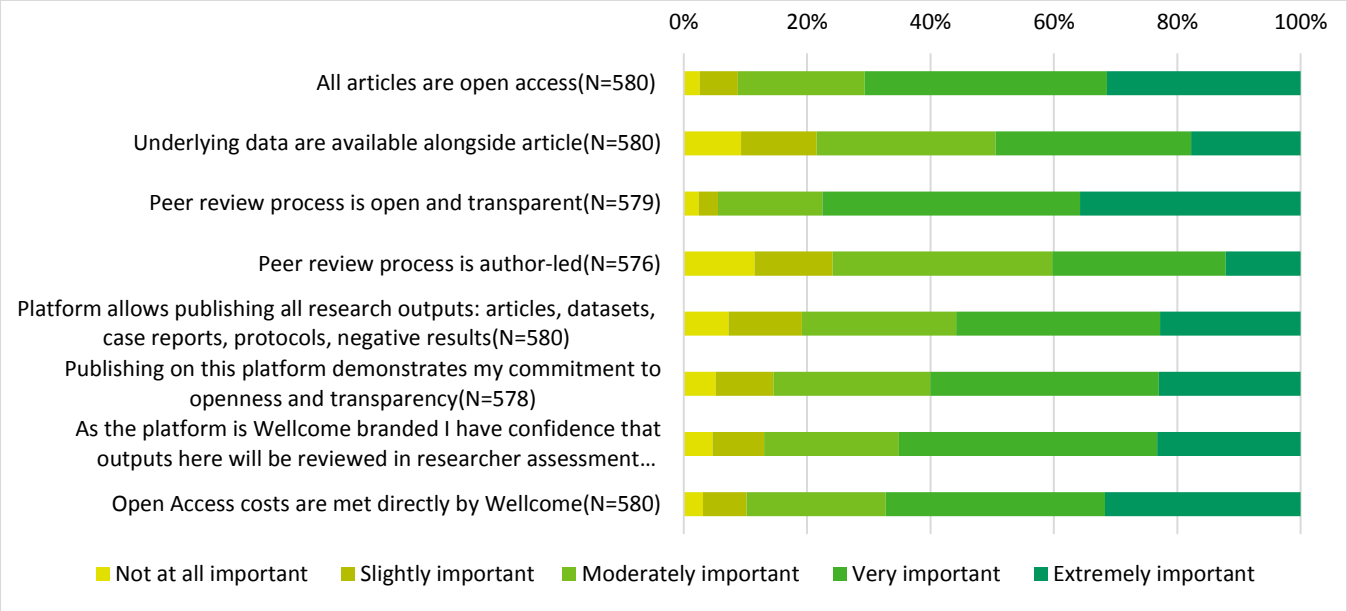
FIGURE 7. IMPORTANCE OF FACTORS WHEN SELECTING PAPERS TO READ AND CITE IN RESEARCH



5.2 Future of publishing

The Wellcome Trust recently launched a new publishing platform, Wellcome Open Research (WOR), which allows researchers to rapidly publish any results they think worth sharing. When asking researchers which features they would value most for this platform, those seen as very to extremely important are an open and transparent peer review process, all articles to be open access, and the costs to be met by the Wellcome Trust (Fig 8). ANOVA tests of the mean Likert scores against research discipline, funding area and career stage show significant dependencies for certain factors, whereby factors are either statistically more or less important for a group of respondents (Table 3). Focus group participants identified outputs they currently publish themselves online, as possibly candidates to publish on WOR in future, such as anthropological reports that provide more detailed context and narrative than papers can.

FIGURE 8 IMPORTANCE OF FACTORS THAT ENCOURAGE RESEARCHERS TO PUBLISH IN WELLCOME OPEN RESEARCH



In general, features that researchers think important for future research publication systems are the ability to see reviewer comments (60 %), a commentary and discussion forum for published papers (51%), data visualisation integrated in papers (47%) and the ability to publish preprint papers online that can afterwards be submitted to a journal (43%). Various focus group participants have positive experiences using bioRxiv for preprints, then submitting papers to established journals. Also figshare and PeerJ preprints have been used for preprints.

TABLE 3. SIGNIFICANT DIFFERENCES ACCORDING TO CAREER STAGE, RESEARCH DISCIPLINE AND FUNDING AREA IN THE LEVELS OF IMPORTANCE GIVEN TO FACTORS THAT WOULD ENCOURAGE RESEARCHERS TO PUBLISH IN WELLCOME OPEN RESEARCH (STATISTICAL SIGNIFICANCE INDICATES WHICH FACTORS ARE STATISTICALLY MORE OR LESS IMPORTANT FOR A PARTICULAR CATEGORY OF RESEARCHERS)

	all articles are open access	underlying data are available	peer review is open and transparent	peer review is author-led	can publish all kind of research outputs	demonstrates commitment to openness and transparency	Open Access costs met by Wellcome
MORE IMPORTANT	early-career researchers	early-career researchers ~ biomedical and clinical researchers ~ genetic and molecular sciences, infection and immuno-biology	early-career researchers	early-career researchers	early-career researchers ~ biomedical and clinical researchers ~ cellular, develop-mental and physiological sciences	early-career researchers	humanities, clinical and population health researchers
LESS IMPORTANT		humanities researchers			humanities researchers		social scientists ~ genetic and molecular sciences , neuroscience and mental health

Based on coded free text responses provided via the survey by 517 respondents to the question “What single thing would encourage you to publish more work in fully open access (OA) journals?”, the principal motivators are open access journals being of high quality and reputation (22%), lower or no publication charge (16%), the publication cost to be covered by funder or institution (13%), career or funding rewards from publishing in such journals (9%) and open access journals having a higher impact factor (7%).

Changes researchers would like to see in the scholarly publishing system, based on 478 free-text responses given to the question “What would you change about the scholarly publication system if able to?”, are principally a more open or transparent peer review system (17%), more rapid publishing (16%) and lower cost (9%).

> *“The open access / data sharing / transparency issues for social science and humanities researchers are very different than the issues for scientific researchers. Wellcome needs to understand those differences and take them into account in these activities”*

5.3. Comparison with ESRC-funded researchers

The contextual differences for both groups of researchers are that Wellcome expects researchers to make their papers, monographs and book chapters open access within 6 months, whilst for ESRC this is within 12 months, and Wellcome has EPMC as defined repository where researchers submit their open access publications; for ESRC they can be submitted to the RCUK Gateway to Research or to an institutional repository.

ESRC-funded respondents publish on average less papers (14 over last 5 years), and publish less papers as open access (59%). Twenty-five percent of respondents have over the last five years published all their papers as open access papers. When papers have not been published as open access papers, this is due to lack of funding to cover the article processing charges (48%), the journal not having an open access option (35%) or since papers are being uploaded to social network platforms such as ResearchGate, Academia, Mendeley, etc. (24%). Only 35% of researchers report having used RCUK institutional block grant funding to cover article processing charges (APC) to publish papers resulting from ESRC grants as open access.

When testing for significant differences between the two groups, we found that Wellcome Trust-funded researchers publish a significantly higher percentage of their papers as open access, compared to ESRC-funded researchers (73 and 59% resp.), but if restricting this only to researchers in the humanities and social sciences (HSS), then the difference is not significant (63% and 59% resp.). Wellcome Trust-funded researchers are more likely to use funding to cover APC, but again the difference is not significant for HSS researchers. Open access publishing and the use of APC is therefore more determined by discipline than by the difference in funding model between the two funders.

Similar as for Wellcome Trust-funded researchers, the most important factors considered by researchers when choosing a journal to publish are journal reputation and audience, followed closely by journal’s impact factor and the quality of the peer review process (Fig 9, versus Fig 6).

Content quality was chosen as the most important factor when choosing papers to read and cite, followed by journal and author reputation (Fig 10 versus Fig 7). What is remarkable, however, is that almost half of the respondents (42.2%) consider that availability of the data underpinning publications is not at all important.

Features that researchers think should be provided in future research publication systems are a commentary and discussion forum for published papers (50%), the ability to publish preprints online that can later be submitted to a journal (41%) and the ability to see reviewer comments (32 %). The latter feature scores significantly lower than for Wellcome Trust-funded respondents. Other features score at a similar level.

FIGURE 9. THE IMPORTANCE OF FACTORS IN CHOOSING A JOURNAL TO PUBLISH WORK

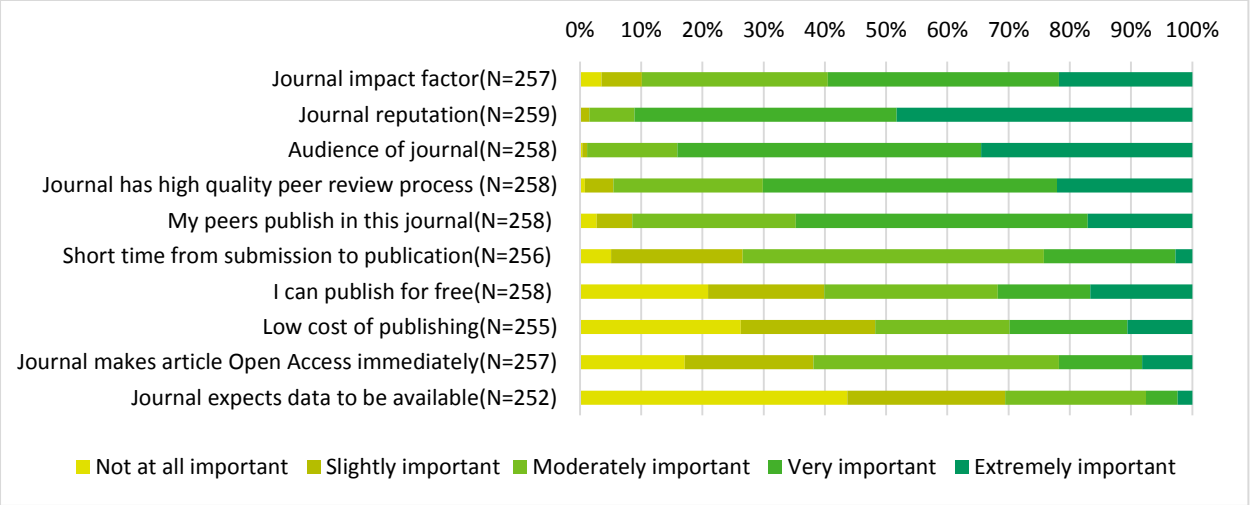
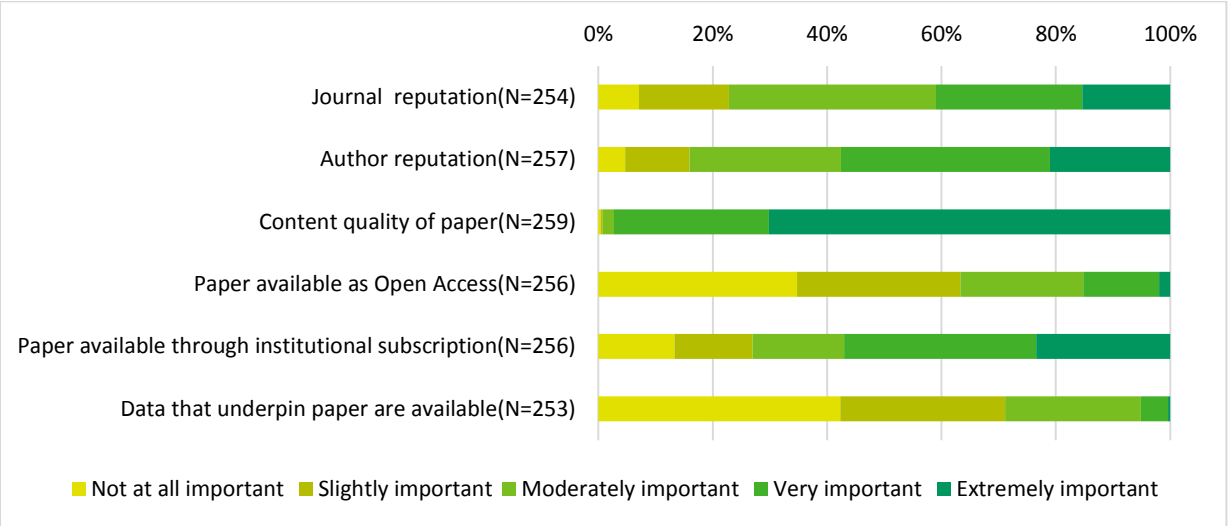


FIGURE 10. THE IMPORTANCE OF FACTORS WHEN SELECTING PAPERS TO READ AND CITE IN RESEARCH



5.4 Actions Wellcome can take

During focus group discussions the following suggestion were made by researchers on actions to take with regards publishing:

- clearer information to grant holders on the availability of open access funds and what it can be used for, for example to publish book chapters, open access books where individual contributors may not be Wellcome-funded;
- in humanities there seems limited choice of open access journals, especially for gold open access; researchers would like a wide list of open access journals in their discipline;
- find ways to enable junior researchers to publish open access when publishing in high impact journals is so important for their career; or provide tailored guidance or options of open access publishing for different career stages;

- Wellcome as an important international funder may have leverage with publishers, in conjunction with other funders, to make strong recommendations to established journals to enable easier open access publishing;
- central database of research protocols (cf. clinical trials protocols).

Overall many positive comments were made about the significant contribution Wellcome has made to advance open access publishing by providing abundant funding for open access publishing and through eLife which has really become a competitor for high impact journals. Participants indicate that if Wellcome wants to push open access publishing more strongly, they could consider mandating that papers can only be published in open access journals.

> *"I would love to see a system implemented whereby any article can be commented by anybody and anybody else can reply to each other's comments …. and comments can be voted as useful or not. Then when trying to read an article one can "download" the comments of other people with a simple (even online) software that allows you to see them inline and after reading one can upload one's comments or criticisms for other people to comment on."*

6. Data sharing and reuse

6.1. Current practices in data sharing

The survey indicates that 95% of respondents generate research data, and 51% of those have made data available to the research community through a repository, data archive, journal, project website, online database or other online form. Each respondent has shared on average four datasets. Note that this sharing excludes informal sharing or sharing upon requests. A range of types of research data are generated, some with characteristics that may pose challenges when sharing data (Fig 11).

Making datasets available is largely independent of the kind of methods used (Fig 12). Chi-square tests between research method and whether data are made available show that researchers using experiments, secondary analysis and simulations are more likely to make data available. For researchers using qualitative methods and surveys there is no significant dependency.

Established researchers make available significantly more datasets than early-career researchers, shown by a one-way ANOVA test between number of datasets made available in the last five years and the number of years the respondent is in research [$F(8, 545) = 2.428$, $p = 0.014$](Fig 13a). There is also a significant correlation between the research discipline and number of datasets made available by respondents, with high numbers of datasets made available in genetics and molecular science and infection and immunobiology [$F(6, 547) = 4.774$, $p < 0.001$] (Fig 13b). There is no significant difference in the number of datasets a researcher makes available and the location of the researcher (UK or abroad).

FIGURE 11. TYPES AND CHARACTERISTICS OF RESEARCH DATA GENERATED BY WELLCOME TRUST-FUNDED RESEARCHERS (N=583)

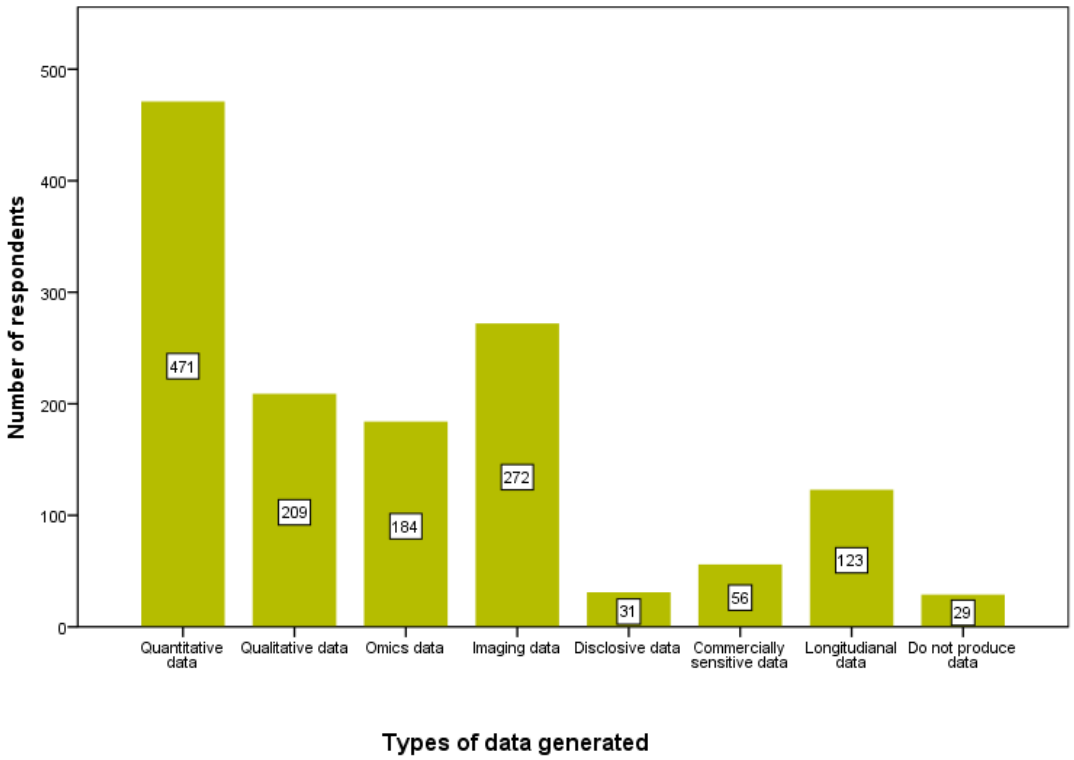


FIGURE 12. DATA SHARING AND CODE SHARING PRACTICES BY RESEARCH METHOD (PERCENTAGES INDICATE THE PERCENTAGES WITHIN THE GROUP OF RESEARCHERS USING THIS METHOD IN THEIR RESEARCH)

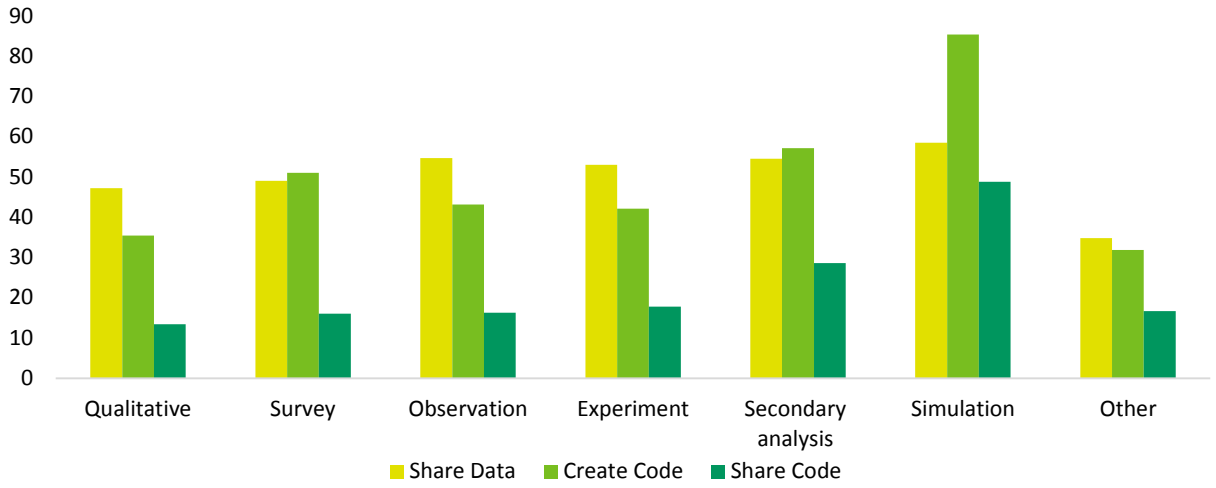
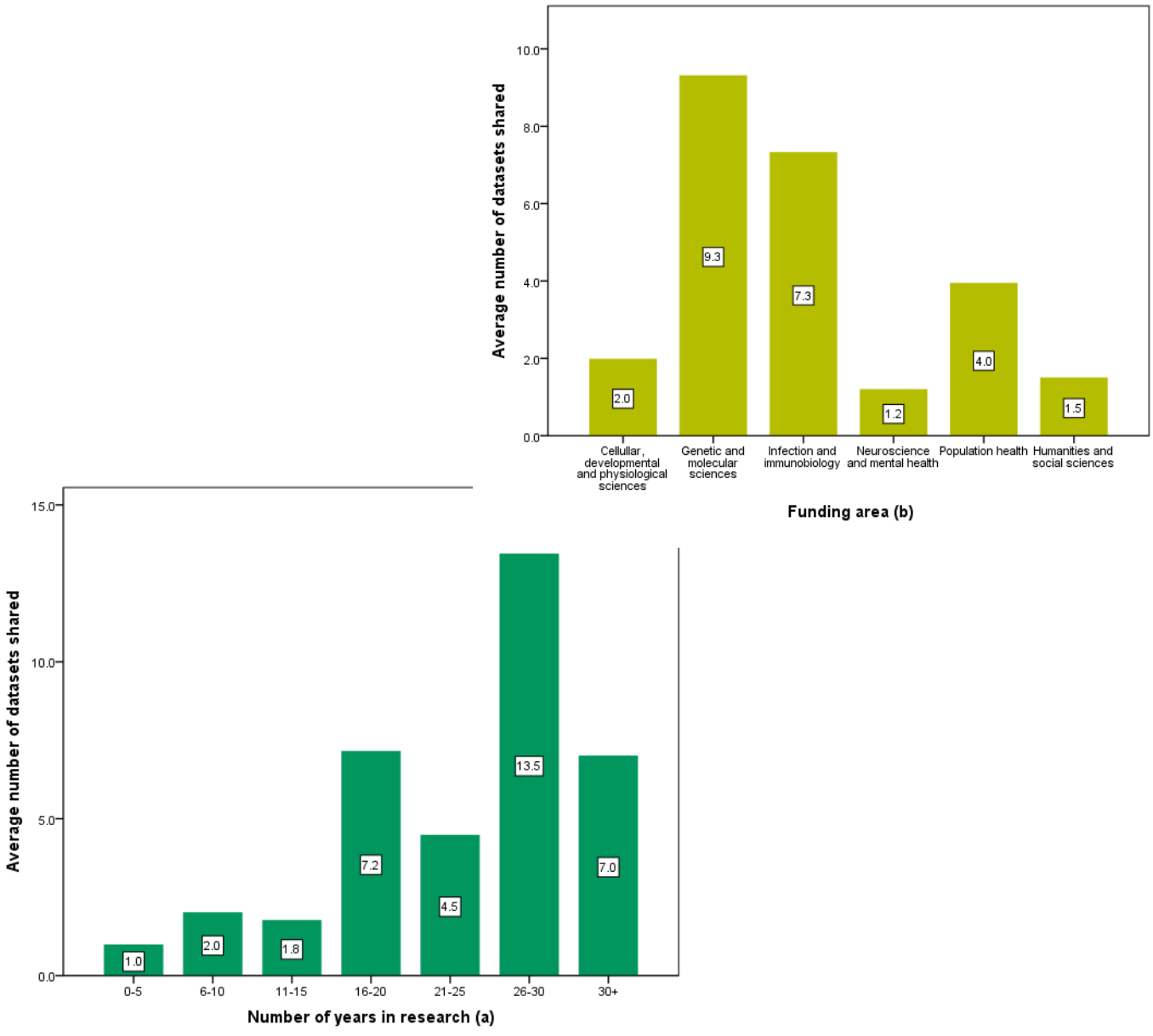


FIGURE 13. AVERAGE NUMBER OF DATASETS RESEARCHERS HAVE MADE AVAILABLE OVER THE LAST FIVE YEARS ACCORDING TO CAREER STAGE (A) AND FUNDING AREA (B)



Amongst the 281 respondents who share data, 48% make datasets available as a full dataset from a research project, 43% as a subset linked to a publication and 33% as a subset. There is no dependence between the amount of data made available (full datasets or subsets) and the research discipline, career stage or location of respondents. There is, however, a significant relation with some research methods used and some types of data generated. Researchers doing surveys are more likely to make available full datasets, and researchers using qualitative methods are more likely to make available subsets. Omics data are more likely to be made available as full datasets or subsets linked to papers. Longitudinal data are more likely to be made available as subsets.

The majority of respondents make datasets available as open access (80%), 19% make data available upon request via an application procedure, 10% restrict access to immediate collaborators and 9% restrict access to registered users. There is a significant relation between the type of data generated and the access level set for shared data: researchers generating omics and imaging data are more likely to make them available as open access. Researchers generating qualitative data and longitudinal data are less likely to make them available as open access, and more likely to restrict access to immediate collaborators or provide access only upon request. Those generating longitudinal data are also more likely to restrict access to registered users. Researchers generating disclosive data difficult to anonymise are more likely to make them available upon request.

Data are made available in community repositories (32%), institutional repositories (45%), generic repositories (15%), and private repositories (12%) and as supplementary materials to journals (10%) (Table 5). Participants point out the limitation that supplementary data in journals may be just small amounts of data linked to graphs in the paper, rather than full datasets, so not necessarily the optimal way to make data available.

During focus group discussions, researchers in LMIC indicated that data may be better published in local and national journals to be of use for policy purposes locally.

TABLE 4. DATA REPOSITORIES IN WHICH RESPONDENTS MAKE DATASETS AVAILABLE, MENTIONED AT LEAST TWICE IN THE SURVEY

Community repositories	**Generic**
RCSB Protein Data Bank (PDB) (N=15)	Figshare (N=11)
Gene Expression Omnibus (GEO) (N=11)	GitHub (N=5)
GenBank (N=8)	ResearchGate
Array Express (N=6)	Dryad (N=2)
European Genome-phenome Archive (EGA) (N=5)	
PRoteomics IDEntifications (PRIDE) (N=4)	
UK Data Service (N=4)	
EMBL European Bioinformatics Institute (N=3)	
EMDataBank (N=3)	
BioModels Database (N=2)	
CARMEN (N=2)	
ChEMBL (N=2)	
European Nucleotide Archive (ENA) (N=2)	
INDEPTH iSHARE (N=2)	
IntAct Molecular Interaction Database (N=2)	
Sequence Read Archive (SRA) (N=3)	
National Center for Biotechnology Information (NCBI) (N=2)	
TriTrypDB (N=2)	
WWARN (N=2)	

6.2. Reasons to share data

The most important reasons for researchers to make their data available are funder requirements, journal requirements, it being considered good research practice, to facilitate collaboration, and to enable validation and replication (Fig 14). Some reasons are strongly determined by career stage and research discipline (proven by ANOVA testing of mean Likert scale scores). Public health benefits, the ability to respond rapidly to public health emergencies and ethical obligations towards research participants are statistically more important reasons for data sharing for early-career researchers. For humanities researchers, all reasons to make data available score very low. Enabling validation and replication, community expectations and good research practice are statistically important reasons for data sharing for biomedical scientists. Journal expectation, public health benefits and the ability to respond rapidly to public health emergencies are statistically important reasons for data sharing for population health scientists. Enabling validation and replication, contributing to academic credentials and journal expectations are statistically important reasons for genetics researchers. An additional reason for making data available, raised during focus group discussions is that making data visible safeguards against fraud accusations (transparency).

The main benefits that researchers have personally experienced from making data available are collaboration (16%) and higher citation rates (14%). Many researchers, however, have not personally experienced any benefit (41%). Benefits are strongly determined by research discipline, and somehow by career stage and location. Biomedical scientists are more likely to experience career benefits and higher citation rates. Population health scientists and clinical scientists are more likely to experience more publications, new collaborations and improvements to public health as benefits. Population health scientists are also more likely to experience more funding opportunities. Humanities researchers are more likely to experience higher citation rates. Social scientists are more likely to not experience any benefits. Researchers in LMIC are more likely to have experienced career benefits, more publications, more funding and improvements to public health from data sharing; and established researchers are more likely to have experienced higher citation rates.

Focus group discussions indicate that immediate sharing of data is often crucial for public health benefits, but the academic career and rewards systems often prohibit this, as papers need to be published first before data can be made available. Some researchers see this as an ethical dilemma in their research, feeling that research should serve society first.

Whilst immediate benefits of data sharing may be perceived as being low, most researchers have not had any bad experiences after making data available to other researchers. Only six percent have had a bad experience: being scooped to publication, incorrect reuse of their data, the time required to supporting reusers, and receiving no acknowledgement from reuse.

6.3. Barriers to data sharing

The main barriers to data sharing are the fear for misuse and misinterpretation of data, the fear to lose publication opportunities and the time and effort required for data deposit (Fig 15). Some reasons are strongly determined by career stage and research discipline, which was substantiated by ANOVA testing of the mean Likert scale scores against those parameters (Table 5). Barriers are clearly less important than benefits for most researchers.

> *"Give much greater academic credit for collection of high-quality data. For example, for clinical epidemiological studies any cretin can download poor quality routinely collected diagnostic data from HES (or similar) and publish over-simplistic and uninformed analyses, but it is a major academic undertaking to collect and slowly come to understand real data."*

FIGURE 14. REASONS FOR MAKING RESEARCH DATA AVAILABLE

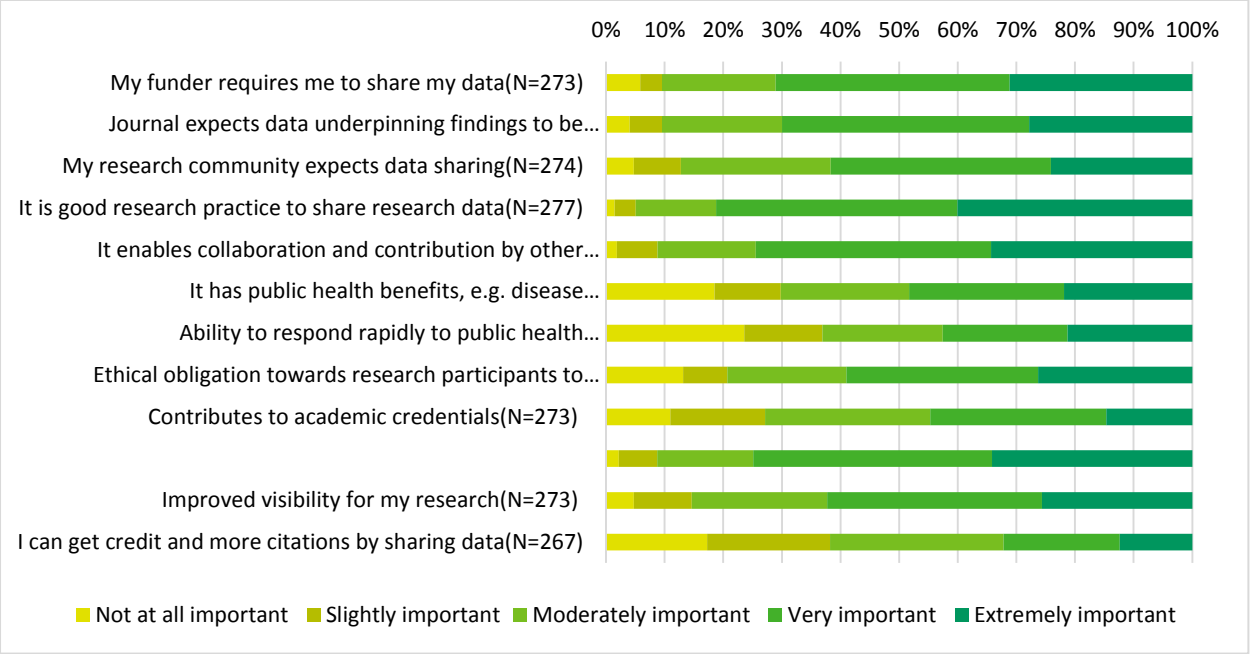


FIGURE 15. MAIN BARRIERS TO MAKING DATA AVAILABLE

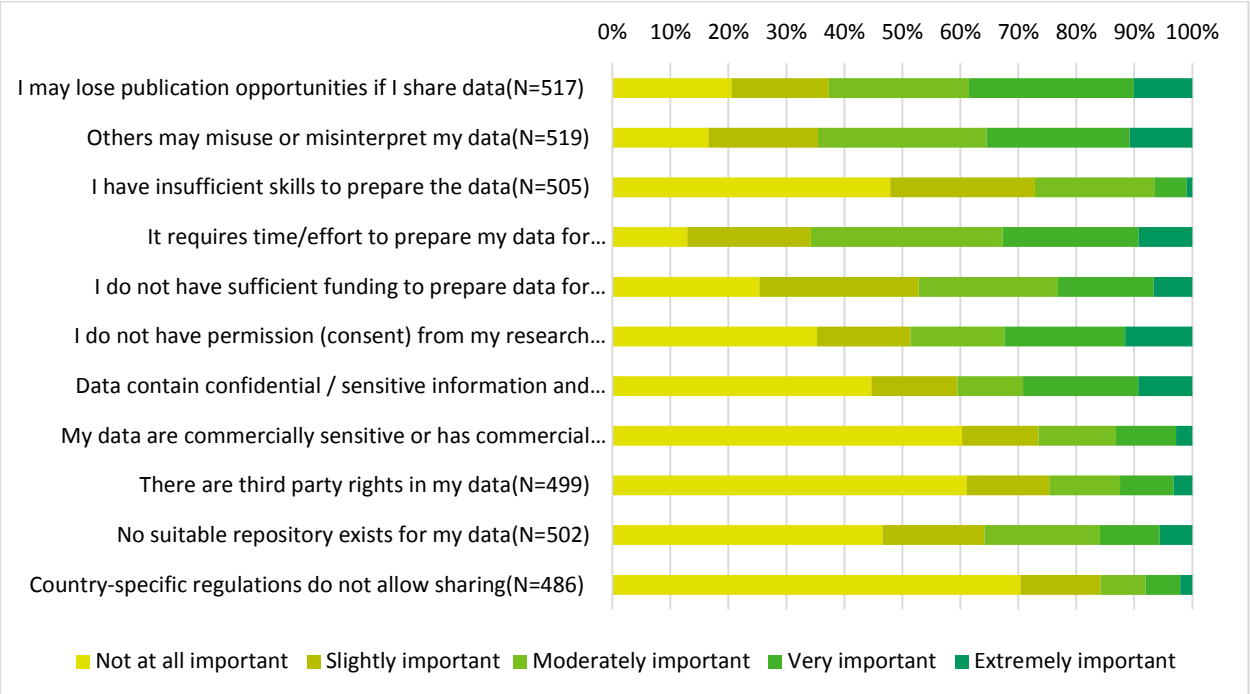


TABLE 5. SIGNIFICANT DIFFERENCES ACCORDING TO CAREER STAGE, RESEARCH DISCIPLINE AND LOCATION IN THE BARRIERS TO DATA SHARING, BASED ON SIGNIFICANCE OF ANOVA TESTS (SIGNIFICANCE INDICATES WHICH FACTORS ARE STATISTICALLY MORE OR LESS IMPORTANT FOR A PARTICULAR CATEGORY OF RESEARCHERS)

	Lose publication opportunities	Fear misuse / misinterpretation of data	Insufficient funding to prepare data	No permission from research participants	Data contain confidential / sensitive information	Third party rights in data	Country-specific regulations prohibit sharing
MORE IMPORTANT	early-career researchers	clinical and social researchers	established researchers ~ LMIC researchers ~ population health and social researchers	LMIC researchers ~ population health and social researchers	LMIC and non-UK HIC researchers ~ population health and social researchers	LMIC and non-UK HIC researchers ~ population health researchers	LMIC researchers
LESS IMPORTANT		biomedical and genetics researchers	biomedical researchers	biomedical researchers	biomedical researchers	biomedical researchers	

6.4. Motivations for data sharing

Main factors that would motivate respondents to make available more of their data are: receiving additional funding to cover the cost of data preparation (63%), data sharing leading to enhanced academic reputation (56%), knowing how other researchers use the data (54%), and if data sharing would be taken into account in future funding decisions and career promotion (53%) (Fig 16).

Motivations show clear statistical dependencies (proved by crosstabulation and chi-square testing) with the researcher's location, funding area, research discipline and career stage (Table 6). This can guide specific actions for the Wellcome Trust to take. Researchers currently not sharing data are not indicating to be motivated by much, except enhanced academic reputation and evidence of citation.

FIGURE 16. FACTORS THAT WOULD MOTIVATE THE RESPONDENT TO MAKE MORE DATA AVAILABLE, AS PERCENTAGE OF RESPONDENTS (N=546)

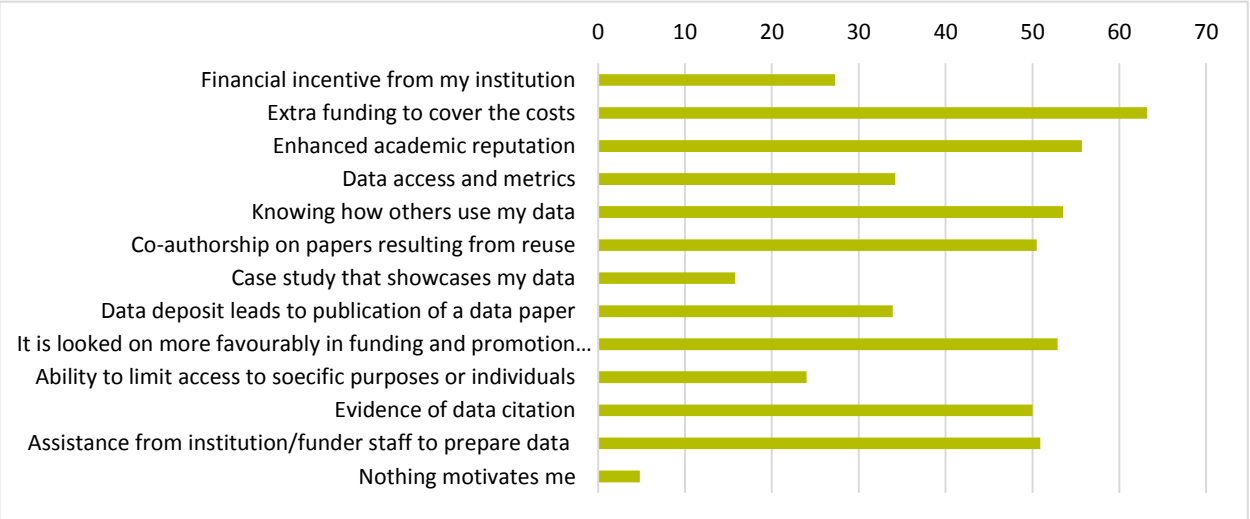


TABLE 6. SIGNIFICANT FACTORS THAT WOULD MOTIVATE PARTICULAR GROUPS OF RESEARCHERS TO MAKE DATA AVAILABLE IN A REPOSITORY (N=546) (SIGNIFICANCE INDICATES WHICH FACTORS ARE MORE OR LESS IMPORTANT FOR A PARTICULAR CATEGORY OF RESEARCHERS)

6.5. Reuse existing data

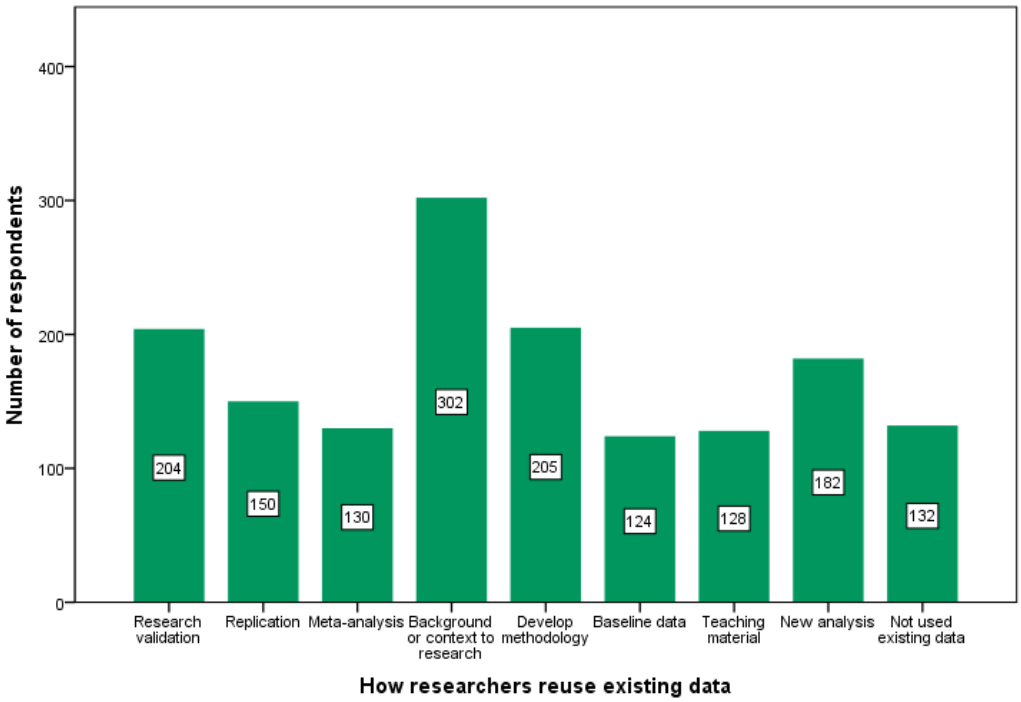
Fifty-two percent of respondents (N=578) have used existing data as background or context to their research, 35% for research validation, 35% to develop their research methodology, and 31% for new analysis (Fig 17). 23% have never reused existing data. Cross-tabulation and chi-square tests show that reuse is very strongly determined by research discipline, strongly determined by career stage and methodology, and to some degree by location (Table 8). Researchers in genetics and molecular science, infection and immunobiology and population health are more likely to reuse data; humanities and social science researchers are less likely to reuse existing data. The kind of reuse is strongly influenced by discipline (Table 7). Early-career researchers are statistically more likely to reuse existing data for replication and to develop methodologies. For established researchers, reuse for teaching material is statistically important. Researchers in LMIC are more likely to reuse data for meta-analysis and as baseline data. For UK-based researchers, replication is statistically more important.

Data are mainly obtained from colleagues/collaborators (49%), community repositories (44%), upon request from the data creator (32%) or from institutional repositories (28%) (N=434). Important aspects in reuse of data is that data are obtained from a reputable source (86%), are high quality (86%) and well documented (76%). Less important are that they are open accessible (55%) and immediately accessible (40%) (N=439).

TABLE 7. STATISTICALLY SIGNIFICANT DEPENDENCIES ACCORDING TO CAREER STAGE, RESEARCH DISCIPLINE AND FUNDING AREA IN HOW EXISTING DATA ARE REUSED (N=578)

	Research validation	Replication	Meta-analysis	Background/context	Baseline data	New analysis	Never used existing data
MORE IMPORTANT	biomedical and clinical researchers - genetic and molecular science, infection and immuno-biology	biomedical and clinical researchers - genetic and molecular science, infection and immuno-biology, neuroscience and mental health	clinical and population health researchers - population health, infection and immuno-biology	biomedical, clinical and population health researchers - cellular, develop-mental and physiological sciences, infection and immuno-biology, population health	population health researchers - infection and immuno-biology, population health	population health researchers - genetic and molecular science, population health	humanities and social sciences
LESS IMPORTANT	humanities and social sciences, neuroscience and mental health	humanities and social sciences	genetic and molecular science, humanities and social sciences, neuroscience and mental health	humanities and social sciences, neuroscience and mental health	humanities and social sciences, cellular, develop-mental and physiological sciences	biomedical and humanities researchers - humanities and social sciences, cellular, develop-mental and physiological sciences	clinical and population health researchers - genetic and molecular science, population health

FIGURE 17. REUSE OF EXISTING DATA BY RESPONDENTS (N=578)



6.6. Comparison with ESRC-funded researchers

The different contextual background for both groups is that ESRC has a research data policy that mandates data being deposited with the UK Data Service within 3 months. This is supported by data infrastructure where all types of datasets created can be deposited (UKDS), that also has different access levels to facilitate controlled access to disclosive and sensitive data. In addition UKDS staff support researchers with preparing and depositing data, as well as guidance and training in data management and data preparation, specialising in handling consent and anonymization to share confidential and sensitive data resulting from research with human subjects. ESRC also have specific funding streams to promote secondary use of large data assets in the social sciences. Wellcome has a data policy that advocates data sharing with suggestions for community repositories [20] where researchers can deposit certain data types (e.g. genome, protein and microarray databases, UKDS), with funding provided to facilitate sharing.

Ninety-four percent of ESRC-funded respondents generate data, and 55% of them make data available, which is comparable to Wellcome Trust-funded researchers. Overall significantly less datasets are shared, with an average of two datasets over the last 5 years (versus 4 dataset for Wellcome Trust-funded researchers). However, when comparing humanities and social sciences researchers, then Wellcome Trust-funded researchers make less dataset available, on average 1.2 over the last five years, but this is not statistically significant. Similar to Wellcome Trust-funded researchers, career stage has a strong influence, and with more datasets made available the longer the person is in research (Fig 18 versus Fig 13).

More often datasets are made available as full datasets (55%), yet significantly less are made available as open access (66%), compared to Wellcome Trust-funded researchers (80%). The difference is not significant when considering just humanities and social science researchers. Access may be given upon request via application (28%) or restricted to registered users (18%). Significantly more qualitative research data are created by ESRC-funded researchers (67% of respondents).

[20] https://wellcome.ac.uk/funding/managing-grant/data-repositories-and-database-resources

Data are made available in community repositories (28%, half of these at the UK Data Service), institutional repositories (52%), generic repositories (10%) or private repositories (20%). More are made available in private repositories compared to Wellcome Trust-funded respondents.

The most important reasons for researchers to make their data available are funder requirements, it being considered good research practice, to facilitate collaboration, to improve visibility of research and to enable validation and replication (Fig 19 versus Fig 14). This is very similar to the reasons stated by Wellcome Trust-funded researchers. Indeed, when testing for significant differences (ANOVA) in the Likert scale scores between humanities and social sciences researchers funded by Wellcome or ESRC, no significant differences are found.

FIGURE 18. AVERAGE NUMBER OF DATASETS RESEARCHERS HAVE MADE AVAILABLE OVER THE LAST FIVE YEARS ACCORDING TO CAREER STAGE (N=243)

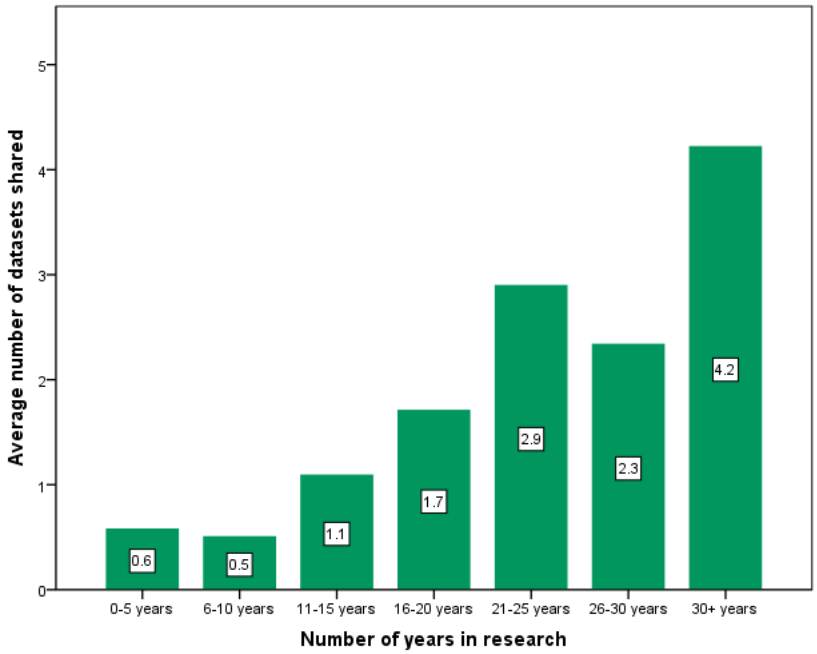
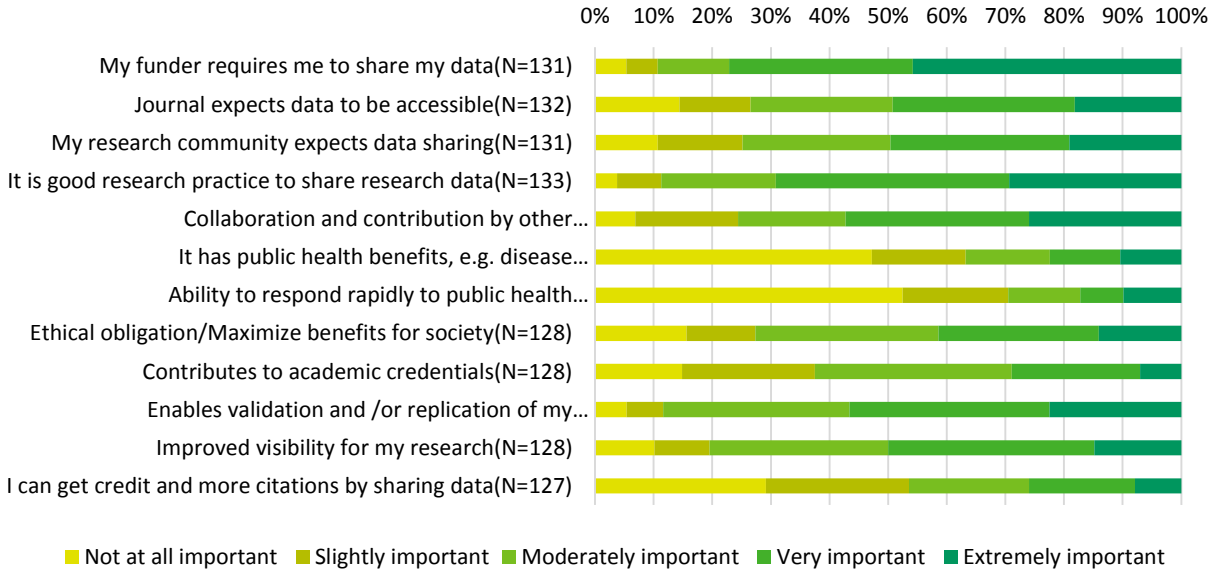


FIGURE 19. REASONS FOR ESRC-FUNDED RESEARCHERS MAKING RESEARCH DATA AVAILABLE



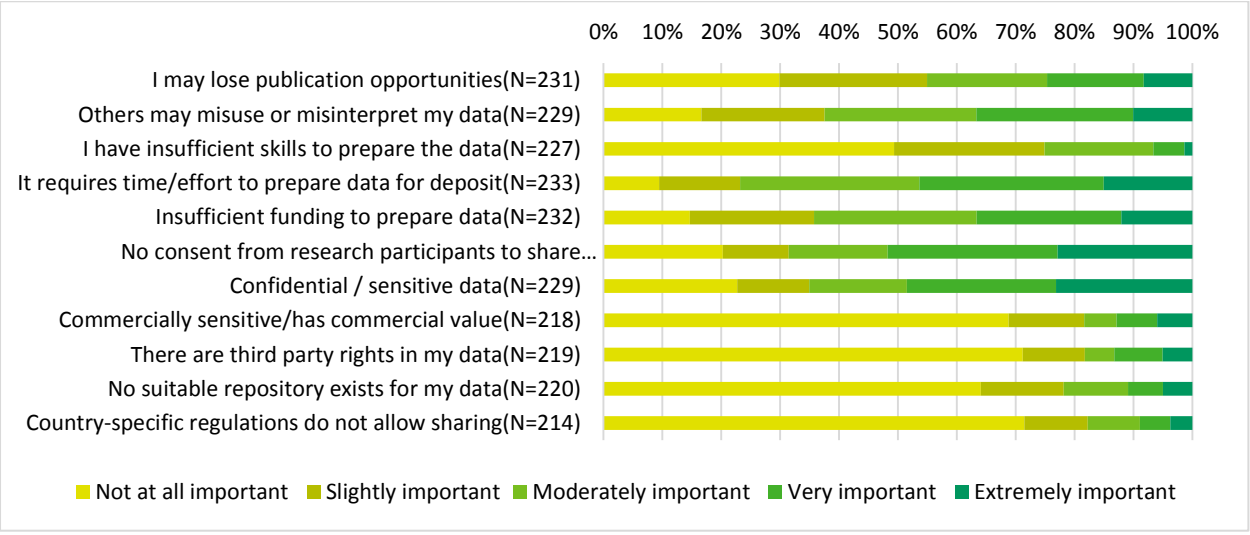
The main benefits that researchers have personally experienced from making data available are collaboration (27%) and higher citation (22%). These percentages are significantly higher than for Wellcome Trust-funded respondents. Many researchers, however, have not experienced any benefit (53%); also this is higher compared to Wellcome Trust-funded respondents. When testing for significant differences (crosstabulation and chi-square testing) in the benefits between humanities and social sciences researchers funded by Wellcome or ESRC, no significant differences are found. When testing between the entire groups covering all disciplines, then career benefits and new collaborations are statistically more important benefits for Wellcome Trust-funded researchers; and they are less likely to not have experienced any benefit from data sharing. Again, most ESRC-funded researchers have not had any bad experiences from making data available to other researchers.

Main barriers to data sharing are not having research participant permission to share data and data containing confidential / sensitive information (Fig 20 versus Fig 15). When testing for significant differences (ANOVA) in the mean Likert scale scores between humanities and social sciences researchers funded by Wellcome or ESRC, only one statistically significant difference is found: the lack of skills to prepare data for sharing is a more important barrier for Wellcome Trust-funded researchers. When testing between the entire groups covering all disciplines, then barriers which are more significant for Wellcome Trust-funded researchers are the loss of publications opportunities and the lack of suitable repository for data.

Main motivations to share more data are funding to cover the cost for data preparation (74%), assistance from institutional/funder staff to prepare data (60%), knowing how others use the data (58%), evidence of data citation (52%) and co-authorship on papers resulting from reuse (48%). These motivations are very different from Wellcome Trust-funded respondents. When testing for significant differences (crosstabulation and chi-square testing) in the motivations for data sharing for humanities and social sciences researchers funded by Wellcome or ESRC, then enhanced academic reputation is a much stronger motivator for Wellcome Trust-funded researchers, whilst additional funding to cover the cost of data preparation is for ESRC-funded researchers.

Reuse of existing data is primarily for background or context of research (48%), for new analysis (35%) and in teaching materials (32%). When testing for significant differences (crosstabulation and chi-square testing) in how existing data are reused by humanities and social sciences researchers, then researchers funded by Wellcome are significantly less likely to reuse data for replication and new analyses, and more are likely to never have reused existing data. ESRC-funded researchers are more likely to reuse data as teaching material.

FIGURE 20. MAIN BARRIERS TO MAKING DATA AVAILABLE FOR ESRC-FUNDED RESEARCHERS



Data are mainly obtained upon request from the data creator (48%), from colleagues (47%) and from the UK Data Service (41%). The same aspects are deemed important in reuse of data as for Wellcome Trust-funded respondents: data are obtained from a reputable source, are high quality and well documented. Data being openly accessible and/or immediately accessible are less important factors for researchers.

> *"Have an honest debate about why authors can't / won't share data: when it has taken > 5 years to gather precious clinical data you don't want to "give it away" when you have ongoing projects using the data."*

6.7. Actions Wellcome can take

When asked what the Wellcome Trust can do to help researchers make more data available, 394 respondents provided free-text responses in the survey, totalling 426 different suggestions. These were coded and could be grouped into six main areas on which the Wellcome Trust could focus. Some suggestions were very generic (e.g. more funding) and were coded to the main topic, whilst other responses provided detailed suggestions. The numbers indicate the number of respondents whose survey responses corresponded to each suggestion. These same areas were explored more during the focus group discussions, with suggestions made during these sessions included in the discussion below.

Guidance, training and support (N=125)

Additional staff support was considered critical to assist researchers in producing and curating high-quality data. In particular there was a recognised need for support related to data preparation and curation (N=51) and data collection (N=2).

There was a broad need for training and guidance on how researchers should prepare data for sharing (N=15), with specific training needs identified on:

- standards and procedures for documenting data, de-identifying content and choosing file formats (N=21);
- identifying suitable repositories for data and guidance on the benefits and disadvantages of each (N=12);
- balancing legal and ethical obligations with the need to make data available (N=8); in particular, it was recognised that there was a need for:
    - engagement with ethics committees to ensure they are aware of sharing obligations when making recommendations on data retention/destruction, particularly related to clinical and personal data;
    - community guidelines on sharing of cohort data which can be difficult to share due to complex data ownership (multiple researchers, institutions, participants) and challenges to obtain all permissions;
- standards that should be applied to ensure data can be easily reused (N=4);
- how to share data that is impossible to de-identify (N=4);
- improved guidance on identifying data that should be kept and defining timescales for data sharing, which address researchers' needs to make full use of their data in publications , in addition to existing community/disciplinary practice considerations;
- best practice for providing credit for data reuse (via citation, co-authorship and other activities).

Capacity building in data management and data documentation at grassroots level was noted as being particularly important for data producers working in Low and Middle Income Countries, who often lack access to resources. Focus group discussion also identified the need to consider data managers' professional career paths and provide guidance on how data could be used for national policy development in LMICs.

Researchers creating qualitative data pointed out during survey comments and during focus group discussions that they consider sharing their data very problematic. However, Wellcome could point to the extensive guidance and expertise the UK Data Service has developed in this area.

During discussions it was clear that much expertise on standards and best practices exists within the research community, so Wellcome can take a role of facilitating debate and development of community standards and practices.

Infrastructure (N=106)

Respondents would like a repository infrastructure for research data which:

- is simple-to-use for data sharing purposes (N=56);
- offers storage that is free (N=22) and suitable for large datasets (N=2);
- allows depositors to define access controls for personal and sensitive data (N=7);
- ensures data held in the system is high-quality (N=5);
- allows the sharing of negative findings and wider results (N=5);
- provides functionality specific to certain resource types, such as imaging data (N=4);
- is developed in collaboration with or built upon existing services, such as Figshare (N=2);
- offers a discovery service suitable for locating existing data (N=5).

Survey and focus group participants recognise that several data repositories appropriate to their data exist and that these perform a valuable role in collating data and metadata; but limitations were noted due to their dependence on short-term funding. In addition, more bespoke repository solutions such as for sharing imaging data, large datasets, sensitive data and negative findings are lacking. It was suggested that resources could be allocated to further infrastructure development for specific resource types, for example via pilot projects. Many institutional laboratories currently accept responsibility for storage of these resources, but do not feel they have the capability or funding to share them in accordance with community needs.

Funding (N=93)

These need for additional funding can be grouped into three categories.

Funding to researchers and / or at institutional level for the preparation and deposit of research data (N=45).

Data preparation was recognised as an important activity to enable data to be shared, but it was often time-intensive to perform due to the need to develop specialist skills and perform relevant tasks. This forces the researcher to spend less time on the research itself. It was suggested that additional data management staff could be funded to work with one or more projects and handle data preparation and sharing activities on their behalf. These staff members could be located at the institution, the funder, or some other organisation.

This topic was discussed in detail during the focus groups, where it was suggested that a dedicated funding stream could be setup specifically to cover data (and code) preparation and deposit, similar to the approach currently taken for open access publishing. Participants feel that open access publishing works very well for Wellcome Trust-funded researchers, since Wellcome provides separate funding to cover APCs, rather than expect projects to allocate it from their existing research budget. If a separate funding stream existed to cover the cost of preparing data for sharing (e.g. via dedicated staff and servers to handle this), researchers would not feel that such expenditure takes funding away from the research

Fund repository development and support networks (N=10) to address the current lack of suitable repositories for specific data types. Suggestions made was that for example to address the lack of data sharing systems for imaging data, many academics labs would benefit from the setup of a OMERO (OME Remote Objects) enterprise server to enable them to manage, analyse and visualise biological image data.

Specifically for Low and Middle Income Countries, the need to support more data management and sharing networks was raised. It was noted that networks such as the ALPHA Network, Human Heredity and Health in Africa (H3Africa), INDEPTH Network, and WorldWide Antimalarial Resistance Network (WWARN) perform an essential role in supporting researchers working in LMICs who would otherwise be isolated, and make a difference in facilitating data standardisation and sharing.

Finally, it was suggested that funding could be allocated to projects that produce or enhance research data suitable for reuse (N=4). This may take the form of small grants for innovative data sharing activities, to enhance existing data held by a longitudinal study and make it suitable for reuse, to maintain well-used resources produced by a previously completed project, or to provide support for specific data reusers, such as those based in LMICs. This was equally elicited during focus group discussions.

Rewards (N=53)

Respondents propose that Wellcome should encourage broad changes in the research community to ensure data sharing activities are recognised as part of career progress evaluation (N=6). This should be built upon a granular approach, which provides additional academic credit for the creation and sharing of high quality data (N=6).

It was also suggested that Wellcome can actively recognise open research practices in their funding decisions and take this information into account when making funding decisions (N=11), for example by asking grants applicants to provide details of data they have made available in the past, how they are being used, and who is using them.

Finally, it was suggested that Wellcome should extend their policy on ensuring data reuse is recognised in research publications by facilitating opportunities for data creators to become co-authors on new publications based upon their data (N=10), implement additional measures to ensure data is correctly cited (N=7) and that appropriate usage credit is provided in the main body of the paper (N=3).

> *"Wellcome could commit to increase visibility of studies which share data but it could also highlight 'success stories' on what is being done with the shared data showing that data sharing is advantageous also for the original researcher beyond a simple citation on a paper"*

Policy (N=35)

The current Wellcome data management and sharing policy was considered to have an important role in encouraging greater sharing of research data, with several respondents expressing a desire for data sharing to be mandated for all research data (N=14). However, there was concern that a blanket policy could be challenging to implement for specific types of research output. Discussion during the focus groups revealed a strong view that qualitative data cannot be shared openly and data sharing expectations must be developed with consideration of the capabilities of the available infrastructure in mind. It was proposed that Wellcome's guidance should be developed further to provide explicit statements on the level of sharing that is expected for specific resource types (N=19), in particular where data sharing is currently seen as problematic, such as for imaging data, social science data, patient-related data and anthropology data. Pilots could be setup to trial it with specific projects, and guidance for the sharing of problematic social science, humanities and patient-related data can be obtained from expertise developed by the UK Data Service.

A related point noted was that many researchers find it difficult to share research data due to conflicts between government policy and Wellcome's data sharing expectations. It was suggested that Wellcome could work with several funding agencies to open a dialogue with host nations and develop government-level sharing agreements that can be applied for research (N=1).

> *"We would find it very useful if there was some kind of 'marriage bureau' whereby Wellcome could introduce collections to potential researchers .... I find it difficult to know how to present our research collections to the academic community."*

Promotion (N=35)

Actively promoting the benefits of making research data available and showcasing best practice in this area was also indicated as an action point. Wellcome can facilitate the tracking of reuse of data, similar to that applied by ResearchGate for downloads (N=8), showcase projects and researchers that are currently sharing research data and make them visible (N=11), and promote success stories of data reuse (N=2).

Wellcome can facilitate networking events for data creators and reusers to bring the two groups together (N=5) and highlight to researchers that their data would be of use to others (N=3).

7. Code sharing and reuse

7.1. Current practices in code sharing

The survey indicated that 40% of respondents (N=236) generate code in their research, of which 43% (102 respondents) have made code packages available for access by others in the last five years, a significantly smaller number of people compared to those that make data available (51%).

Creating and sharing code is strongly determined by the kind of methods used (Fig 12), and the research discipline (Fig 21). Chi-square tests between research method and whether code is created and/or made available show that researchers using surveys, secondary analysis and simulation are more likely to create code. Those using secondary analysis and simulations are more likely to share code, whereas those using surveys are less likely to share code. For researchers using qualitative methods and experiments there are no significant dependencies. Out of our sample, researchers who perform simulations (57%), secondary analysis (50%) and experiments (42%) share most of the code produced, whereas those who use qualitative and survey methods shared less (38% and 31%, respectively).

Chi-square tests between funding area and whether code is created and/or made available show significant dependency for code creation, but not for code sharing. Respondents within genetic and molecular science, neuroscience and mental health, and population health research are more likely to produce code (Fig 20). Whilst the number of code packages made available seems higher for genetic and molecular science research, followed by infection and immunobiology and population health researchers, and is low for humanities and social science researchers, these differences between groups are not statistically different (Fig 22). ANOVA tests carried out to test for significant differences in the number of code packages shared between groups of respondents according to number of years in research (career stage), research discipline, funding area and location show that no significant differences exist.

FIGURE 21. PERCENTAGE OF RESPONDENTS TO PRODUCE CODE, SHARE CODE, AND PERCENTAGE OF THOSE THAT SHARE OUT OF THOSE WHO PRODUCE (N= 236)

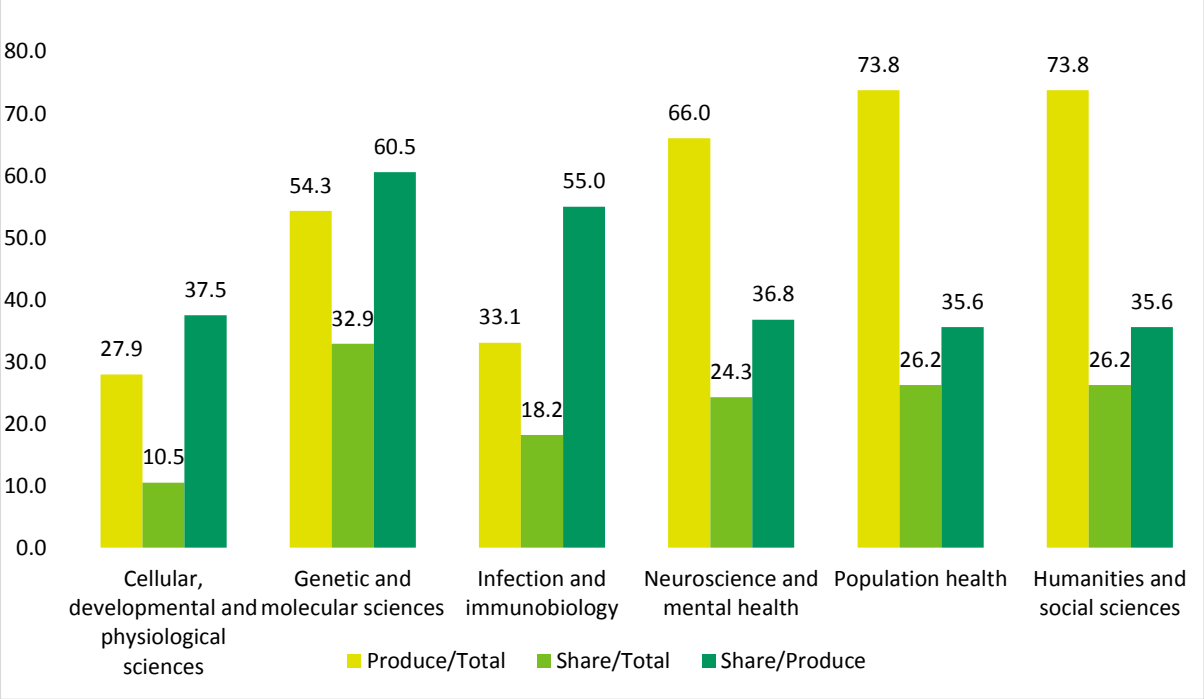
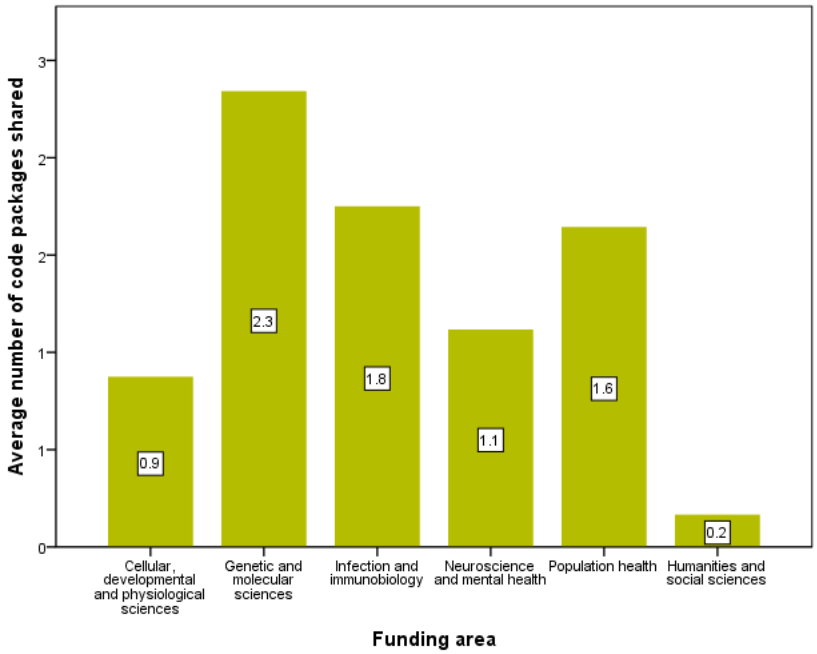


FIGURE 22. AVERAGE NUMBER OF CODE PACKAGES MADE AVAILABLE IN THE LAST FIVE YEARS BY FUNDING AREA (N=234)



Code is shared primarily via generic repositories such as GitHub, BitBucket, CRAN, Figshare, Dryad and the BioModels database (50%). It is also shared via institutional (34%), disciplinary (27%), and private (13%) repositories or via journal systems (10%) (Table 8).

For comparison, Wellcome Trust-funded respondents who identified themselves as working in Humanities and Social Sciences shared fewer code packages than ESRC-funded respondents (0.8 vs 2.5 respectively). However, this is not a statistically significant difference.

TABLE 8. REPOSITORIES IN WHICH RESPONDENTS MAKE CODE AVAILABLE, MENTIONED AT LEAST TWICE IN THE SURVEY

Repository
GitHub (N=22)
Journal supplementary material (N=12)
Comprehensive R Archive Network (CRAN) (N=3)
Sourceforge (N=3)
FigShare (N=3)
Dryad N=2)

7.2. Reasons for code sharing

Respondents indicated their code sharing activities were primarily motivated by a desire to comply with good research practice and enable other researchers to collaborate and contribute to the work (Fig 23). Other factors rated as being extremely or very important by respondents focused upon the role of code sharing in enabling research validation, replication, as well as enhancing research visibility. Funder requirements for code sharing was considered to be the least important or only slightly important factor in the choices provided, followed by journal sharing expectations. Testing for dependence between reasons to make code available with research discipline, career stage or location (ANOVA test of mean Likert scale scores) found very few significant dependencies, mainly due to the small number of respondents per subcategory (N=95 to 101).

Good research practice is statistically more important for genetics and infection and immunobiology researchers; improved visibility of research is statistically more important for genetics and neuroscience researchers.

Considering current reasons for sharing code, most respondents identified personal benefits that they had gained by making their code available, however this was not the universal response. Many respondents indicated that making their code available had led to new collaborations (38%), publications had received higher citation rates as a result of accompanying code being made available (34%), and that they had been able to produce a greater number of publications (23%). However, many respondents did not believe they had gained any benefit from making code available(34%) (Fig 24).

Conversely, only 12% of respondents indicated that they had a bad experience when sharing code, noting the cost and time effort of maintaining code and providing user support, code re-users not acknowledging or citing the source, or violating the licence conditions.

For comparison, the majority of ESRC-funded respondents did not recognise any personal benefits from code sharing activities. However, of those that did, higher citation rates and new collaborations were the most common factors identified.

FIGURE 23. REASONS FOR MAKING CODE AVAILABLE IN A REPOSITORY OR OTHER ONLINE FORM

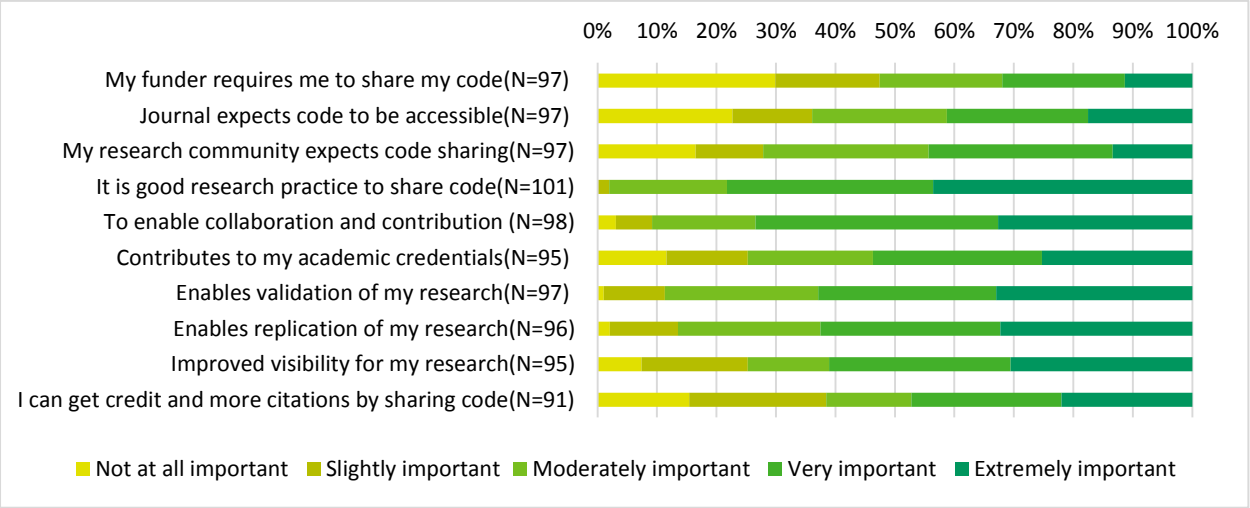
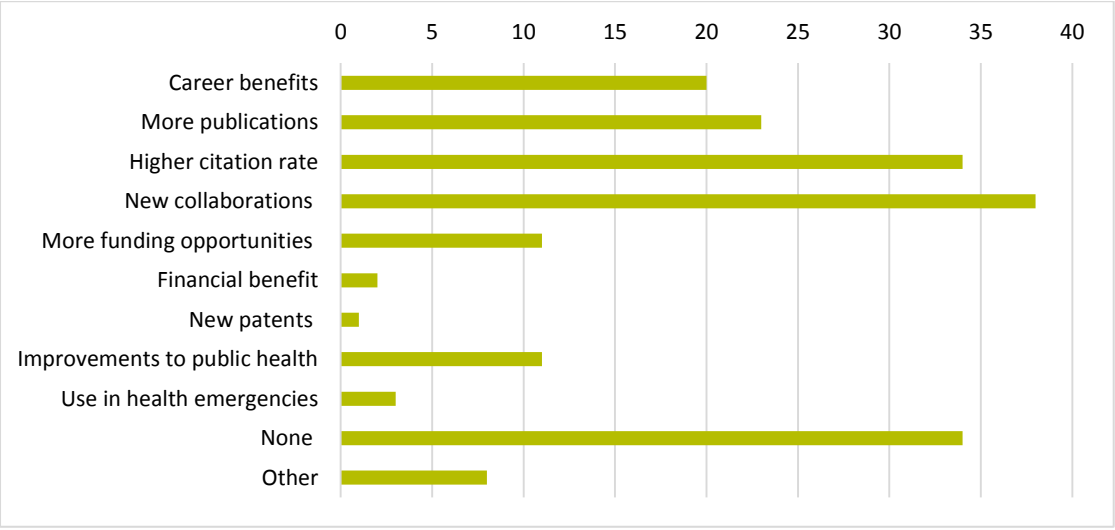


FIGURE 24. PERCENTAGE OF RESPONDENTS THAT HAVE GAINED PERSONAL BENEFITS BY CODE SHARING (N=100)



7.3. Barriers to code sharing

The primary barriers to code sharing were associated with the time and effort necessary to prepare code for access, insufficient funds available to prepare code, and concern that they lacked the necessary skill to prepare code for access and use by others (Fig 25). The majority of researchers did not consider code sharing to affect their publication opportunities, believed they possessed a suitable repository in which to host code, and did not consider their own or others intellectual property or potential patent applications to be a factor that would limit code sharing. When testing for dependency between barriers to code sharing and location (crosstabulation and chi-square test), three barriers are significantly more important for LMIC researchers (although some caution is needed due to low response numbers): desire to patent, protecting intellectual property and fear for misuse/misinterpretation.

FIGURE 25 RESPONDENT EVALUATION OF CODE SHARING BARRIERS

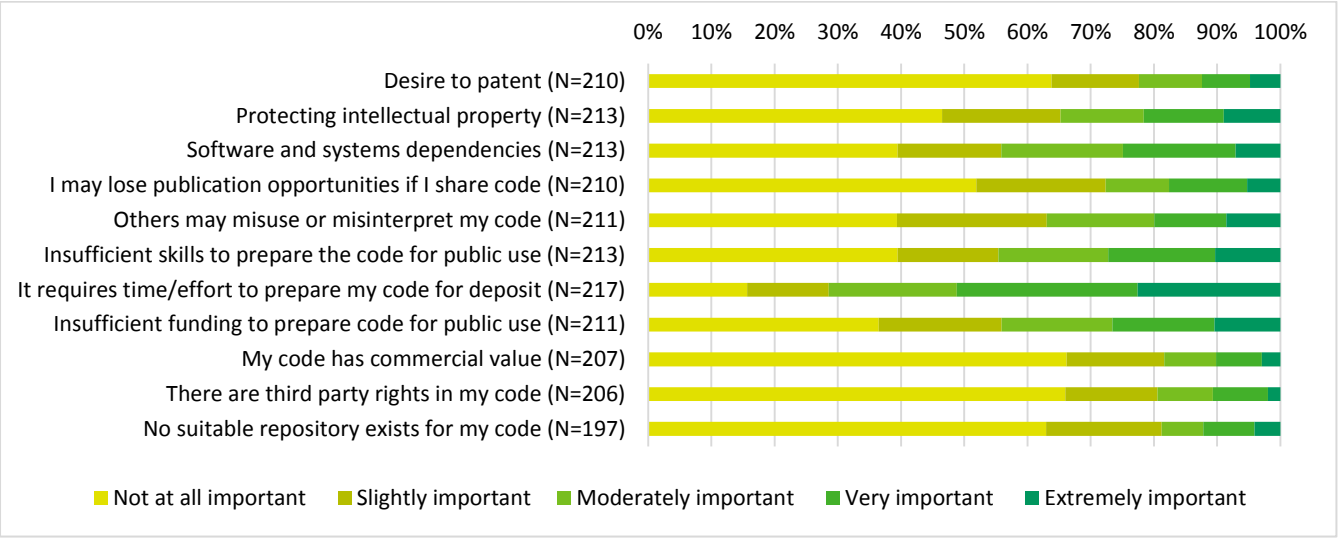
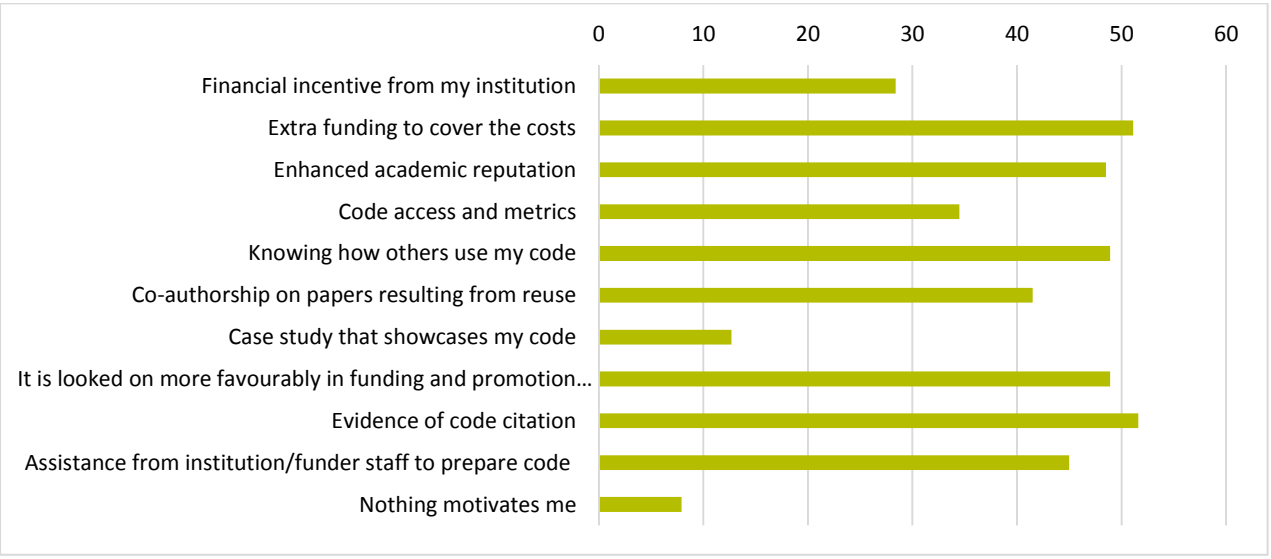


FIGURE 26 FACTORS THAT WOULD MOTIVATE THE RESPONDENT TO MAKE MORE CODE AVAILABLE, AS PERCENTAGE OF RESPONDENTS (N=229)



7.4. Motivations for code sharing

Although respondents did not recognise significant barriers to code sharing, they identified several factors that would encourage them to share more code in the future (Fig 26). Most notably, they wished for it to be looked upon more favourably in funding and promotion decisions, requested evidence that their code is cited by other researchers, additional funding to cover associated costs, as well as assistance from institutional or funder staff in preparing code for sharing. The latter was emphasised in free text responses:

> *"It takes a lot of effort to prepare code to a level that you can distribute and there is a long-term commitment to support the code once it is distributed. For example, you have to maintain it as new versions of operating systems are developed or compilers/interpreters are made. "*

Testing for dependencies (crosstabulation and chi-square tests) shows that motivations for code sharing are closely linked with a respondent's career stage and numbers of years that he/she have being doing research. Motivations that are statistically more important for junior researchers are enhanced academic reputation, knowing how others use code, co-authorship, code sharing being looked upon favourably in funding/promotion decisions, and assistance from institution/funder staff to prepare code. Demand for extra funding to cover sharing costs increases in importance as researchers spend more time in research. By comparison, those who had spent 0-19 years in research requested evidence that it would enhance their academic reputation, but this decreased among those who had spent two decades in research. There are no significant dependencies with research discipline or funding area. For researchers currently not sharing code, there is a significant dependency for assistance from institution/funder staff to prepare code as motivator for them to share code.

The notion of rewards was expanded upon by one respondent to the survey:

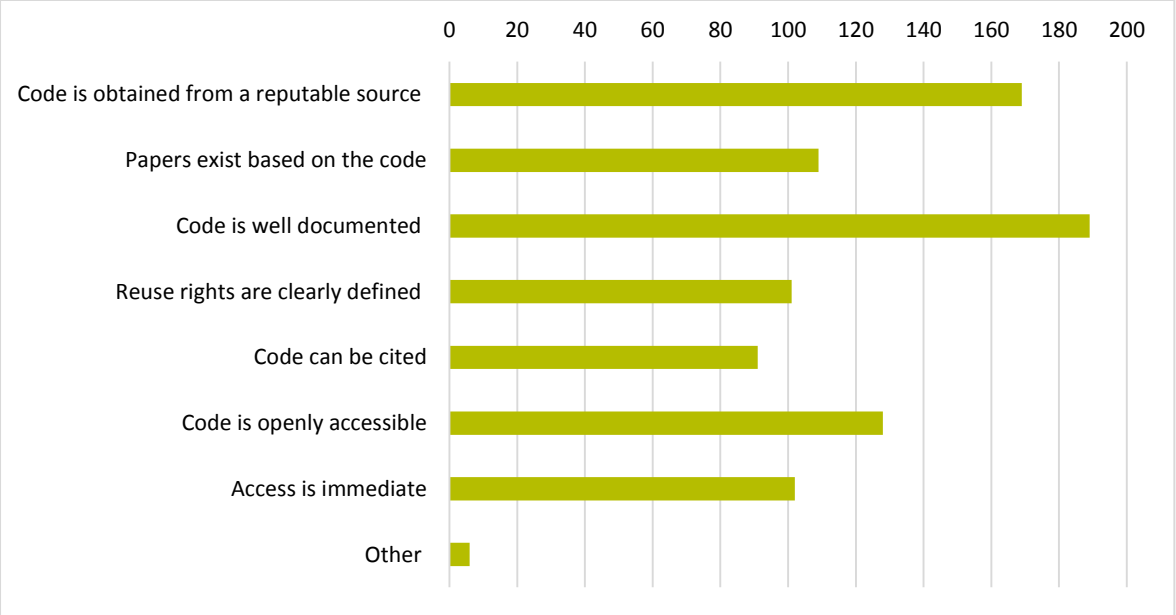
> *"Give some indication that sharing code is valued when funding decisions are made. Editing code from the state where it works on my computer to where it can be used by everybody takes a huge amount of time. In addition to making the code better / more robust, making it public also requires a significant amount of documentation. There is little credit given for this effort, especially when the code is supporting a specific paper (rather than code for a tool that will be widely used by the community)"*

7.5. Use of existing code

In total, 37% (of 583 respondents) have used existing code obtained from elsewhere. Primarily, this is obtained from peers and colleagues (63%) or from disciplinary / community repositories (44%). Wellcome Trust-funded researchers are more likely to reuse code than ESRC-funded respondents (28%), as substantiated by crosstab and chi-square test, but when only comparing humanities and social sciences researchers, then Wellcome Trust-funded researchers reuse significantly less code (8%).

The reuse of code is primarily motivated by the existence and extent of documentation describing the code (well documented), followed by an evaluation of the source from which it was obtained (reputable source), and that it is openly accessible (Fig 27).

FIGURE 27. FACTORS CONSIDERED IMPORTANT WHEN USING EXISTING CODE (N=217)



7.6. Actions Wellcome can take

When asked what the Wellcome Trust can do to help researchers share more code, 149 respondents provided free-text responses. These were coded and could be grouped into six main areas on which the Wellcome Trust could focus. Some suggestions were very generic (e.g. more funding) and were coded to the main topic, whilst other responses provided detailed suggestions. The numbers indicate the number of respondents whose survey responses corresponded to each suggestion. These same areas were explored more during the focus group discussions, with suggestions made during these sessions included in the discussion below.

Guidance, training and support (N=42)

Many researchers participating in the survey and focus groups recognise the role of code in their work, but do not consider themselves experts in its creation. They expressed a need for greater clarity on Wellcome's expectation for code development and sharing (N=2). Training on preparing code for sharing, for example via software carpentry (N=6), guidance on code development standards to be applied (N=3), on documentation practices, on licence models (e.g. GPL) and on the use of specific tools (e.g. recording processing steps in ImageJ) were raised as suggestions.
Wellcome could also provide additional support for projects trying to prepare code for publication (N=17), or staff who can prepare and publish code on the researchers' behalf (N=7).

> *"It's a lot of effort to get your code into a shape where you feel like you would want to put it in front of somebody else or make it useful for somebody else. And that kind of causes a step towards kind of basic literacy, which people of my age have kind of missed out on at school, although people of my children's age now get it as core curriculum. So we are kind of in a bit of a hole and I would say our postdocs and PhD students are in that same hole as well. They are basically they are illiterate from a data and maybe even programming point of view, even if they are using their smart phone all the time.* (Focus Group 4)"

Funding (N=31)

A small number of respondents indicated that additional funding should be provided to support code sharing activities. Specifically, there was a recognised need for funds to support the preparation of code for sharing during the project life (N=16), as well as ongoing maintenance over time (N=3).

These issues were also raised by several people in the focus groups, with the specific need for capacity building in Low and Middle Income Countries, where research projects struggle to recruit and retain staff with programming expertise, due to their desire to work in more financially lucrative sectors.

Infrastructure (N=30)

Respondents and focus group attendees were aware of, and used, code sharing platforms such as GitHub to develop and share their code. Participants questioned whether GitHub was a suitable long-term repository solution for academic research, noting that many web-based services commit to providing long-term storage and have withdrawn it at a later date. Code sharing infrastructure developments proposed were:

- a suitable repository for code developed by researchers, such as a Wellcome GitLab (N=22);
- encourage use of a single sharing platform (N=3);
- invest in creation of deposits tools that allow code (and other resources) to be uploaded to relevant repositories (N=2).

> *" I think that to me it makes common sense that if you've put a lot of effort into it, writing some code if you can put it onto a web platform, you know, it's there and you don't have to worry too much, if people want it they can just get hold of it."* (Focus Group 1)

Rewards (N=22)

Suggestions were made to recognise code sharing activities in funding decisions (N=6), to ensure that credit is given to researchers who share their code (N=6), encourage authors to cite code in research outputs (N=4) and provide resources for use by researchers wishing to protect rights associated with code (N=3).

> *"I think on a social or economic level ensuring that due recognition and credit is given to those people who do a lot of work in that arena and maybe the outputs that they publish may look more like lots of GIT commits rather than a big paper."* (Focus Group 4)

Promotion (N=16)

Wellcome could promote code sharing by producing resources that may be used to monitor code reuse metrics (N=4), providing networking opportunities for code developers and code re-users (N=3), showcasing examples of code sharing best practice, e.g. via case studies (N=3), encouraging greater recognition of code outputs within the academic promotion structure (N=2) and promote the role of code manager or scientific programmer in research (N=1) "*with the core responsibility of translating research prototypes into re-usable and friendly codes by other researchers"* and investigate the career path of developers in research e.g. in conjunction with the Software Sustainability Institute.

> *"...it's how do you get someone like I was at the time, a post doc, get a bunch of time writing code and make sure that that's not the end of the line for their career because it wasn't a paper. It wasn't a paper so what do they put on their CV? How do I convince someone that it was worth something? (*Focus Group 4)"

Policy (N=6)

A small number of respondents indicate that the Wellcome Trust should introduce a mandate for code sharing associated with research, similar to that currently applied for data (N=6).

8. Open research in general

During focus group discussions, participants made various suggestions for the Wellcome Trust to promote open research practices in general, such as showcasing “champions of open research” via an award or prize that researchers can add to their CV, and via open research workshops for early-career researchers

Participants want to see explicit valuation of open research practices, data sharing and code sharing in grant and job applications. During focus groups discussions, senior researchers who have experience of serving on Wellcome Trust funding panels indicated that whilst Wellcome does indicate that panels should not focus on impact factors and should consider a wide range of quality indicators, in reality panels may not always follow this ‘ethos’ and still focus just on classic impact factors and take decisions based on the science, not on openness. Wellcome may want to firm up on panel practices. Panels could also reassure postdoctoral researchers that they look beyond impact factors to the real contributions researchers make to science, e.g. by requesting reviewers to read and evaluate best papers as a measure of research quality, rather than simply check impact factors. Panels could also request statements of value multiplication from making data and code openly available, e.g. if code is being used multiple times by other researchers across different applications. It may not always be possible to measure this via citation. Wellcome application forms are considered good, asking for key publications (not just impact factors), but this could be expanded so applicants can showcase a wider range of outputs, such as websites, patient leaflets, code, etc.

Wellcome should look more at the wider range of ‘impact’ of research, and move away from evaluating publications and traditional outputs, instead measuring the wider impact of research on society, the applicability of research. Currently researchers feel that the ‘currency’ of success in their careers and in winning grants are publications and the novelty of research. It should instead be the quality of results and how meaningful those are. For example, contributing data to international health indicators (WHO) and collaborations with international entities should be important measures of research success. Different modes can be developed to assess quality of research and data, e.g. through audit of research practices or standards for quality research and data publishing

Participants feel that Wellcome could influence the Research Excellence Framework (REF) to use metrics other than journal impact factors to assess research excellence, to ensure that open access publishing, data sharing and code sharing are being valued as measures of high quality research. It is worth noting here that outputs such as high quality datasets can already be submitted to REF, but in 2014, only 16 datasets were submitted out of over 1900 submissions.

Can Wellcome Trust lobby university academic administrators to look beyond impact factors for promotion, i.e. look at wider aspects of publishing and making data available?

9. Discussion

Overall, respondents are practicing open research in many ways. An overview of researchers' open research practices, views, experiences and motivations are listed here, for the main areas of publishing papers, sharing data and sharing code.

Particular aspects of open research are very much determined by research discipline, career stage and the location where a researcher is based or carries out research. Statistically significant dependencies and differences are summarised here.

During focus group discussions it became clear that open research practices have been increasing lately. Several participants mentioned they had recently started doing open peer review, publishing open access books, publishing preprints, etc. Overall they had positive experiences, such as the increase of citation rates and the speeding up of science when publishing preprints. At the same time, the disciplinary differences were noticeable: in some research areas preprints are unheard of and established journals would refuse to accept a preprint paper already available online; in other disciplines they are well-established with journals actively sourcing preprints for submission from bioRxiv.

9.1. General open research findings

Publishing

With regards publishing, this research shows that in general:

- journal reputation, audience high quality peer review and impact factor are key factors when researchers publish; open access publishing is less important in comparison;
- in selecting papers to use, content quality, journal and author reputation and institutional subscription are the main deciding factors; papers being available as open access and supplementary data being available are less important;
- many researchers publish papers as open access, also thanks to funding provided by the Wellcome Trust, whereby >70% of all papers are published as open access and 30% of researchers publish everything as open access;
- open and transparent peer review, all articles being open access and costs covered by Wellcome are considered to be the most important features for Wellcome Open Research;
- future publishing systems should facilitate reviewer comments being visible, provide a commentary and discussion forum, enable data visualisation in papers and allow publish preprints that can afterwards be submitted to journals.

Data sharing and reuse

With regards making research data available and reusing existing data, this research shows that in general:

- half of researchers make research data available so they can be used by other researchers, as full datasets or subsets;
- research made available on average 4 datasets over the last 5 years;
- data are made available mostly as open access via institutional and community repositories;

- data are reused as background and context to research, for research validation, to develop methodologies and for new analysis; a quarter of researchers have never reused existing data;
- data for reuse are obtained from colleagues, repositories or directly from the creator;
- data for reuse need to be from a reputable source, of high quality and well-documented; open and immediate access are less important;
- main reasons to make data available are funder and journal requirements, it being considered good research practice, to facilitate collaboration, and to enable validation and replication;
- main benefits researchers have experienced from sharing data are new collaboration and higher citation rates, however, most researchers have not experienced any benefits, but neither have they had any bad experiences from sharing data;
- main perceived barriers to sharing data are fear for misuse or misinterpretation, fear to lose publication opportunities, and time and effort to prepare and deposit data;
- main motivations to make more data available in future would be funding to cover the cost of data preparation, data sharing to enhance academic reputation, knowing how others will use the data and data sharing being taken into account in future funding and career promotion decisions.

Code sharing and reuse

With regards making code available and reusing existing code, this research shows that in general:

- less than half (41%) of respondents produce code in their research; and 43% of these make their code packages available;
- researchers who produce code tend to share it;
- participants have different interpretations of research code – some view it in the context of software outputs, but do not necessarily consider processing scripts (such as stata.do files and batch files within this definition; it is therefore possible that a larger number of respondents produce code-like outputs, but did not complete the code section;
- when promoting the value of code as a valid research output, Wellcome may wish to consider incorporating process scripts into its code definition and advocacy activities;
- main motivations for code sharing are good research practice and ensuring research reproducibility, which is noticeably different from publication and data sharing activities (meeting funder and journal requirements); this position may be encouraged in the future through open science events;
- overall code sharing seems unproblematic for researchers;
- main barriers for code sharing are the time, effort and expertise required to prepare code for sharing; desire to protect intellectual property and other factors are minimal barriers;
- if Wellcome were to build capacity around code development, commissioning training events to build expertise and enhance guidance on where/how to publish code, it may encourage a greater number of researchers to make code available.

9.2. Open research findings by career stage and location

Particular characteristics of publishing papers and other works, data sharing, code sharing, data reuse and code reuse are strongly determined by a researcher's career stage and location. This table summarises which characteristics and factors are statistically significant for a particular category of researchers, and which barriers, incentives and motivations matter more for those groups, based on the survey findings. A more detailed description is provided in Annex 1.

	Early-career researchers	**Established researchers**	**LMIC researcher**	**UK-based researcher**
Factors important in publishing	papers are open access	journal reputation peers publish in journal of choice		
Have made research data available	less	more		
Use Wellcome funds for APCs	less likely	more likely		
Features important for Wellcome Open Research	all papers open access wide range of outputs data with papers author-led, open peer review commitment to openness outputs can be submitted to REF		Wellcome covers cost	
Reasons for sharing data	public health benefits respond to health emergencies ethical obligation to participants			
Why reuse data	replication develop methodologies	teaching materials	meta-analysis baseline data	replication
Benefits of data sharing		higher citation rates	career benefits more publications more funding public health improvements	
Barriers to data sharing	less publication opportunities no skills to prepare data	no money to prepare data	fear misuse / misinterpretation no money to prepare data no participant permission to share data 3rd party rights country-specific regulations	

	Early-career researchers	Established researchers	LMIC researcher	UK-based researcher
Motivations for data sharing	enhanced academic reputation know how others reuse data co-authorship on reuse papers higher citation rates publication of data papers	extra funding to cover costs	know how others reuse data case studies showcase data publication of data papers access controls to data	favorable for funding / promotion decisions
Barriers to code sharing			desire to patent IP protection fear misuse / misinterpretation	
Motivations for more code sharing	enhanced academic reputation know how others reuse code co-authorship on reuse papers higher citation rates publication of data papers favorable for funding / promotion decisions	extra funding to cover costs		

9.4. Open research findings by research discipline

Particular characteristics of publishing, data sharing, code sharing, data reuse and code reuse are dependent on research discipline and Wellcome funding area. Particular characteristics of publishing papers and other works, data sharing, code sharing, data reuse and code reuse are strongly determined by research discipline. This table summarises which characteristics and factors are statistically significant for a particular category of researchers, and which barriers, incentives and motivations matter more for those groups, based on the survey findings. A more detailed description is provided in Annex 1.

	Biomedical researcher	Clinical researcher	Population health researcher	Humanities researcher	Social science researcher
Research methods mostly used	experiments simulations	observation experiments	qualitative methods surveys secondary analysis observations simulations	qualitative methods secondary analysis other methods	qualitative methods surveys observations

	Biomedical researcher	Clinical researcher	Population health researcher	Humanities researcher	Social science researcher
Data types / characteristics mostly generated	quantitative omics imaging	quantitative omics imaging commercially sensitive longitudinal	quantitative longitudinal disclosive	qualitative no data	qualitative disclosive longitudinal
Use Wellcome funds for APCs	more likely	more likely			less likely
Features important for Wellcome Open Research	wide range of outputs data with papers	wide range of outputs data with papers Wellcome covers cost	Wellcome covers cost	Wellcome covers cost	
Reasons to share data	validation replication visibility of research	public health emergencies	journal expectations respond to health emergencies		
Why reuse data	validation replication background/context	validation replication meta-analysis background/context	meta-analysis background/context baseline data new analysis	less reuse	less reuse
Benefits of data sharing	higher citation rates career benefits	more publications new collaborations public health improvements	more funding opportunities more publications new collaborations public health improvements	higher citation rates	none
Barriers to data sharing		fear misuse / misinterpretation	time/effort to prepare for deposit no participant permission to share data confidential/sensitive data 3rd party rights		fear misuse / misinterpretation no money to prepare data no participant permission to share data confidential/sensitive data 3rd party rights
Motivations for data sharing		co-authorship on reuse papers access controls to data assistance to prepare data	co-authorship on reuse papers access controls to data assistance to prepare data	access controls to data know how people use data	

Funding area	Cellular, developmental and physiological sciences	Genetic and molecular sciences	Infection and immunobiology	Neuroscience and mental health	Population health	Humanities & social sciences
Research methods mostly used	experiments	experiments simulations	experiments	experiments simulations	qualitative methods surveys secondary analysis observations	qualitative methods surveys observations secondary analysis
Data types / characteristics mostly generated	omics imaging	quantitative omics	omics imaging	quantitative imaging	quantitative qualitative longitudinal disclosive	qualitative disclosive
Use Wellcome funds for APCs		more likely	more likely			less likely
Features important for Wellcome Open Research	wide range of outputs	data with papers	data with papers		Wellcome covers cost	Wellcome covers cost
Have made research data available	less	more	more	less		less
Reasons to share data		enhanced academic credentials				
Why reuse data	Background /context	validation replication new analysis	validation replication meta-analysis background/context baseline data	replication	meta-analysis background/context baseline data new analysis	no reuse
Barriers to data sharing				time/effort to prepare for deposit	time/effort to prepare for deposit no participant permission to share data confidential/sensitive data 3rd party rights	fear for misuse / misinterpretation of data

Funding area	Cellular, developmental and physiological sciences	Genetic and molecular sciences	Infection and immunobiology	Neuroscience and mental health	Population health	Humanities & social sciences
Motivations for data sharing	extra funding to cover costs know how others reuse data co-authorship on reuse papers sharing leads to data paper sharing considered in funding/promotion decisions	extra funding to cover costs sharing considered in funding/promotion decisions	know how others reuse data sharing leads to data paper case studies showcase data	extra funding to cover costs co-authorship on reuse papers sharing leads to data paper sharing considered in funding/promotion decisions	extra funding to cover costs know how others reuse data co-authorship on reuse papers case studies showcase data	know how others reuse data case studies showcase data
Produce code		more		more	more	less
Reasons to share code		good research practice enhanced visibility of research	good research practice	enhanced visibility of research	more likely	

9.5. Comparison with ESRC-funded researchers

Thanks to the parallel survey with ESRC-funded researchers, open research practices, barriers and motivations could be compared with a group of social science researchers who have for over 15 years had a mandatary data sharing policy and data infrastructure to support it; and more recently different open access publishing requirements (requirement to only make publications open access within 12 months, instead of 6) and a slightly different funding model (APC only being covered via institutional block grants). The most useful comparison is comparing ESRC-funded researchers with humanities and social sciences (HSS) researchers funded by the Wellcome Trust (rather than with the full group).

This comparison shows that whilst in general Wellcome Trust-funded researchers publish more open access papers, this is not the case for HSS researchers. And whilst in general Wellcome Trust-funded researchers use more Wellcome funding to cover APCs for open access publishing, this is not the case for HSS researchers.

With regards data sharing, Wellcome Trust-funded HSS researchers do not make significantly less data available than ESRC-funded researchers and neither are there differences in the reasons for sharing data or the benefits researchers experience from sharing data. Lack of skills to prepare data for sharing is, however, a more important barrier for Wellcome Trust-funded HSS researchers; and the lack of suitable data repositories and the fear to lose publication opportunities are more important barriers for Wellcome Trust-funded researchers in general. Wellcome Trust-funded HSS researchers see no significantly different motivations to increase data sharing.

Wellcome Trust-funded HSS researchers are more likely to never have reused existing research data, compared to ESRC-funded researchers, and are less likely to reuse data for new analysis and replication.

With regards code, Wellcome Trust-funded HSS researchers reuse less code than ESRC-funded researchers, but show no other significant differences in code sharing.

Literature

Baker, M (2016) Is there a reproducibility crisis? NATURE 533 (452)

Barbui (2016) Sharing all types of clinical data and harmonizing journal standards. BMC Medicine (2016) 14: 63. doi:10.1186/s12916-016-0612-8

Carr, D (2015) Sharing Research Data to Improve Public Health. Wellcome Trust. https://blog.wellcome.ac.uk/2015/04/08/sharing-research-data-to-improve-public-health/

Expert Advisory Group on Data Access (2014) Establishing incentives and changing cultures to support data access. https://wellcome.ac.uk/sites/default/files/establishing-incentives-and-changing-cultures-to-support-data-access-eagda-may14.pdf

Federer LM, Lu Y-L, Joubert DJ, Welsh J, Brandys B (2015) Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. PLoS ONE 10(6): e0129506. doi:10.1371/journal.pone.0129506

Pisani E and AbouZahr C. (2010) Sharing health data: good intentions are not enough. Bulletin of the World Health Organization 88(6):462–466. doi: 10.2471/BLT.09.074393

Piwowar, H.A. (2011) Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. PLoS ONE 6. (http://www.plosone.org/article/info:doi/10.1371/journal.pone.0018657)

Rathi, V. et al (2012) Sharing of clinical trial data among trialists: a cross sectional survey. *BMJ* 2012; 345 doi: http://dx.doi.org/10.1136/bmj.e7570

Ross E. (2014) Perspectives on Data Sharing in Disease Surveillance. London: The Royal Institute of International Affairs. https://www.chathamhouse.org/sites/files/chathamhouse/home/chatham/public_html/sites/default/files/20140430DataSharingDiseaseSurveillanceRoss.pdf

Sane, J and Edelstein, M. (2015) Overcoming Barriers to Data Sharing in Public Health A Global Perspective. Chatham House. https://www.chathamhouse.org/sites/files/chathamhouse/field/field_document/20150417OvercomingBarriersDataSharingPublicHealthSaneEdelstein.pdf

Savage, C.J., and Vickers, A.J. (2009) Empirical study of data sharing by authors publishing in PLoS journals. PloSOne, 4(9): e7078. Doi:10.1371/journal.pone.0007078

Sayogo, D.S. and Pardo, T.A. (2013) Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. Government Information Quarterly, 30(1): 19-31. doi:10.1016/j.giq.2012.06.011

Tenopir, C., Allard, S., Douglass, K., Aydinoglu ,A.U,, Wu, L., Read, E., Manoff, M., and Frame, M. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6. doi:10.1371/journal.pone.0021101

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. (2015) Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLoS ONE 10(8): e0134826. doi:10.1371/journal.pone.0134826

Van den Eynden, V., Knight, G. and Vlad, A. (2016). Open Research: practices, experiences, barriers and opportunities. [Data Collection]. Colchester, Essex: UK Data Archive. 10.5255/UKDA-SN-852494

Van den Eynden, V. and Bishop, L. (2014). Sowing the seed: Incentives and Motivations for Sharing Research Data, a researcher's perspective. Knowledge Exchange. http://www.knowledge-exchange.info/event/sowing-the-seed

van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014;14: 1144. doi:10.1186/1471-2458-14-1144

Wallis, JC, Rolando, E, Borgman, CL (2013) If we share data, will anyone use them? Data sharing and reuse on the long tail of science and technology. PLoS ONE 8(7): e67332 http://dx.doi.org/10.1371/journal.pone.0067332

Walport M. and Brest P. (2011) Sharing research data to improve public health. Lancet 377(9765): 537–539. doi:10.1016/S0140-6736(10)62234-9

Wellcome Public Health Research Data Forum. https://wellcome.ac.uk/what-we-do/our-work/public-health-research-data-forum

Youngseek K. and Stanton, J.M. (2012) Institutional and Individual Influences on Scientists' Data Sharing Practices. Journal of Computational Science Education 3(1): 47-56.

Youngseek, K and Adler, M (2015) Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. International Journal of Information Management 35(4): 408–418. doi:10.1016/j.ijinfomgt.2015.04.007

Acknowledgements and contributions

We wish to thank the Wellcome Trust for funding this exciting and timely research project, which we have thoroughly enjoyed, and in particular David Carr and Robert Kiley for leading this initiative on open research, and providing great insight and advice on the relevant topics. The project would have been impossible without the hundreds of researchers who contributed rich and detailed information via the surveys and during focus group discussions.

Veerle Van den Eynden (UK Data Service) and Gareth Knight (London School of Hygiene and Tropical Medicine) lead the project, developed the methodology, described and interpreted the results and wrote the report, with Veerle carrying out and interpreting statistical analyses of survey results and qualitative coding of free-text responses. Anca Vlad (UK Data Service) took responsibility for designing and administering the survey and focus groups, cleaning and analysing survey data and preparing graphs. Barry Radler (University of Wisconsin), Carol Tenopir (University of Tennessee), David Leon (LSHTM), Frank Manista (Jisc), Jimmy Whitworth (LSHTM) and Louise Corti (UK Data Service)provided expert advice towards the planning and fine-tuning of the research methodology, survey questions and focus group discussion topics, commented on the survey results, and provided expertise for the report and recommendations.

Annex 1

Open research findings by career stage

Early career stage researchers

- Consider papers being open access as very important in publishing
- Find it important for Wellcome Open Research to publish all papers as open access, to publish data alongside papers, to facilitate the publishing of a wide range of outputs and provide author-led open / transparent peer review, whereby publishing on this platform demonstrates commitment to openness and transparency, and publishing outputs can be reviewed in research assessment exercises
- Are less likely to use Wellcome Trust funding to cover APCs
- Are less likely to have already made research data available
- Are more likely to make research data available for public health benefits, to respond to health emergencies and as ethical obligation towards research participants
- Are more likely to reuse data for replication and to develop their methodologies
- Perceive main barriers to data sharing to be the possible reduction of publication opportunities and their absence of skills to prepare data for deposit
- Are motivated to make more data available in future by potential enhancement of their academic reputation, by knowing how the data may be used by other researchers, and if data publishing should facilitate co-authorship, higher citation and the publication of data papers
- Are motivated to make more code available in future if this enhances their academic reputation, by knowing how the code may be used by others, if code sharing could facilitate co-authorship on resulting papers and is looked at more favourably in funding and promotion decisions

Established researchers

- Consider journal reputation and their peers publishing in their journals of choice as very important in publishing
- Are more likely to use Wellcome Trust funding to cover APCs
- Have made more data available
- Are more likely to have experienced higher citation rates from data sharing
- Are more likely to reuse data as teaching materials
- Perceive main barrier to data sharing to be the lack of funding to prepare data
- Are motivated to make more data and code available in future if extra funding was provided to cover the cost

Open research findings by location

LMIC researchers

(including UK-based researchers doing research in LMIC)

- Are more likely to reuse existing data for meta-analysis and as baseline data
- Are more likely to have experienced career benefits, more publications, more funding and improvements to public health from data sharing
- Perceive main barriers to data sharing to be the fear for misuse / misinterpretation of data, the lack of funding to prepare data for deposit, the lack of permission from research participants, the confidentiality/sensitivity of data, third party rights in the data and country-specific regulations

- Are motivated to make more data available in future by knowing how the data may be used by other researchers, case studies showcasing their data, data deposit leading to publication of a data paper and the ability to restrict access to specific purposes / individuals
- Perceive main barriers to code sharing to be the desire to patent, protection of intellectual property and fear for misuse or misinterpretation

UK-based researchers

- Are more likely to reuse data for replication
- Are motivated to make more data available in future if this was looked upon more favourably in funding and career promotion decisions

Open research findings by research discipline

Biomedical researchers

- Are more likely to use experiments and simulations in their research
- Are more likely to generate quantitative, omics and imaging data
- Are more likely to have used Wellcome funding to cover APCs
- Find it important for Wellcome Open Research to facilitate publishing a wide range of outputs and for data to be available alongside papers
- Are more likely to make research data available for validation, replication and to make their research more visible
- Are more likely to reuse data for research validation, replication, background and context
- Are more likely to have experienced career benefits and higher citation rates by sharing their data
- Perceive few barriers to data sharing
- Need no noticeable motivations to make more data available in future

Clinical researchers

- Are more likely to use observations and experiments in their research
- Are more likely to generate quantitative, omics, imaging, commercially sensitive and longitudinal data
- Are more likely to have used Wellcome funding to cover APCs
- Find it important for Wellcome Open Research to facilitate publishing a wide range of outputs, for data to be available alongside papers, and for Wellcome to cover open access costs
- Are more likely to make research data available to respond to public health emergencies
- Are more likely to reuse data for research validation, replication, meta-analysis, background and context
- Are more likely to have experienced more publications, new collaborations and improvements to public health from data sharing
- Perceive main barriers to data sharing to be fear for misuse / misinterpretation of data
- Are motivated to make more data available in future by co-authorship on papers resulting from data reuse, the ability to control access for specific purposes / individuals and assistance from institutional or funder staff to prepare data for deposit

Population and public health researchers

- Are more likely to use qualitative methods, surveys, secondary analysis, observations and simulations in their research
- Are more likely to generate quantitative, longitudinal and disclosive data
- Find it important for Wellcome to cover open access costs for Wellcome Open Research

- Are more likely to make research data available due to journal expectations and to respond to health emergencies
- Are more likely to reuse data for meta-analysis, background and context, as baseline data and for new analysis
- Are more likely to have experienced more funding opportunities, more publications, new collaborations and improvements to public health from data sharing
- Perceive main barriers to data sharing to be the time/effort to prepare data for deposit, the lack of permission from research participants, confidential / sensitive information in the data and third party rights in the data
- Are motivated to make more data available in future by co-authorship on papers resulting from data reuse, the ability to control access for specific purposes / individuals and assistance from institutional or funder staff to prepare data for deposit

Humanities researchers

- Are more likely to use qualitative methods, secondary analysis and other methods in their research
- Are more likely to generate qualitative data or no data
- Find it important for Wellcome to cover open access costs for Wellcome Open Research
- Are more likely to never have reused existing data
- Are more likely to have experienced higher citation rates from data sharing
- Are motivated to make more data available in future via access controls and knowing how other people use their data

Social science researchers

- Are more likely to use qualitative methods, surveys and observations in their research
- Are more likely to generate qualitative, disclosive and longitudinal data
- Are less likely to have used Wellcome funding to cover APCs
- Are more likely to never have reused existing data
- Are more likely to not have experienced any benefits from making research data available
- Perceive main barriers to data sharing to be the lack of money to prepare data for deposit, the lack of permission from research participants, confidential / sensitive information in the data and third party rights in the data and fear of misuse / misinterpretation
- Have no significant motivations to make more data available in future

Open research findings by funding area

Cellular, developmental and physiological sciences

- Are more likely to use experiments in their research
- Are more likely to generate omics and imaging data
- Find it important for Wellcome Open Research to facilitate publishing a wide range of outputs
- Make less research data available
- Are more likely to reuse data for background and context info
- Perceive few barriers to data sharing
- Are motivated to make more data available in future by extra funding to cover costs, knowing how other people use their data, co-authorship on papers resulting from data reuse, data deposit leading to publication of a data paper, and data sharing to be considered in funding and promotion decisions

Genetic and molecular sciences

- Are more likely to use experiments and simulations in their research
- Are more likely to generate quantitative and omics data
- Are more likely to have used Wellcome funding to cover APCs
- Find it important for Wellcome Open Research for data to be published alongside papers
- Make more research data available, mainly as open access
- Contribution to academic credentials is an important reason to make data available
- Are more likely to reuse data for research validation, replication and new analysis
- Perceive few barriers to data sharing
- Are motivated to make more data available in future by extra funding to cover costs and data sharing to be considered in funding and promotion decisions
- Are more likely to produce code
- Share on average much code
- Are likely to share code as good research practice and to improve visibility of their research

Infection and immunobiology

- Are more likely to use experiments in their research
- Are more likely to generate omics and imaging data
- Are more likely to have used Wellcome funding to cover APCs
- Find it important for Wellcome Open Research for data to be published alongside papers
- Make more research data available
- Are more likely to reuse data for research validation, replication, meta-analysis, background and context info, and as baseline data
- Perceive few barriers to data sharing
- Are motivated to make more data available in future knowing how other people use their data, case studies to showcase their data and data deposit leading to publication of a data paper
- Share on average much code
- Are likely to share code as good research practice

Neuroscience and mental health

- Are more likely to use experiments and simulations in their research
- Are more likely to generate quantitative and imaging data
- Make less research data available compared to other disciplines
- Are more likely to reuse data for replication
- Perceive main barriers to data sharing to be the time and effort to prepare data for deposit
- Are motivated to make more data available in future by extra funding to cover costs, co-authorship on papers resulting from data reuse, data deposit leading to publication of a data paper, and data sharing to be considered in funding and promotion decisions
- Are more likely to produce code
- Are likely to share code to improve visibility of their research

Population health

- Are more likely to use qualitative methods, surveys, secondary analyses and observations in their research
- Are more likely to generate quantitative, qualitative, longitudinal and disclosive data
- Find it important for Wellcome to cover open access costs for Wellcome Open Research
- Are more likely to reuse data for meta-analysis, background and context, as baseline data and for new analysis

- Perceive main barriers to data sharing to be the time/effort to prepare data for deposit, the lack of permission from research participants, confidential / sensitive information in the data and third party rights in the data
- Are motivated to make more data available in future by extra funding to cover costs, knowing how other people use their data, co-authorship on papers resulting from data reuse and case studies showcasing their data
- Are more likely to produce and share code

Humanities & social sciences

- Are more likely to use qualitative methods, surveys, observations and secondary analyses in their research
- Are more likely to generate qualitative and disclosive data
- Are less likely to have used Wellcome funding to cover APCs
- Find it important for Wellcome to cover the cost of open access in Wellcome Open Research
- Make less research data available
- Are more likely to never have reused existing data
- Perceive main barriers to data sharing to be the fear for misuse / misinterpretation of data
- Are motivated to make more data available in future by knowing how other people use their data and case studies to showcase their data

Veerle Van den Eynden is a manager in the Producer Relations and Research Data Management team at the UK Data Archive, providing expertise, guidance and training on data management and sharing to researchers across the UK, leading data sharing research and development projects, and coordinating the archiving of research data from ESRC grants with the UK Data Service.

Gareth Knight is Research Data Manager at the London School of Hygiene & Tropical Medicine, who led a Wellcome Trust ISSF project which established a Research Data Management Service for the institution, with Research Data Management policy, web-based RDM guidance, training and one-to-one support for researchers and an institutional research data repository, LSHTM Data Compass.

Anca Vlad, Collections Development Officer at the UK Data Archive, manages and oversees the deposit of research data from ESRC grants into the ReShare data repository and coordinates the review of data for disclosure risk, confidentiality, documentation and completeness.

Barry Radler is researcher at the University of Wisconsin's Institute of Aging, and has over 20 years of experience conducting survey and marketing research. He is responsible for the DDI Lifecycle metadata underpinning the discovery portal for the longitudinal Midlife in the United States study (MIDUS) and interested in how adoption of metadata standards can improve the entire research data lifecycle, from conceptualization and development to analysis and sharing.

Carol Tenopir, Chancellor's Professor of Information Sciences at the University of Tennessee, is a co-lead of the Usability & Assessment Working Group of the National Science Foundation-funded DataONE project, that develops cyberinfrastructure and a culture of data sharing among earth and environmental scientists. She also leads research projects on the needs, behaviours, and practices regarding data sharing among key stakeholder groups.

David Leon is an epidemiologist at the London School of Hygiene & Tropical Medicine, who has been funded by the Wellcome Trust for over 15 years to conduct studies in Russia; research that has included the establishment of a metadata website to facilitate data sharing. He also chairs the School's oversight group on open access and data sharing.

Frank Manista is Jisc open access support coordinator and OpenAire open access focal point for the UK, working with library open access teams to assist in complying with funders' mandates, monitoring expenditure and engaging with researchers on understanding the direct and indirect costs of publishing.

Jimmy Whitworth (London School of Hygiene & Tropical Medicine) is a public health physician who worked at the Wellcome Trust for 10 years, during which time he led a multi-funder initiative to promote data sharing of public health research data from low and middle income countries.

Louise Corti is an Associate Director of the UK Data Archive and Head of the Producer Relations and Collections Development teams.

Citation: Van den Eynden, Veerle et al. (2016) Towards Open Research: Practices, experiences, barriers and opportunities. Wellcome Trust. https://dx.doi.org/10.6084/m9.figshare.4055448