Written evidence submitted by The Royal Statistical Society (DDA0037)

1. Summary

1.0.1. The Royal Statistical Society (RSS) is the learned society for statisticians and data scientists. Our members' work depends on being able to access data for research purposes, to develop innovative evidence bases informing decision-making and delivering public benefits.

1.0.2. Policy around digital data and the right to privacy is a difficult area, at least in part because attitudes to privacy vary widely with context. Trust in organisations that collect personal data is enhanced by the perception that the interests of the organisation are aligned with – or at least, not in conflict with – the interests of the data subject. It is also important that people have some meaningful control both over what data is collected, and the uses to which it may be put. Therefore, governance must include clear structures for accountability either to, or on behalf of, people whose data has been used.

1.0.3. The RSS's view, as expressed in our Data Manifesto, is that greater data sharing between government departments for statistics and research purposes will strengthen our public services – improving health, education, housing, transport and the lives of the public. Researchers and statisticians in the public and private sectors require improved access to government data for research purposes, particularly person-level records. Safeguards to protect people's fundamental rights – such as privacy – should be built into any data sharing at the outset.

1.0.4. In this submission we set out what this means in the context of the specific questions that the Select Committee is interested in and make the following recommendations:

Recommendation 1: *Build on the success of the Covid-19 Dashboard to use more data for the public good by keeping the public informed about the functioning of the state – including, for example, more areas of the health and social care system.*

Recommendation 2: *Invest in training for researchers about the ethics of using pre-existing data for research, including developing case study materials as resources and improving understanding of what is required for a data access application.*

Recommendation 3: *UK Research and Innovation (UKRI) should set clearer expectations for research projects that it supports and monitor how research data access is offered.*

Recommendation 4: *Ensure that any understanding of scientific research that may be used to allow researchers to access data is broad enough to encompass social and economic research.*

Recommendation 5: *Support universities and other research institutions in identifying the correct existing legal grounds for research rather than creating new legal grounds for university research.*

Recommendation 6: *The government's data strategy should include a strategy for public engagement to build an understanding of how data is or should be used and to foster trust in the safeguards employed to protect the public.*

Recommendation 7: *Assign strategic oversight of the various bodies involved in governance to an independent and authoritative body to help coordinate and communicate activity.*

2. Potential benefits of and existing barriers to data sharing

2.0.1. In general, we take the big picture benefits of data sharing and improving research access to data to be well understood: data sharing for statistics and research purposes allows the compiling and use of statistics for the public good – leading to better policymaking, strengthening trustworthiness and democracy as well as strengthening public services to improve health, education, housing, transport and the lives of the public. The concept of statistics for the public good is an important one – and one that is important to the work of the Office for Statistics Regulation (OSR)[1] – but the term “public good” is very wide and work is required to understand how the statistical system can most effectively serve it. There is a role for parliamentarians to engage in helping to set out what the public good looks like for this purpose.

2.0.2. As the RSS is a membership organisation – made up of statisticians and data scientists many of whom try to make use of government data to understand a range of very specific issues – we have asked our members to give some examples of the potential benefits that they can see from their own research as well as to highlight some of the examples of barriers that they have faced.

2.1. Potential benefits

2.1.1. There are a number of examples of ways in which data can be used to improve public services:

2.1.1.1. The Covid-19 pandemic has clearly illustrated several ways that data sharing can be used to help respond to a public health emergency. For, example, the Covid-19 Dashboard has grown to draw together data from a variety of sources to report on over 200 metrics per day and has been a vital resource in guiding the UK’s response to the pandemic.

2.1.1.2. The pandemic has also shown an appetite for local data – the public have been keen to understand the picture in their very local vicinity. There is an opportunity to provide data at a much more local level than we are currently used to. There has been work to support local authorities build up their data capability in the context of Covid-19, which has found that question development is as much a barrier as other skills, and some local authorities have more capacity than others. Connected Health Cities and the Great North Care Record are good examples of local progress, but they have been constrained to greater access to only health data for only direct purposes.

2.1.1.3. Grouped data – information on groups of individuals – has the potential to be used more than it currently is. This approach can be used to estimate the prevalence of rare traits, such as was successfully used to estimate the prevalence of HIV. As well as this sort of benefit to understanding specific medical conditions, the technique can be applied much more widely.

2.1.2. As well as providing benefits to particular services, there is also a wider sense in which data sharing can promote the public good by informing public debate and encouraging trust through transparency. This is particularly beneficial in areas where there are particular concerns about privacy – such as in the health and social care system – or where there is an issue where evidence is disputed – eg, around economic issues tracking job vacancy rates, labour shortages, delivery times for certain goods, balance of trade figures.

2.2. Barriers

2.2.1. Providing access to publicly held data is a public service, and – given the potential benefits of this – it is reasonable to expect that the process should be smooth and as straightforward as possible. Some parts of government have made great improvements to the way that their services are accessed – eg, the DVLA have made ordering a new driving license much more straightforward – and while accessing data cannot be as straightforward as other government services, there is considerable room for improvement. Our members have reported a number of areas where they have had difficulty accessing datasets. A sample of the feedback we have received is below:

[1] Some of their work is detailed in a recent article *What do we mean by ‘statistics that serve the public good’?*

2.2.1.1. We have received mixed views on how the UK Data Service (UKDS) and UKRI-funded research centres (eg, Consumer Data Research Centre) are functioning to allow access to datasets. Some members report that administrators within the data provider (or their own university's lawyers) can unintentionally block or otherwise hamper speedy access to research datasets. Others report having made good and repeated use of UKDS and continuing improvements at secure access centres. Because of the variety of experiences, this is a tricky issue to resolve – but improving researchers' understanding of what is required for a data access application would likely help.

2.2.1.2. Data Protection Officers have an important role in safeguarding data – particularly in health trusts. This is an important role and in general the people in these roles take very seriously their duty to ensure that people's data is protected. However, especially in health and mental health trusts, their focus and narrative are based on a fear of breaching the information governance rules. Without shifting the narrative on data away from this and onto how we can use valuable patient data to improve lives, this is thought likely to remain the case.

2.2.1.3. Data owners do not always seem to be fully engaged with the governance structures, which in turn may reflect that the costs and the benefits of data sharing fall in different places. There is often an unwillingness to sign agreements that allow data sharing and/or research applications due to a lack of confidence/understanding in assessing the risk involved. There is a role here for UKRI to set clearer expectations for research projects that it is involved with and monitor how research data access is offered.

2.2.1.4. One member's efforts to access the Mental Health Dataset maintained by NHS Digital is illustrative of some of the barriers faced. The Data Access Request Service is not set up to supply a regular monthly or quarterly feed of data to an analyst, researcher or statistician and is more suited towards a one-off supply. Further, to harvest value from this dataset, it is necessary to run a series of algorithms or complex queries against it. NHS Digital are not set up to do this and the only way to do this is for the analyst to be able to access the raw data.

Recommendation 1: Build on the success of the Covid-19 Dashboard to use more data for the public good by keeping the public informed about the functioning of the state – including, for example, more areas of the health and social care system.

Recommendation 2: Invest in training for researchers about the ethics of using pre-existing data for research, including developing case study materials as resources and improving understanding of what is required for a data access application.

Recommendation 3: UK Research and Innovation (UKRI) should set clearer expectations for research projects that it supports and monitor how research data access is offered.

3. Comments on the government's current approach

3.0.1. The emphasis that government is currently placing on data is welcome – in addition to the policy documents listed in the Committee's call for evidence, the government's integrated defence review and research strategy also have data at their centre. The RSS considers data to be a major competitive advantage for the UK: the use of data by the ONS is world-leading and they are increasing access for researchers as well as attempting to join up various government data sets; and, the UK's strong higher education sector means that we have the capability to invest in the skills pipeline, ensuring that we have a workforce able to make use of our data resources.

3.0.2. In our view, the various strategies do not make a clear analysis of the UK's advantages in this area. This is important – if we have the ambition to lead in this area it is important that the government appreciates all of the strengths that the UK has. The government's current approach seems to focus on technology but does not seem to appreciate the strength in foundations, training and ethical practices around statistical use of data.

3.0.3. The RSS's view – as set out in our Data Manifesto – is that greater data sharing between government departments for statistics and research purposes could strengthen our public service. It is also important that researchers in the public and private sectors are given access to government data when there are clear benefits to society. It is important that privacy is appropriately safeguarded – but the social benefits of accessible data are such that the system for accessing data needs to be designed with privacy and accountability properly integrated so that rapid access to data can be given with minimal risk of inadvertent harms resulting. This is especially so, as noted in the introduction, given that people's attitude towards data sharing is so heavily context dependent.

3.0.4. We have responded to a number of government consultations on these topics: most recently responding to 'Data: A New Direction' and the 'National Data Strategy'. Here we will summarise the concerns expressed in this document that are relevant to the Committee's inquiry.

3.1. Data: a new direction

Research access to data

3.1.1. We welcome the government's efforts to clarify the legal grounds for researchers to access data but think that the government's approach to this – through legislation – is not the best way to go about it. First, we would question their plan to create a statutory definition of scientific research – where we do not think that there is a suitable definition available. Any option that we have considered seems either too permissive or too restrictive. For example, the definition proposed in the consultation – 'technological development and demonstration, fundamental research, applied research and privately funded research' – strikes us as both too permissive and too restrictive. It seems to consider any research whatsoever as scientific research, so long as it is privately funded. While at the same time it is not clear enough about its domain to set out whether study is restricted to the physical and natural world, or whether it also includes the social world. From our perspective, it is vital that social and economic research counts as scientific research because many students and researchers develop their statistical skills in this field and go on to apply them widely after they graduate or leave academia.

3.1.2. We also question the government's plan to create a new, separate lawful ground for university research. There are two challenges here that concern us. First, it is important to consider public perception. It would be potentially damaging if the new lawful grounds could be portrayed as allowing essentially any university research – if the public feel that the safeguards that are in place on personal data use and reuse are not sufficiently robust, there is a risk that people will be less willing to consent to their data being used in the first place. And this, in turn, would mean that any analysis conducted would be much less useful as the data it was based on would be highly selective. On the other hand, if – to maintain public confidence – overly robust safeguards are introduced, then there is a risk that useful research is discouraged. Of course, non-university bodies, public and private, conduct data-driven research too, and much research is collaborative. Any new lawful ground would have need to recognise this; in which circumstances the concerns above remain relevant.

3.1.3. It is not at all clear to us what type of safeguards would be sufficiently reassuring to the public while also not discouraging potentially useful research. The general problem is that the required safeguards are contextual and will change over time – so researchers need to not only rely on legal grounds but also be conscious enough of the broader context to engage with it, understand the public's views and change safeguards accordingly. It is difficult to mandate that behaviour through legislation, so the answer needs to involve both some legal safeguards as well changes to working practices – where there is perhaps a role for intermediaries like the RSS and other professional bodies. Improving guidance to universities so that they can precisely identify the correct existing legal grounds and looking to work with partners to improve working practices would seem to be a preferable approach.

Public Understanding

3.1.4. The government's approach as outlined does not place sufficient emphasis on improving public understanding of issues around the use of their data – especially around the concepts of consent and legitimate interest. One of the key challenges in this area is that public understanding of these issues seems to be lacking.[2]

3.1.5. Improving public understanding needs a more strategic intervention than changing the law or improving guidance – it requires a strategy for public engagement to build an understanding of how personal data is used. There are examples of good practice in this – earlier last year the Geospatial Commission conducted a public dialogue around location data and Understanding Patient Data is also supporting dialogue with the public around the trustworthy use of patient data. Encouraging this type of work – for different use cases and data types – as well as drawing it together and sharing learning should be a central part of the government plan

3.2. National Data Strategy

3.2.1. In our most recent work on the National Data Strategy, we partnered with the Ada Lovelace Institute, the Centre for Public Data, the Institute for Government and the Open Data Institute to organise a series of events covering the four pillars of the strategy. We jointly published a document Getting data right: perspectives on the UK National Data Strategy 2020, which details the key insights gathered through the series. In our individual response to the strategy, the RSS focused on data skills. While improving statistical and data skills would help improve commissioning and governance, we take it that this inquiry is not intending to focus on these issues, so we will not expound our views on this here.

3.2.2. The aspect of our response which is relevant is around the lack of emphasis in the strategy on the responsible and appropriate use of personal data by government, academia and the private sector for a wide variety of purposes (ranging from the desire to increase profitability, to improve service efficiency and/or to address major research issues where there is a public interest). While the Centre for Data Ethics and Innovation has an important role and the government's data ethics framework is welcome, we believe that there are ways in which the data ethics agenda could be strengthened.

3.2.3. As the Cabinet Office has now assumed responsibility for government use of data, we would like to see its ethics functions more strongly linked to the UK Statistics Authority (UKSA). We would like to see the adoption of some of UKSA's good practice – as set out in the Code of Practice for Statistics – of being open and external facing.

3.2.4. We view the self-assessment tool that UKSA have developed and that is mentioned in the National Data Strategy as exemplary and would strongly encourage an uptake in its usage. It is welcome that the ONS Secure Research Service and the UKDS Secure Lab have this self-assessment as a mandatory part of their process. Sometimes researchers can struggle to complete it because due to a lack of understanding about the ethical issues involved in, eg, data reuse. This agenda could be developed further by extending it to include:

3.2.4.1. The provision of online ethics training for the research and statistical community across Government, academia, and the commercial sector. The training that is currently available is at an introductory level and more advanced training suitable for those in leadership roles is needed.

3.2.4.2. An ethics user support service for the research and statistics community to provide ethics advice at the research design phase.

3.2.4.3. While guidance on cross cutting ethical issues in research and statistics is now being published, there needs to be greater engagement with a wider range of people – currently the views of users who are accessing data dominate these documents.

[2] We are not aware of quantitative research that has been conducted into this area – but the Open Rights Group have conducted interview-based research *Public Understanding of GDPR* which suggests that there is a high level of awareness that data processing can require consent, but a low level of understanding of what consent means.

3.2.5. This is an area where there is an opportunity for the UK to be distinctive and offer international leadership on data ethics. The ambition of ensuring "that UK values of openness, transparency and innovation, as well as the protection of security and ethical values, are adopted and observed globally" is praiseworthy: signalling our commitment to independent ethics and use of data in the public interest may have soft power benefits in the international scene. Indeed, this has been recognised in the independence of our regulation of statistics by OSR, which is more strongly codified than in other countries, and the role of professional bodies in establishing the data science accreditation standards.

Recommendation 4: Ensure that any understanding of scientific research that may be used to allow researchers to access data is broad enough to encompass social and economic research.

Recommendation 5: Support universities and other research institutions in identifying the correct existing legal grounds for research rather than creating new legal grounds for university research.

Recommendation 6: The government's data strategy should include a strategy for public engagement to build an understanding of how personal data is or should be used and to foster trust in the safeguards employed to protect privacy.

4. Statistical perspective on the ethics underpinning data sharing in a health context

4.0.1. Statistical work to analyse health data and identify measures that might improve people's wellbeing is dependent upon effective data sharing. In a health context, it is important that statisticians have access to information about individuals: while statisticians are only interested in aggregating the details of any single person in a dataset, it is important that the data being used has origins that are well understood enough to make and test assumptions about the distribution of variables observed and their relationships. What matters is the underlying relations between the variables of interest – so, for this reason, confidentiality should always be able to be respected in statistical work.

4.0.2. There is a distinction between privacy and confidentiality:

4.0.2.1. Privacy applies to a person and the types of questions that are relevant here are around how people in a study are identified and what methods should be used to collect information about people.
4.0.2.2. Confidentiality applies to data and the key questions there pertain to identifiable data, eg, how the data is maintained and who has access.

4.0.3. Statistical research often involves reusing data – and, because this involves questions about how people are identified and what consent they have given, there are important questions about how this type of statistical research is compatible with data protection. After all, statistical research cannot rely upon getting specific consent from every person in a desired dataset – it may be impossible to achieve this where information on large populations is required, with impacts on costs and also potential bias.

4.0.4. One possible ethical basis for this type of work is the concept of contextual integrity:[3] ie, the idea that any further use of data ought to respect the context in which data was originally offered and using it for purposes beyond those originally intended requires ethical justification. At the most straightforward level, this would mean that consent can be sought for research purposes generally, without needing to gain consent for every individual

[3] This is a concept developed by Helen Nissenbaum in her book *Privacy in Context, Technology, Policy and the Integrity of Social Life*.

research project, as research projects go through other processes, such as research approval boards, to protect people from harm.

4.0.5. However, in a health context people do not standardly give consent for their data to be used for research purposes. Instead, it is generally understood that, in this context, people reasonably expect their data to be used for direct delivery of services and the management and audit of those services only. Service evaluation and research – the use cases in the proposed sharing mechanism for care data and GDPR -- typically entail access to the data for other specialist users, so are not necessarily consistent with the context in which data was originally obtained.

4.0.6. There are two types of argument that are advanced for why it is ethical to use a person's data in this way: the first is based on the concept of reciprocity, the second on solidarity.

4.0.6.1. Reciprocity: the idea here is that as individuals have benefited from past research, it is justifiable for their data to be used to help other people in the future.

4.0.6.2. Solidarity: in the context of a group of people with a common concern, this means a sense of responsibility for others and for the group as a whole – the argument is that individuals have a responsibility to this group and so where there is an opportunity for their data to improve outcomes for the group as a whole it is justifiable to use it.[4]

4.0.7. We suggest that the concept of solidarity is the more effective way to understand the ethical basis for using data for research purposes. The argument does not rely on individuals being able to see that they have benefited from this type of research in the past and it captures a more general conception of improving the future of the health system for society, rather than just direct instances of treatment. The case is perhaps more abstract than the case that is based on reciprocity – and that is perhaps why it can seem easier to make a case based on reciprocity – but it provides a wider rationale.

4.0.8. There is a clear argument, based on solidarity, that NHS data is a national and even – as discoveries using NHS data have benefitted the world – global, asset that should be used for the public good both within the UK and internationally. The Covid-19 pandemic perhaps makes it easier to make the case to the public for this view of health data than it would otherwise be – as there are so many examples from the past two years of how this data has helped manage the crisis. There is perhaps an opportunity to refer to these examples to make the case that while privacy is an important consideration it is not helpful to regard it as an absolute. It has been especially beneficial that the public has been able to see information about the health system in the Covid-19 Dashboard – and seen how this has fed into government decision making – and we would recommend that this general practice (of routinely sharing data about the health system) continues even after the worst of the Covid-19 crisis has passed.

4.0.9. Visibility of how people's personal data is being used should build trust and understanding in both the health and social care system as well as in how health data is used. The project People Like You has done work to show people how commonalities in individuals' data can be used to get identify wider patterns. Trust is a two-way relationship and ensuring that people who are being asked to share their health data for altruistic reasons can see that data feeding into information about how the healthcare service is meeting their needs can only improve trust.

[4] This is detailed in Barbara Prainsack and Alena Buyx's *Solidarity in Biomedicine and Beyond*

5. The effectiveness of existing governance arrangements

5.0.1. Governance arrangements tend to be designed to control the use of data in respect of a single specific context, without too much consideration of the wider picture. This is especially clear in the case of health where the use of clinical data for things other than direct health purposes – such as statistical research – faces resistance from clinicians, governance and administrative personnel. Statisticians increasingly link together data of different types and sources to do research, which means contending with multiple processes and confusions about the beneficiaries of research.

5.0.2. Progress has been made in making the public benefit case for research and sharing information with the public about the work, but administrative processes can still be challenging to navigate. Communication about the nature of data and its utility remains challenging, and the National Data Strategy makes no attempt to tackle this issue. Many organisations – eg, Which?, Ada Lovelace Institute and ESRC – are attempting to tie this together, but it would be better for an organisation like the Centre for Data Ethics and Innovation (CDEI) to have been leading on this.

5.0.3. Governance bodies are numerous, each guarding norms in particular settings, and trying to maintain the legal arrangements that are in place – however this can result in their taking an excessively risk averse stance. In some cases, eg the METADAC project, progress can be made – but in this example genetic data was linked with social survey data by bringing more data into the more sensitive setting: there remain difficulties in cases where two separate data owners need to negotiate. There are some innovations which are promoting sharing data by negotiating on the basis of the benefits as set out in more recent legislation – eg, the UK Statistics Authority's Research Accreditation Panel was introduced to fulfil new duties under the Digital Economy Act and is proving successful in this respect. But even the local health data sharing models come up against barriers locally when professionals work just outside the system such as, eg, if a speech therapist is employed directly by a school.

5.0.4. As research relies on more comprehensive data, larger samples and linking between services, local and sector governance arrangements are insufficient. The work done by the Office of National Statistics (ONS) with Health Data Research (HDR) UK and Administrative Data Research (ADR) UK is putting in place systems to address this, but the processes are very onerous for individual researchers and PhD students in particular. In broad terms, the governance bodies are looking to their own legal responsibilities to restrict access to data rather than facilitating access that supports public benefit, especially across established boundaries. The government strategies have either stepped around the aims and practice of research, or confined it within a domain, as in health. Neither approach matches the political ambition to realise the value of data.

5.0.5. A proliferation of guidance documents has not assisted research as policy reports are large and lack standing while codes of ethics are impractical or too high level to be useful. More recently the UKSA Centre for Applied Data Ethics has started to publish guidance documents which need to be supported with ethics case studies. At present, researcher training lacks curriculum material of the kind that is standard in engineering programmes[5] and ought to be developed but activity is limited by liability concerns. There is a lot of activity in producing guidance but not obvious coordination, which looks like it is being driven by organisations being driven by a desire to avoid ceding authority.

5.0.6. The RSS has found in its own project developing professional standards for data science – a project which is included in the National Data Strategy – that there are practice gaps in respect of what responsible innovation looks like. Standards will include programming and analytical skills as well as business data strategy - a further strand on what is required for ethics is in development, at the stage of sharing current practice. This is an area

[5] The Royal Academy for Engineering has information on these materials.

where we originally hoped the CDEI would be involved but have established their focus on digital technology rather than for practitioners using data responsibly. Even before these standards, RSS members have a Code of Conduct which is compulsory in respect of Chartered Statisticians, and there is further expectation at international level through the International Statistical Institute code and the UN's Fundamental Principles of Official Statistics. We note that RSS members have led at these levels very recently, and the models of the OSR and the National Statistician's Data Ethics Advisory Committee (NSDEAC) are very well regarded.

5.0.7. The big challenge of governance is around who benefits from reuse of data, and the determination of which projects go ahead. At present, the best example of an ethics committee is NSDEAC, but that has quarterly meetings and may require extensive revisions, so is best suited to government projects. Most other organisations have poor processes to oversee the reuse of data for research purposes, which have left researchers stymied, overwhelmed or anxious. Reuse of data within organisations tends not to be restricted in a coherent way, with sharing beyond the organisation restricted and without regard to the public value. More complex arrangements, for limited sharing within some kind of accredited club for private sector and research uses, do exist and are often branded 'data trusts' although there are many models of these data institutions and it is unclear which models, other than classic data sharing contracts, are proving effective in different contexts.

5.0.8. This is a difficult problem, and it is not clear how it will be solved solely by data trusts or other new data institutions. As a starting point a proper stakeholder analysis needs to be conducted and some part of government ought to have strategic oversight. It is not clear which organisation is best suited for this. Usually, governance is best placed in contextual bodies – ie, those within a sector who properly understand it. In some sectors, this is obvious (eg, Ofgem for energy, Ofcom for telecoms) and in some (eg, health and social care) it is less obvious – but in principle various different contextual bodies might reasonably want to take different approaches to data governance. What is important, from our perspective, is that a pre-existing, independent and authoritative body has strategic oversight of all the organisations and stakeholders involved to help coordinate and communicate activity. This could be the Information Commissioner's Office – which seems natural from one perspective because it would not require creating a new organisation and it is the regulator for protecting people from harm due to data collection, sharing or use. However, careful thought would be needed as to whether it is wise to combine their existing role with broader data governance: it would be a large job and also risks reducing the emphasis on harm reduction. So, some other body may well be more appropriate.

Recommendation 7: Assign strategic oversight of the various bodies involved in governance to an independent and authoritative body to help coordinate and communicate activity.

January 2022