The opportunities and challenges to sharing data across the Osteoarthritis research community

Compiled by Jill Evans, Paul Biggs, Mark Elliott

March 2020

Institute of Digital Healthcare
WMG, University of Warwick
Coventry, UK
CV4 7AL

Contact: Mark Elliott
Email: m.t.elliott@warwick.ac.uk

Contents

Executive Summary

Background

Osteoarthritis (OA) research covers a broad range of sub-disciplines, working largely in siloed groups. The resulting spectrum of datasets is heterogenous and there is no central repository, nor any best practice framework or minimum dataset requirement. However, there is growing evidence to suggest that large datasets are of significant benefit to the OA community, and could contribute to expedited advances in understanding the disease. OA research typically suffers from small sample sizes which may affect the ability to derive meaningful insights from the data.

Recent advances in imaging and wearable technology have generated new opportunities in research, including in the field of OA, to access large datasets and potentially pool them to create so-called 'big data'. Determining if and how this might be utilised by OA researchers is a challenge which if addressed in the near future could lead to much faster advances in OA research and treatment. One avenue of interest to OA researchers is machine learning, which may transform the field by taking this big data and developing algorithms and pattern identification previously not possible due to time or resource constraints. By doing so, there is potential for earlier identification and stratification of OA to be achieved in a reliable way.

However, such advances will not be possible if data are not shared between researchers or collected purposively for repositories. To do this requires either homogeneity of data to begin with, or a feasible way of homogenising heterogenous data which is not prohibitively time consuming. Data pooling also requires anonymisation processes which satisfy data protection legislation in whichever countries and organisations in which the data will be used. Anonymisation is in itself a further challenge, as these processes can result in loss of granularity of data. It is unknown to what extent OA researchers share data or access databanks, and how they currently attempt to address these challenges (if at all).

The current project, funded by the OATech Network, aims to find out what the current practices are within OA research in the UK with regards to data sharing and accessing big data. This project involved speaking directly to OA researchers and clinicians to determine not only what the current approaches are to data collection and sharing, but also to ask what the desired future directions are for the field and whether a best practice framework might be achievable.

Methodology

Interviews were conducted with nine researchers working in osteoarthritis, one representative from a large databank and two commercial representatives. Semi-structured questions asked participants about their experiences collecting, sharing and using data within OA research, and their opinions on data sharing and associated issues. In addition, a case-study interview was held with a clinician who had set up a UK-wide Cerebral Palsy database, to gain knowledge of data-sharing from a different clinical field.

A questionnaire was distributed to the OAtech Network mailing list, aiming to build upon the feedback gained from the interviews and to gather data about OA research from a wider sample. Unfortunately, the target sample was not achieved, however a summary of results is presented in this report.

Findings

It was clear from the interviews that whilst there are a range of data collection processes and methods, there are many common factors and goals as well. Primarily, data approaches are driven by the research question and goals, as well as the resources available. Participants were generally open minded about data sharing and new approaches but were pragmatic about the logistics of implementing changes. The following are the main themes within the feedback:

Machine learning: There was a good level of agreement that machine learning and artificial intelligence offers opportunities to achieve analyses that would not be possible by humans alone, or that would be prohibitively time consuming otherwise. It was also stated that people with the appropriate knowledge are essential in order to develop the correct methods and approaches needed to tackle large datasets and extract meaningful insights. This is both in terms of programming and tools to implement approaches, and also in terms of understanding the research aims and what exactly is being sought within the data. Therefore, collaboration between the OA and data science research communities is essential for success in this area.

Stratification of OA: Although there were clear advantages recognised in the use of large datasets for machine learning, specific discussion on pooling data for stratifying types of OA appeared to be met more scepticism. Participants felt that with current resources and technology, this would represent a large amount of effort which could be better spent on other research endeavours. Moreover, challenges in stratification related to the complexity of the condition due to early stage OA symptoms varying to such a degree it would be difficult to capture reliable data.

Use of current large databanks/datasets: The concept of using data repositories was viewed positively, either for increasing sample size and thus improving statistical power, and for reducing replication of others' previous work. The application and access processes when using these datasets were generally viewed as appropriate on a governance level, but there were varied experiences in terms of ease of access. Hence, further guidance or tutorials could be provided to researchers to ensure these facilities are fully utilised in OA research studies. The OA Initiative was a particularly relevant dataset that was mentioned several times in discussions. A detailed interview with a representative from the SAIL database provides an insight into the use of data repositories.

Main barriers to data sharing

Time and cost of data preparation: The high cost of formatting, documenting and hosting shared datasets creates a large barrier, even if there is a desire to share data. The subsequent reward and acknowledgement is seen as minor and therefore, there is currently little benefit to researchers sharing their datasets, which may have taken years to accumulate with other researchers. On the other hand, some people recognised that all the effort and hard work justifies sharing, to ensure time and resources aren't wasted on other groups collecting the same data in future. Needs a top-down approach driven by publishers and funders.

Data storage and management: It is clear that the issue of mass data storage, particularly in readiness for sharing, is a new concept in the field of OA and as such is not yet governed by any best practice guidelines determining an appropriate and capable data storage solution, and generating the funding for it. The main difficulties within the issue of storage were seen as data security and being able to accommodate the size of the datasets.

Ethical and data protection compliance: While many older studies may not have included any procedures to re-use and share datasets in their ethical approvals, there is opportunity for all

studies in the future to have some statement around this, that would facilitate future sharing. GDPR rules have made it more important to put specific procedures in place, however. This is likely to be an area where the production of guidelines/best practice would really benefit the OA research community and increase the opportunity for datasets to be reused and shared in the future.

Governance: There is an additional requirement to understand and possibly control how the data will or may be used in the future. It was highlighted that the original researchers behind the dataset should always be acknowledged in any future use. There were concerns that governance may be resource intensive and costly to implement.

Opinions on future data sharing approaches

Contribution to a central databank: Participants were open to the idea of submitting data to data repositories. It was felt that this would improve collaboration, completeness of datasets, and the potential to discover new insights more efficiently. However, given the availability of the OAI database, it was felt that anything new would need to provide a different angle to existing robust databases. This was seen as particularly important given the resources required to run a new database. An interview with the coordinator of a UK database for Cerebral Palsy is included to give a detailed insight into setting up and running such a repository.

Standardisation/Frameworks: Although there was strong agreement that there are currently no formal guidelines or frameworks covering best practice around standardising data collection approaches, it was also considered something that would be highly challenging to achieve. It was noted that even where the same standardised measures are used in different studies, they may not be used in the same way or at the same time points. However, the route to achieving a standardised approach could be possible if driven by a large centralised organisation, for example the MRC, who have coordinated the UK Biobank. It is clear therefore that there is a trade-off between standardization and flexibility to collect the variables required in the way required.

Future work

1). Create best practice guidelines for ethical approvals and data protection that ensures research data collected in the future has everything in place to be shared with other researchers.

2). Related to the above, investigate storage and management facilities that facilitate data sharing whilst retaining appropriate levels of control. This could be through national databanks or localised (University) storage facilities.

3). Provide guidance/tutorials for accessing (inter)national databanks, such as OAI, UK Biobank and SAIL. Include how to cost and write these into funding bids and methods of using such secondary to combine with or validate new primary datasets.

4). Ensure there are collaborative opportunities between OA researcher and data science, e.g. through link ups with Alan Turing institute etc. Researchers felt positively about innovative opportunities for collaboration such as sandpit events and links with experts from other areas of expertise, and felt that collaboration should be facilitated rather than enforced through a one-size-fits-all approach.

5) Provide training and guidance on nomenclature within OA, including clinical codes and terminology which could enable researchers to more easily search and make use of data from a wider range of sources. Encourage streamlining of terminology where possible in order to harmonise as many datasets as possible.

1. Literature Review and Databank Summary

1.1 Introduction

Osteoarthritis (OA) is known to be a heterogenous condition characterised by a wide variety of clinical factors, and is a significant health challenge worldwide (Palazzo et al., 2016). Current approaches to treatment remain focused largely on symptomatic relief, however this approach does not yield consistent outcomes due to the variance in pathologies from patient to patient. There is growing evidence that OA may be stratified into disease subsets which may then be targeted differentially and potentially provide improved scope for successful treatments, as well as predictive data allowing for earlier interventions (Driban et al., 2009). Stratification of OA may also present opportunities in research, by harmonising data collection approaches and focusing efforts on disease predictors which may be analysed using modern large-scale methods such as machine learning (Kingsbury et al., 2016).

Recent years have seen major advances in imaging techniques and machine learning, improving upon traditional data collection and analysis methods in OA. Traditionally, radiography has been and continues to be, the standard for OA imaging. However, radiographs have been criticised as having poor accuracy and an inability to detect early OA-related changes. Magnetic Resonance Imaging (MRI) has been demonstrated to provide detailed examination of whole joints and also offers several benefits as a research tool (Changhai et al., 2013). MRI is a non-invasive method of visualising structural changes in the joint from a much earlier stage than radiography, and provides a substantially more sensitive measure of disease progression. These features mean that MRI is a useful tool for expediting research participant screening and selection, consequently reducing timescales for clinical trials. The progression of OA is typically slow, and clinical trial design options are generally to observe participants for lengthy periods and risk high attrition rates, or begin with large sample sizes in order to satisfy power requirements, which is challenging from a recruitment perspective. Stratification using MRI may help to identify participants with the greatest likelihood of a rapid disease progression at an early stage, allowing for an improvement in clinical trial condition allocation (Hunter, 2009). Such improvements may also lead to increased reliability of study data due to larger sample sizes, potentially reducing distortion of effect size results which can be problematic in OA research (Nüesch et al., 2010).

In addition to aiding participant screening and expediting clinical trials, the use of MRI provides a potential opportunity for machine learning in OA stratification. Machine learning is increasingly of interest in health research due to advances in data generation, including the widespread use of wearables both on a consumer and clinical level. Whilst currently the quality of data from such sources is not always sufficient for robust clinical research, machine learning processes are showing great promise in extracting meaningful insights from large datasets through predictive modelling and data mining (Halilaj et al., 2018). Some of these observations are visible and predictable before radiography would detect the same changes, suggesting that earlier interventions are possible (Liebl et al., 2014). Machine learning algorithms run on MRI data have also been shown in some cases to correlate with patient-reported measures of OA as determined using the Western Ontario and McMaster Universities Arthritis (WOMAC) questionnaire (Ashinsky et al., 2017).

Alongside the advances in imaging techniques and data generation, recent years have seen an unprecedented genesis of passive data borne from consumer grade wearable technology, due in large part to the introduction and popularity of activity monitors and smart watches. Huge

amounts of data are now generated constantly by wearers of increasingly sophisticated technology, giving rise to opportunities for researchers to use these measurements as supplementary data to research and primary care datasets. Consumer data may not yet be robust enough to support clinical research, but can still provide valuable insights. It is entirely possible that future iterations of this type of technology may yield data which is robust enough for clinical analysis, therefore it seems prudent at this stage to consider it when planning big data solutions. Consumer grade wearables are already being used successfully in research, providing not only passive data measures which might otherwise be difficult to obtain in a complex condition like OA but also real-time self-report feedback, potentially reducing recall bias (Beukenhorst et al., 2019).

1.2 Challenges in Osteoarthritis data

One of the bigger challenges when considering moving toward the use of large datasets is the heterogeneity of the data, and the resulting difficulties in sharing and harmonising these data in order to pool smaller studies into larger repositories. Even when dedicated effort is made to harmonise multiple heterogenous datasets in OA, challenges remain, particularly when attempting to harmonise data in different languages or using different classifications. Post-hoc harmonisation, whilst still the best option in the absence of access to purposively homogenised data, is time- and resource-consuming and may still not yield robust results. The European Project on Osteoarthritis (EPOSA) experienced such challenges when attempting to combine data from five multinational longitudinal studies (Schaap et al., 2011). The EPOSA study found that the lack of agreement on data collection instruments and procedures between OA researchers was a key factor in heterogeneity of data, and concluded that there is an urgent need for such agreement in order to facilitate pooling of cohort datasets. The researchers felt that longitudinal large scale pooling is possible, but not while such levels of heterogeneity exist.

Currently, there are no standardised guidelines or frameworks in terms of data collection taxonomies in OA, and thus comparisons are problematic. Modelling and predicting OA progression may be aided by standardised data which could make more longitudinal studies possible, and there have been some encouraging applications of this theory in recent years. (Kraus et al., 2015). However, there are a number of practical considerations for sharing data in a large-scale collaborative fashion. Such efforts would need to be regulated, similarly to the governance applied to clinical trial data in general. Even if data quality challenges are met, the governance and ethical challenges remain and are substantial. Large scale data sharing endeavours will need to navigate patient and participant consent issues, as well as guarantee confidentiality and safety of data. This type of undertaking may not be feasible by researchers alone, and instead may be more achievable when a dedicated framework is developed, with a monitored study registry (Peat et al., 2014). In order to ensure that taxonomical and data management standards are being upheld sufficiently to compare data across multiple studies, one option is to update clinical trial registration protocol to include clear data points and collection methods, as well as replicable analysis plans. In this structured and governed approach, trial researchers could also take the opportunity to obtain permission from participants to add anonymised data to a central database *a priori*, and provide records of the required data protection and anonymisation, all in one place, managed by a central authority (Taichman et al., 2016).

1.3 Opportunities from big data and machine learning

The growing amount of robust medical imaging data available to clinicians and researchers provides new opportunities for innovations in analytical techniques in order to derive insights on a large scale. Machine learning is one such emerging approach, and is showing promising reliability in managing complex and vast volumes of data efficiently (Yan et al., 2013). In healthcare, data mining is an increasingly important type of machine learning, having been demonstrated to be useful when working with heterogeneous data. Data mining has many applications to healthcare, including predictive modelling, pattern identification, clustering, association analysis, and more. Data mining may provide answers to the challenges of working with huge volumes of complex data with deeply hidden information, in a crucially expedited way (Chen et al., 2015).

Another way in which machine learning can contribute to healthcare, and specifically OA, is by identifying disease biomarkers at earlier stages. The field of 'omics' (genomics, epigenomics, transcriptomics, proteomics, metabolomics and lipidomics) is paving the way for this diagnostic approach. The ability to determine the presence or absence of specific biomarkers may enable earlier diagnosis, thereby potentially identifying osteoarthritis before the patient begins to experience pain. Biomarkers and omics investigation may even lead to sufficiently accurate disease prediction as to detect pre-indicators before disease onset (Ren & Krawetz, 2015). Machine learning has been successfully used with omics data to enable improved classification accuracy when compared with other feature selection methods (Swan et al., 2013).

1.4 Databanks

In recent years, databanks have begun to emerge, a number of which have demonstrated the potential for success in the set-up and maintenance of an open-access approach, governed by a central authority. Databanks can be set up in a number of ways; data safe havens, used in particular with anonymised patient data, allow access to datasets via a portal login. All data stay held securely on central servers and are never given out to researchers – instead, analyses are carried out inside the portal, or the responsible organisation give out pre-analysed data (e.g. the Secure Anonymised Information Linkage (SAIL) database). Alternatively, researchers can apply for accounts or completed mandated training and gain access to the data, sent to them in a secure format for analysis off site. Many databanks can be used by any bona fide researcher acting in the public interest. Using such large scale repositories can lead to innovations not possible with smaller cohorts, such as developing diagnostic algorithms which rely on accurate big data. The Fracture Risk Assessment Tool (FRAX®) is a predictive algorithm to calculate a patient's 10-year risk of bone fracture, which was developed using population-based cohort data from the USA and latterly from multiple other country-specific cohorts. Though the FRAX® is not a databank in itself, it is a good example of how shared large-scale data can be used as building blocks for new approaches to disease prognosis and treatment planning.

1.4.1 Examples of databanks

The UK Biobank is a national and international health resource which has recruited and monitored 500,000 participants for over a decade, collecting various health-related measures and samples at repeated time points. The Biobank is open access and datasets are available to researchers for any studies in the public interest. Studies can be either exploratory or hypothesis-driven, and can be general or condition-specific. Since the Biobank is open to any researchers acting in the public interest, there is great potential for the dataset to provide insights which would otherwise be prohibitively costly, including large-scale investigation of OA

risk factors (Harvey et al., 2013). Using this resource, Zengini et al. (2018) were able to identify nine new OA loci, using combined genotype data from up to 327,918 individuals – a far larger cohort than a single study might expect to achieve. In addition to the sample size, confidence in the findings was greatly increased by the homogeneity of the data, having been processed and validated prior to the study. Not only is the quality of the dataset greatly improved by having one central organisation managing it, but this further expedites research by removing this step from smaller research teams with limited resources and funding.

The Osteoarthritis Initiative (OAI) is an eleven-year longitudinal study of knee OA, which aims to build a large scale data repository of imaging, biochemical, genetic and risk markers from a cohort of men and women aged 45-79. Anonymised study datasets are available for download at no cost, as well as study protocols and assessment schedules, allowing researchers to match their own study measures and methods in order to create comparable data. The OAI is currently only collecting data in the USA, therefore it is not clear how comparable the results may be in the UK, however the structure of the initiative and the methods used may be of interest to UK researchers, and the study is a useful example of a successfully implemented large-scale longitudinal study with the inclusion of data sharing practices. Since the OAI has been established for several years, there are now a number of studies which have successfully used its data, providing evidence that the use of open-access shared data can contribute positively to the field of OA. For example, Waarsing, Bierma-Zeinstra & Weinans (2015) used the data to explore subtypes of knee OA through cluster analysis of 600 knees. The study was able to conclude that there exist distinct subtypes of clearly differentiated symptoms and causes in knee OA. Not only was the study able to easily test this hypothesis on a large cohort of ready-to-use data, but crucially this was done far more quickly and cost effectively than without access to the OAI data.

Given the complexity of OA in terms of presentation and progression, opportunities to investigate data collected at various stages of disease progression and from a cross section of subjects, large databanks may have a crucial role to play in stratification. Investigation of pain phenotypes in OA can be even more challenging, due to variances in whether patients experience pain at all, at what stage of OA this occurs, individual tolerances, and at what stage medical help is sought. Kittelson, Stevens-Lapsley & Schmiege (2016) used latent cluster analysis (LCA) on data from the OAI to explore whether OA could be stratified into defined pain phenotypes. The study was able to use over 3,000 participants, with several in depth and reliable measures taken at multiple time points. The results supported the hypothesis that specific pain subgroups exist within OA, and that diagnosing patients using clinical symptoms alone may be insufficient for effective treatment plans.

The Secure Anonymised Data Linkage (SAIL) databank was founded in 2006 and now holds over 25 billion records about 4-5 million people. There are no specific OA datasets, however it may be possible that records held in SAIL could provide useful general population demographic information, and data about other health conditions which may be related to OA. A small number of rheumatoid arthritis projects are listed under the ‘projects using SAIL’ section of the website, which is encouraging for potential OA applications. The datasets are not open access and are only accessible after a two-stage application and approval process (with some advanced datasets also requiring further permission from the data custodians), and must relate to a funded project.

The Farr Institute was funded from 2013 to 2018 and was a collaboration between 21 health research partners in England and Wales. This partnership, amongst other big data projects, led

to the inception of the UK Secure eResearch Platform (UKSeRP), an electronic data access portal which allows safe and secure access to shared data, whilst maintaining data owners' control over their contributed data. The platform allows a fully governed approach to data linkage and provides easy and safe access to large scale data without the need for individual researchers or organisations to implement complex data management plans. The UKSeRP also catalogues and quality checks all data uploaded, thereby streamlining the process of accessing and using the data.

The 100,000 Genomes Project, launched in 2012, is a large-scale genomics project which aims to provide the NHS with a set of fully sequenced human genomes in order to observe medical insights and kick start a UK genomics industry. Initially, the project is focused on cancer and rare diseases, as these have been shown to greatly benefit from genomic investigation (for example understanding differentially that certain types of cancer are receptive to particular treatments, which would not be effective on other types of the same cancer). The project is due to deliver its results to the NHS in 2019, and it is anticipated that this will lead to a genomics-led approach to treatment in some areas, with a view to expanding this approach to other viable conditions and diseases in the future.

The Genomics England Clinical Interpretation Partnership (GeCIP) has been set up to facilitate the analysis of the data by groups of researchers, who will conduct their own research projects as well as contributing to the specific aims of the project as a whole. There are a number of pre-approved 'domains', under which researchers can apply to join the GeCIP. Whilst there is no arthritis category at present, there are some cross-cutting domains which may be of interest to researchers investigating stratification and machine learning; *Quantitative Methods, Machine Learning and Functional Genomics* and *Stratified Healthcare and Therapeutic Innovation*. Proposals for new domains are accepted, and researchers can apply to access the GeCIP dataset provided they meet the eligibility requirements (e.g. affiliation with an academic research institution, completion of data governance training) and pledge to contribute to the knowledge base for genome interpretation via active participation in the community. There does not appear to be a fee to access the data, however researchers are required to have funding in place for their projects and no funding is provided by the GeCIP.

The Clinical Practice Research Datalink (CPRD) provides longitudinal large scale primary care and linked data from UK GP practices nationwide, and aims to support prospective and retrospective public health research. The pool contains de-identified electronic health records (EHRs) from 11 million patients in the UK, accessible by researchers in an almost real-time format. Researchers are permitted to access the data against specific study protocols, which must first be approved by the Independent Scientific Advisory Committee (ISAC).

The Health Improvement Network (THIN) database is a collaborative effort between a software developer and UK GPs, which has collated 11.1 million anonymised patient records, as well as over 158,000 comments, all of which are accessible by researchers pending application, a variable charge and full MREC approval. Medical diagnoses and treatments, as well as demographic data, are all captured within the database and therefore it may be of interest to disease-specific researchers for contextual information or machine learning. However, the database is derived from EHRs which may be incomplete, therefore conclusions drawn from analysis of the data may be misleading (Petersen et al., 2018).

1.5 Conclusions

Osteoarthritis research covers a broad range of sub-disciplines, working largely in siloed groups. The resulting spectrum of datasets is heterogenous and there is no central repository, nor any best practice framework or minimum dataset requirement. However, there is growing evidence to suggest that large datasets are of significant benefit to the OA community, and could contribute to expedited advances in understanding the disease. OA research typically suffers from small sample sizes which may affect the ability to derive meaningful insights from the data.

Recent advances in imaging and wearable technology have generated new opportunities in research, including in the field of OA, to access large datasets and potentially pool them to create so-called 'big data'. Determining if and how this might be utilised by OA researchers is a challenge which if addressed in the near future could lead to much faster advances in OA research and treatment. One avenue of interest to OA researchers is machine learning, which may transform the field by taking this big data and developing algorithms and pattern identification previously not possible due to time or resource constraints. By doing so, it is possible that earlier identification and stratification of OA may become possible in a reliable way.

However, such advances will not be possible if data are not shared between researchers, or collected purposively for repositories. To do this requires either homogeneity of data to begin with, or a feasible way of homogenising heterogenous data which is not prohibitively time consuming. Data pooling also requires anonymisation processes which satisfy data protection legislation in whichever countries and organisations in which the data will be used. Anonymisation is in itself a further challenge, as these processes can result in loss of granularity of data. It is unknown to what extent OA researchers share data or access databanks, and how they currently attempt to address these challenges (if at all).

The current project, funded by the OATech Network, aims to find out what the current practices are within OA research in the UK with regards to data sharing and accessing big data. This project will speak directly to expert OA researchers to determine not only what the current approaches are to data collection and sharing, but also to ask what the desired future directions are for the field and whether a best practice framework might be achievable.

2 Research Methodology

2.1 Study design

A multi methodological design was used for this study, having been identified as the most appropriate way to collect data from multiple sources. The study was granted favourable review by the Biomedical & Scientific Research Ethics Committee (BSREC) and the Health Research Authority prior to any data collection taking place.

The following methods were used:

Interviews: One-to-one interviews were conducted with expert participants each representing various sub-disciplines within OA. One-to-one interviews were seen as the most appropriate methodology for these participants partially due to their varying locations and availability, and partially to allow each participant to speak upon their area of expertise freely and individually.

Interviews were conducted either in person at the participant's workplace, by telephone, or via video conferencing. Participants were provided with a participant information leaflet prior to taking part and given the opportunity to ask questions. Informed consent was taken by the researcher, on paper for in-person interviews, and electronically for remote interviews. Interviews lasted up to one hour, and participants were asked questions from a semi-structured guide, so as to allow them to give their feedback on pre-determined topics but also to provide the freedom to elaborate or speak about relevant subjects specific to their area.

Focus groups: Focus groups were planned for the project but due to logistic reasons could not take place. Group discussions were intended to be held with clinicians, allied health professionals and research & development (R & D) staff, from two NHS Trusts (Coventry and Warwickshire University Hospitals and Royal Devon and Exeter NHS Foundation Trust). Focus groups were chosen for this set of participants as it was felt that discussion between participants would yield more depth of feedback than one-to-one interviews. Although the focus groups did not take place, it is hoped that they may be conducted in the future either as an extension of the current project or as part of a future project.

Questionnaire: A 23-item questionnaire was distributed to members of the OAtech Network, to obtain qualitative and quantitative feedback from the wider OA community on a range of data-related topics. The questionnaire was designed based on findings from the one-to-one interviews and therefore incorporated expert feedback.

The questionnaire was conducted online using Qualtrics, and was distributed via the OAtech Network mailing list by an administrator. Responses were anonymous and no personal information was collected. Participants were shown a consent statement prior to beginning the questionnaire and were required to agree before proceeding, with lack of consent agreement resulting in being screened out.

2.2 Recruitment

Interview participants were purposively sampled based on their professional experience and roles. As the OAtech Network is divided into themes (list themes here), theme leaders were invited to take part and provide feedback on their area of expertise. Other academic and commercial representatives outside of the network were also invited to participate based on their experience within OA data collection and sharing or large scale data management. Interviewees were also sought from large UK data repositories (e.g. UK Biobank) in order to provide information as to how these may be of use to OA researchers.

The questionnaire was sent to the OAtech Network mailing list with a brief explanation of the type of participants required. Participants self-selected on this basis, and a screening question was also included at the beginning of the questionnaire.

Detailed participant information including final numbers and demographics is provided in the results section.

2.3 Analysis

Qualitative data from interviews and focus groups was analysed thematically. Audio files were transcribed, and each participant was assigned a unique speaker code. Identifying information was redacted from transcripts prior to analysis to achieve pseudonymisation. A researcher assessed each transcript for major and minor themes, and selected verbatim quotes to illustrate the participants' agreement or disagreement with them. The quotes were then collated by theme and assessed as a group to determine the overall feedback and the level of agreement from participants. As participants varied considerably in their professional background and levels of experience with specific areas of OA research, themes were still considered of interest even when they were not discussed by all participants. The semi structured nature of the questions was also taken into consideration when conducting the analysis, as some discussion points emerged only when deviating from the set questions.

Some interviews were analysed alone, as they were considered to be significantly separate from the main participant group. These included one interview with a representative of a databank, one interview with a researcher who had themselves set up a large database in another area of research, and two representatives of commercial OA companies. For ease of understanding of the results, three additional sections in the results have been added for these interviews.

The questionnaire was analysed as supplementary information to the interviews and focus groups. Since the sample size was not sufficient to use inferential statistics, the results are presented as descriptive statistics and should be interpreted as additional information to support or oppose the thematic findings.

Questionnaire responses were collected anonymously and aggregated for analysis.

3 Results

The following sections describe the findings from the interviews and questionnaires.

3.1 Interviews - researchers

3.1.1 Participants' experience and research interests

Table 1 provides a brief outline of each participant's background and research interests, in order to contextualise their feedback. Participants were sought from several different specialist areas of OA research, though due to cross-disciplinary working there are several areas of overlap (e.g. rheumatology). Please note, IP008 is deliberately omitted from this section due to being considered more suited to a separate area of analysis (see case study section).

Participant speaker code	Research interests and current activities, in their own words
IP001	*"I am primarily interested in physical function, not so much from a gait lab perspective but from a clinical perspective."*
IP002	*"I'm quite keen on if we can find early markers of OA, physical markers, can we develop diagnostic tools and can we treat it before it becomes a job that's only fit for a surgeon, so can we stop people having surgery"*
IP003	*"The role of small non-coding RNAs, specifically snow RNAs in cartilage ageing and osteoarthritis and also the role of extracellular vesicles in the pathogenesis of osteoarthritis and in joint homeostasis."*
IP004	*"Back pain, physiology, pathology, biochemistry of intervertebral disc and articular cartilage. […] Pathogenesis, bio markers, cell therapy."*
IP005	*"I'm a rheumatologist by training. I'm now an academic rheumatologist and a director of a musculoskeletal research institute […]. My interests in osteoarthritis are about improving therapies for osteoarthritis, be they pharmacological or non-pharmacological, symptom or structure modifying. […] I have strong imaging interests as well."*
IP006	*"My principal activity is mapping in the human genome, polymorphisms that are risk factors for osteoarthritis and*

	understanding what those polymorphisms do to gene function, protein function, cell, and tissue function."
IP007	*"I'm a rheumatologist a day a week, and four days a week at the university I'm an epidemiologist, so interested in population health research. [...] And I've done smartphone research, where we've had people tracking their symptoms on a daily basis for a range of different kind of research questions, a fairly large cohort of people with osteoarthritis."*
IP009	*"My interest is in clinical translation osteoarthritis. So, taking the findings from the laboratory and testing them in humans, in clinical studies, and in clinical trials. I have a particular interest in biomarkers, particularly prognostic biomarkers of disease. […] I'm particularly interested in early disease, identification of early disease, but also identifying people who are going to do badly, so that would be following people for longer in their disease course."*

Table 1: Research interests of interview participants, in their own words

3.1.2 Data collection and methodology

Participants had experience in working across a range of disciplines, and data collection methods were varied. Methods and approaches included randomised controlled trials (RCTs), lab analysis, cohort studies, imaging, big data secondary analysis, patient report, gait and movement assessment, and access to clinical data.

> *"Quite often we set questions round about clinical intervention, so we quite like randomised trials where possible. […] We've got a couple of active RCTs, knee replacement for example, whether one implant performed better than a different implant."*

> *"So, we've got simple clinical questionnaire data, so your basics on pain and perceived disability and what you can and can't do and your activity levels, alongside full optical motion tracking data doing… of people with EMG and people doing walking, sit to stand, balance tasks, stairs."*

> *"The big data that I collect is from multi-omics projects so they include NMR or nuclear magnetic resonance metabolomics data, mass spectrometry proteomics data, RNA sequencing data and small RNA sequencing data. So they are sort of global ways of looking at different molecules. [...] So there's lots of big datasets and the datasets can come from not just human tissue but also animal tissues and mainly horse."*

> *"We would do two types of trials. So, we work in fields of pharmaco epidemiology, clinical studies and interventional studies. So, I guess in the pharmaco-epi, it involves using big databases like CPRD. The clinical trials is of two sorts – investigator initiated or company pharma sponsored trials – and*

we do both symptom management work and the structure, understanding structural imaging biomarkers. And I use the term 'biomarker' generally to mean imaging rather than wet biomarkers which I don't do much with."

"We would have quite a lot of clinical data, particularly at the baseline point of the study where we're collecting demographic data. […] The sort of things that we would examine in clinic, you know, is there fluid in the knee, is there crepitus, is there joint line tenderness, those sorts of things. We will also collect imaging information, and often be doing that for the purpose of the study, often x-ray and MRI are the go-tos. And […] at each time point we will be sampling often blood, sometimes synovial fluid, sometimes urine, and then generating biomarker information from some of those."

Though a wide range of methodologies and approaches were used across the participants, there was strong agreement that the methods should fit the research question and that there would typically not be a 'standard' approach. Participants agreed that the methods chosen depended most heavily on what information is required, and then to some extent, the resources available and the availability of access to patients or participants.

"The primary thing we would look at is cohort studies. You need much less of a definitive primary end point for that but we'll obviously have a research question, you can explore a lot of data at the same time."

"It really depends on the study, […] and I think sometimes in the sort of classical, the trials you were talking about, often we would be recruiting people who have very bad, say, symptomatic knee OA, so they are coming into the clinic, seeking treatment from a surgeon or from a rheumatologist, or maybe from their GP and we've identified them at that stage, and they're approached, obviously, given information and decide whether they want to take part. So, that would be the classical situation. Often in the at-risk cohorts, we're actually sort of going out and actually finding people and it may be people who don't have osteoarthritis, but they have had a knee injury, or they are a woman who is perimenopausal, or something like that. We're identifying them as an at-risk group and enrolling them."

"It depends what you're trying to do. In some ways you're at the clinical face and trying to assess treatments and behind that you're trying to bring on new treatments and new developments and then behind that again you're trying to understand other ways of monitoring the disease process… understanding about the biology of it and monitoring treatments and prognostic indicators. So the whole spectrum really of health and disease. […] We know what we want and we gather that information."

It was also felt that by maintaining a flexible approach to methodology and being driven by the research question rather than tried and tested protocols, new opportunities can emerge. For example, advancements in wearable technology can present new ways to collect data which can be used to implement more traditional methods such as validated questionnaires. This participant reported success in delivering questionnaires using wearables, and felt that

particularly for self-report data this could be a useful tool due to being able to prompt the participants to complete the questionnaires at a precise time point. This, it was commented, could improve accuracy by reducing recall bias for example when using pain scales.

> *"We're often trying to kind of say, well we know that there's this opportunity from this particular type of technology, and we know that there's all these sorts of clinical questions. Which of those many opportunities and those questions can we pair together to make kind of a useful combination of this particular problem, [what] can be solved with this particular technology opportunity? So [our] […] study, for example, was a pairing together of the opportunity of collecting self-reported data and physical activity data with that problem in OA that we want to measure both pain and activity at the same time."*

3.1.3 Use of standardised and validated measures

Despite the wide variety of research methods and study types, there were some similarities in the use of standardised validated measures, particularly with those researchers working directly with patients. Participants most frequently mentioned the Knee Injury and Osteoarthritis Outcome Score (KOOS; Roos et al., 1998), the Oxford Knee Score (Dawson et al., 1998) and the Oxford Hip Score (Wroblewski, 1996). Other measures which some participants were using included the Western Ontario and McMaster Universities Arthritis Index (WOMAC; Bellamy et al., 1988), the Health Assessment Questionnaire (HAQ; Maska, Anderson and Michaud, 2011), the Forgotten Joint Score (Azzi et al., 2014), the Tegner Acitivty Scale (Tegner and Lysholm, 1985) and the Patient-Reported Outcomes Measurement Information System (PROMIS; Ader, 2007). Several participants also reported using visual analogue scales (VAS) for pain assessment.

> *"Definitely KOOS and […] visual analogue pain scales, and I think it was the Tegner activity scale."*

> *"I mean, I see things like visual analogue scales for things like pain, and then there's patient global assessments, I mean we go to OA, things like the KOOS and so on; disability questionnaires like the HAQ."*

> *"We've done a lot of work around satisfaction and expectation and they're very simple, single question items. So, you can kind of create your own there, but we've been using the same one for a long time which gives a bit of credibility to it and they tend to overarch quite well. [...] We also, we've used the Oxford knee score […], I've been very involved with the Forgotten Joint Score in the last five years.[...] We've just done a project looking at the new PROMIS score, the PROMIS general 12 item score."*

> *"There are lots of reasons for using different scores, number of questions, applicability, you know, ease of use for patients and researcher but there are certain ones that always come up. So, Americans use the Knee Society Score. They always do, and there are various issues, methodological issues with that because there are patient report and clinician report elements to it.*

[…] Or you've got the European camps all use the WOMAC or the KOOS, whereas the Brits tend to focus on the Oxford hip and knee score, because that's what we've used for the last 20 years and even if you think there are maybe some issues with it in terms of, say, sensitivity to change, actually you can compare to all of this massive literature, databases out there already and see how you're sitting compared to historical or other parameters. There's a lot of strength in using the same thing repeatedly."

Participants were generally agreed that using standardised tests is useful within OA research, but many felt that the current options are not without limitations and using them sometimes requires compromises to be made. For example, some felt that some of the questionnaires are outdated, citing a change in activity levels in more recent years due to older people living longer. Since older people are generally the demographic most likely to experience OA, being able to accurately measure their pain and mobility is important. Additionally, it was noted that language may become outdated and translations of questionnaires may cause problems. Some questionnaires, such as the WOMAC, were seen as potentially expensive. One participant described further validation that their research team had conducted, to attempt to explore how valid tools are outside of the field in which they were originally developed.

"For population health studies, [...] questionnaires are a bit crude and poor recall, and so on, but with consumer technology, we have the opportunity of objectively measuring physical activity, and so [...] we collect symptoms on a smartwatch watch face multiple times per day, alongside pulling out raw sensor data from the back of the watch."

"We were using… ideally using validated tools and we're either performing further validation in different cohorts and very often a tool has been developed but not really validated in, say, revision arthroplasty for example, and we might look at that and see is it still relevant in this field, or if that was designed in that language is it still relevant in this language? So, we sort of use the strength of data to make sure that the tools we're using are still reliable in their different contexts."

"Some people don't like to use WOMAC anymore because it costs too much to use, and so they use an alternative tool called the KOOS"

The Oxford score was seen as a good example of a commonly used questionnaire with limitations, particularly for researchers assessing patients pre- and post-intervention. It was explained that these limitations may cause skewed results when considering patient outcomes, and that the Forgotten Joint Score may be more appropriate in some circumstances. However, it was noted that this does not automatically mean that one questionnaire is better than another universally, and as with the study design, researchers should consider their needs when choosing the most appropriate tool.

"I think the Oxford score is really interesting. […] It's a great, great resource, but it was designed [...] in a patient pool prior to joint replacement to show how much better they get, or to test how much better they get after joint replacement. So the measurement changes from pre-op to post-op. So […] you get a massive skew towards better results post-surgery because on the

whole, everyone has got better. As you contrast that 25 years later possibly, where you've got a much more active elderly population who are expecting more out of life and doing a lot more at an older age, so already the older questions are a little bit limited, they focus on things like getting up, walking up and down the road and getting up from a table, you know, getting out of a chair. […] A lot of the joint replacement patients that I see in clinic now, they want to play golf, or curling, you know, tennis sometimes, they're a much more active group. […] 80 percent of folk tend to do quite well. So, you contrast that, say, to a more modern score, like the Forgotten Joint Score, ask different questions, you know, "How aware of your joint are you under certain situations?" and it was designed on a post-op population. […] But that doesn't mean the Forgotten Joint Score is better than the Oxford knee score, it just means they capture strengths at different points."

3.1.4 Size of datasets

There was a large amount of variance in the sizes of the datasets collected or used by participants, which largely appeared to depend on the nature of the study. Understandably, for studies which involve time consuming data collection methods such as sample collection, lab analysis or the application of markers, researchers tended to report smaller sample sizes. Access to resources and funding were also major considerations in these studies, and participants reported sometimes feeling that their results were somewhat underpowered. However, this was seen as very typical for this type of work and it was felt that smaller sample sizes are an inevitability with certain types of research. One participant, who had achieved larger sample sizes on labour- and resource-intensive studies described the difficulty in doing so, and the associated compromises, time and cost required.

"The problem with doing omics is it's expensive. So to give you a rough guide, to run ten samples on RNA sequencing, so to look at all the protein coding genes in say ten samples, will cost you about £7,000 depending on where you get it sequenced. So it's expensive. So you normally can't afford to run hundreds of samples."

"The smartwatch study that we did, we studied just 26 participants."

We're working on some bigger data sets now for the first time, but most of what I've done is in relatively smaller numbers of people, the tens to the hundreds, rather than the thousands.

"We've got a database of [...] I think it's about 200 people, some have imaging as well some don't. But we've got two sets of studies and one the data quality is higher but we don't have EMG, but we've recruited them through MRI so we can access their imaging. We have about 100 controls, a load of patients at different stages and a load of injured, and then we have about 35 where we have recent MRIs linked to their motion analysis. […] Collecting a database of over 200 people is really quite traumatic so we didn't worry so much about the follow-ups in many ways. More recently we've

collected another 35 from this recruiting them from MRI department and that means we can link it, because what we found is we couldn't prove that our normals didn't have any signs of OA and we didn't know exactly how advanced the OA was. […] That took about three years, three, four years to collect that. It's just getting the people in and keeping the lab quality and the time it takes. [...] It's harder with the older age groups and sometimes with the younger age groups because they have to take time off work to come in, so it's just… it's a lengthy process. […] What we find is it takes ages to marker somebody up and get them ready to test, and then the testing doesn't take that long. […] And of course the labs got to be free and not being used by… there's this whole load of logistics go into it and there's always something."

There was also some variance in what would be considered a 'large' or 'small' sample, depending on the aims of the research and the sensitivity of the analysis. One participant explained that due to this, very large sample sizes can be required for some studies, and having fewer can lead to inaccurate results.

"The differences can be relatively subtle, so many two or three percent differences in frequency, and if your sample size is large enough such small differences become highly significant. […] So when you do these genetic screens, which is the beginning of this type of analysis where you're trying to identify risk polymorphisms, you need tens of thousands of DNA samples. If you've found the polymorphism and you think you know which gene it's targeting, you need several hundred samples from patients to then try and work out what's going wrong."

For other types of research however, sample sizes were much larger, sometimes into thousands. Those researchers who reported larger samples described not only having easier measures to collect (for example questionnaires, or routine clinical imaging), but also the ability to pool these measures once taken. For those using validated questionnaires, there was also the opportunity in some cases to increase their sample size by accessing existing large databases.

"For electronic health records, if it's primary care data, you can get thousands of participants. […] The Clinical Practice Research Datalinkcovers a population of about eight million people, and then depending on what population within the general population you're trying to study will determine how many patients you have. Secondary care data is less widely available for research. We're accessing our electronic health records from a single local hospital.

"So RNA sequencing basically sequences everything in your sample. So you can be trying to map against 20,000 different genes in each of your samples. So they are very large datasets."

For those using physical specimens, pooling samples and building a biobank may not be a viable option due to using the samples up completely.

"We get hundreds but we actually do use them all up, so we're continuously processing these samples and using them in our experiments, so we don't really have a biobank of tissue specimens, but we do use hundreds of these samples. So we'll routinely publish studies in which we've used 200-300 samples from patients. […] that's very good when you're wanting to functionally characterise a gene that you think is implicated in the disease. When you want to identify the DNA polymorphisms themselves that are causing the disease, you have to investigate tens of thousands of samples."

3.1.5 Minimum datasets

Participants were asked whether there are any existing minimum datasets used in their area of research, or whether doing so in the future might be possible. There was strong agreement across all disciplines that there are not currently any formal guidelines or frameworks covering minimum data collection requirements. Some participants observed that there are some common data collection methods across different studies, although they were not aware of any central resource providing information on which researchers are using which methods.

"I think there's so many inconsistencies, how people capture the data, the capture rates, the type of data and we don't seem to have any standards or guidelines to say this is the bare minimum."

"I think with the natures of our studies, we do have a lot of information, so there will be key things; age, gender, BMI, usually ethnic origin, handedness, footedness, usually, […] certainly in the injury area, most people are using KOOS for knee injury, so I know that's luckily a shared asset across lots of the cohorts."

"You'll want different things for different studies, but if you always had that core group of facts then they could all be on one system regardless of the ancillary datasets if the database was designed correctly."

It was felt that since most OA research is designed on a study-by-study basis and methodology is determined by the research question, using minimum datasets would not be practical. It was felt the first and foremost, the study design must be appropriate for the question and this often means different measures and methods are considered to be of highest importance.

"A core data set might look quite different in a clinical trial of knee OA to hand OA to an observational cohort to a cohort that was designed for predictive modelling, so they may have very different things that they would consider absolutely essential. Or a cohort that doesn't have OA yet to a cohort that already has OA. […] if we're going to say, mandate a core set, […] you have to be really clear what settings you are requiring that in and that is appropriate for all the people you are talking to, but I wouldn't be against it. […] I think rather than saying it's a mandated, that this is a guideline, that just having considerations about that this is good practice and these are the people who

signed up to it, and these are the joint areas and the types of research that this might be relevant to."

Additionally, it was noted that even where the same standardised measures are used in different studies, they may not be used in the same way or at the same time points. Preferences for adapted or personalised uses of equipment such as movement marker placement was also a consideration which may make comparing datasets difficult.

"As soon as you start bringing in longitudinal factors, people are seeing people at different times and might have done some of these outcomes in different ways."

"I suppose what you get a lot with optical tracking is everyone wants their own unique marker set because they all think theirs is better, but it then means that there's lots of data out there that's maybe not quite so easy to cross reference and link together."

The participants were divided on whether they felt that a set of core variables could be determined and implemented on all studies. Some felt that this would not be possible for the reasons outlined above, however some noted that the feasibility of this would be improved by being managed by a large organisation such as the MRC or a research council. Participants in support of the idea of minimum datasets with a view to post-hoc data linkage felt that being able to re-use data would be a positive step, particularly in studies which are very resource intensive or costly. This view however, was held only in relation to datasets which could reasonably be collected using standardised approaches and for which post-hoc merging of data would make sense in terms of advancing research knowledge.

"The MRC have set up the bio bank, haven't they, which is a good exemplar of what can be done, so maybe if they set out some key facts and data points that could be collected by each centre collecting samples. You could have a common core that different centres could use, that might be a way to improve it. […] Or just request that if you're collecting samples and patient information, please always collect these things and then you could base your database around that and have other parameters as add ons."

"If you're going to spend three years collecting data it would be nice to know that it could be used beyond what it was collected for because it's such an expensive thing to have the equipment, to bring the people in, to pay their travel expenses, the researchers' time. It would have been nice that we could have optimised the data set. [...] So I think it's more just having some core parameters that when you do you record this because that's really important for the analytics people or for linking data sets, I think that's the biggest message we need to get out of some of these things."

"If you've already got the data collected, if it's not standardised, then your options are either to analyse the data from these discrete sources, based on however you've collected them, and perhaps meta-analyse the results from those studies, if you're trying to answer the second question; or alternatively, mapping them to some sort of common standard, or common data model,

that would allow you to then run a unified analysis script on various different data sources, but you've mapped them to make them look the same. […] I think where you can standardise things, that's useful. It doesn't always make sense to have it all collected in the same way, but where you can, and everyone agrees what the standards should be, then that's useful."

Alongside discussing the idea of standardising data collection, one participant also noted that even when using clinical data there are inconsistencies which make data pooling difficult. In particular, they felt that the nomenclature used across OA is poorly defined, and the ICD-10 (International Classification of Diseases) codes can be too varied for effective database searching. A number of reasons were cited for this, including different paths to diagnosis and different presentations of OA. The participant suggested that a framework could be developed to streamline the codes used, and provide guidance on recoding OA for clinicians so that researchers may more easily use the data.

"There's different nomenclature that people use, subgroups, phenotypes, subsets, various sort of classifiers from that point of view. […] A recent barrier we've had, it's a perennial thing, actually, is just around coding of osteoarthritis in the NHS. So, if you have a knee replacement, there is a code associated with that, and that's fine. But a diagnosis of OA, particularly an early diagnosis, is not well coded and I think they've multi, that's multifactorial, some of it is about the use of the term and when people apply that term, and some of it is just the heterogeneity around the possible codes of things you might call… "Oh, this person has some knee pain," to, "They have gonarthrosis," that's knee osteoarthritis, but a term none of us would ever use but is an ICD-10 code, you know, to various other sort of things. […] so if you are wanting to search for patients who might be eligible for studies, it's a bit of a minefield and not an efficient way. […] if you're running a study in diabetes or cardiovascular disease, you've got much more efficient ways of searching for people. […] There are about three different primary care systems, and we can't change that, but I think probably having some kind of musculoskeletal framework or osteoarthritis framework that encouraged people to use particular codes, to have some guidance there, use them early and be consistent would be really great."

Another participant also felt that clinical data collection could be improved in order to facilitate research, and had been working on this from a structural point of view. They also noted that making changes to clinical data collection would have similar challenges to the minimum dataset approach discussed above, but felt that some positive changes would be possible.

"For the secondary use of data that's already been collected, we are only able to use whatever's been collected in routine clinical practice. […] we're looking to try and structure the data collection in clinical care so that it's then also more useful for research as well as improving the, kind of, clinical utility of the data that you collect, but that's quite hard to make those changes within an electronic health record system, particularly for one small specialty within a very large hospital. But we're trying to do that, and getting some progress there."

3.1.6 Stratification of OA

Participants were asked what they felt were important factors to consider when aiming to stratify OA. However, the majority of participants felt that this was not something which they would consider a primary aim of their work due to the perceived challenges in doing so. It was felt that although stratification of OA might be useful in the future, the adaptations to research required to satisfy the requirements of stratification would risk undermining the research itself. For example, some participants felt that the size of the datasets needed in order to perform the required statistical analyses would not be possible. Creating datasets of the magnitude needed would require either multiple studies with precisely the same data collection measures and study design, or post-hoc data harmonisation or linkage that might compromise the validity of the data. Participants felt that with current resources and technology, this would represent a large amount of effort which could be better spent on other research endeavours.

> *"We accept that some people may have the disease because of a particular genetic profile that may be different from the genetic profile of other people with the disease, so we're cognisant of the potential for stratification but it's not an overriding issue with us. So we don't apply it as some kind of core policy or core principal of the work that we do, but we are aware of the need to be aware of it."*

> *"The more information, the more that you know about the patients and the more patients you have, because obviously you're trying to stratify on multiple things, you need even bigger datasets, the more likely you are to stratify it."*

> *"I think stratification is helpful, and if we think about in terms of pooling data long term, you can't go into the depths that you need and actually I think in terms of linking data, stratification is probably the wrong thing to worry about right now."*

One suggested solution to these challenges was to improve communication between researchers in terms of who is collecting what data, and where opportunities for linkage might present themselves naturally. The participant highlighted that data harmonisation and/or linkage is not necessarily a primary goal as this might lead to pooling datasets with too large a degree of heterogeneity. However, data pooling makes far greater sense when multiple researchers are conducting similar research and working towards the same goal. In these cases, an argument can be made for data pooling, and this could be achieved by first establishing who is doing what.

> *"I think data analysis is quite a good example. So, there are loads of different gait labs up and down the country, they will all have fairly standardised protocol, they will all have healthy controls, they will all have some OA patients and I think OA Tech have been together since our project, where they are trying to get a linkage of that and try and pool ten patients from this lab, and 20 from that lab, and trying to get a much bigger, sort of normal data set of gait analysis of OA patients. That is a fabulous way of pulling it all together, and the way they do that, go to all the gait labs and say, "Do you have any standardised, I don't know, walking analysis of osteoarthritis*

patients? Can we pool them with all of these other labs?" Now, that is a great way of doing it but we didn't worry about what the protocol was, how many cameras were involved, what angle. Actually, it is a case of saying, "Who has got what? Can we start off by saying who has got what?" And someone has done a step down task, someone has done a stairs task, someone has done running, but that doesn't matter. Has everyone got a core thing? And then from there we can start and pull it together. That, to me, makes a lot of sense."

Further challenges in stratification related to the complexity of the condition. Two participants described the difficulties in agreeing when someone has early stage OA, and therefore classifying it. For stratification, this is important in order to determine risk factors and disease progression, but symptoms vary to such a degree with OA that this is challenging to capture in reliable data. This was seen as not only a challenge in applying algorithms and statistical analyses, but also in engaging later stage clinicians and gaining their trust in the results.

"I think it's finding a way of agreeing early stage, because all anyone will agree [on is] advanced stage using surgery. I think we don't have a good working definition of any of the stages, so everyone's talking at tangents. Then the clinicians don't like it when we talk [about] early [stage] because they say, 'well, you can't do that', yet I think we have to have a standard nomenclature and actually much clearer understanding, because at the moment we're talking about a disease that's defined by symptoms and that's the problem."

"One of the challenges is always defining the conditions and in musculoskeletal disease that can be difficult."

"It means that when we use algorithms to stratify people the clinicians go 'well, how did you know it was early?' Because they don't trust maths because they don't understand it, it means that there's extra layers of barriers and we all end up working against each other instead of helping each other come up with an agreed term of something. Because all the clinicians say, 'well you can't define it early stage arthritis therefore it doesn't exist' well it must exist because it's a progressive disease, but within our own community we shoot each other down by not having those agreements."

The definition of stratification itself was also seen as complex. It was observed that researchers are likely to see their own area as the most important factor upon which to base stratification efforts. Within such a broad symptom-based condition, this then may lead to continued siloed working and a lack of common goals.

"Stratification factors for me are quite specific to my area […] I think everyone agrees that stratification is a great idea, and it's almost essential to try and move things forward but what are you actually stratifying? What is osteoarthritis? It is a rhetorical question but what is it? We can give you a clinical definition, but are we interested in osteophytes? Are we interested in joint restructure? Are we interested in bony changes? Are we interested in

cartilage changes? Are we interested in other changes in terms of the muscle and the vasculature? You know, disease process and the genetic changes?"

One participant felt that patient outcomes are a relevant stratification factor, and was primarily interested outcomes and response to treatment.

"I think different people have different risks of outcome, be they because they've been exposed to a risk factor like injury or if they have OA of a particular stage, and I believe that we should be able to try and use data, whether that's clinical data or molecular data or imaging data or combinations of those things, and often I think it will probably be mixed models of those things to actually predict outcome. So, I would see it, yes, in terms of stratification I would see it as relevant to outcome. So, I think it has to have a meaning, and for me, a lot of their stratification is about either where will you end up, or what treatment will you respond to? And I think, from a clinical point of view, I think they'd be my two big questions."

Conversely, another felt that stratification of OA is not possible, and cited their analyses of the Osteoarthritis Initiative, having found no subsets of note within the data.

"we've analysed large amounts of the OAI and we don't see many subsets."

3.1.7 Attitudes towards, and experience of, machine learning in OA

The majority of participants had at least a basic understanding of machine learning within OA, and most felt positively about it conceptually. There was a good level of agreement that machine learning and artificial intelligence offers opportunities to achieve analysis that would not be possible by humans alone, or that would be prohibitively time consuming otherwise. Another major advantage identified was the ability to test a hypothesis and train an algorithm on larger datasets, but then refine it on the smaller datasets which are more typical and achievable within OA research. Being able to develop machine learning tools sensitive enough to reliably detect results in small samples was seen as a very positive opportunity. Machine learning was also seen as potentially beneficial to commercial companies, who could use it to expedite trials of their products. All of the perceived benefits of machine learning were felt to have the potential to positively impact patient outcomes. Some participants felt that the application of machine learning might be more impactful in diagnosis than prognosis.

"It's very useful and it can tell us things that we don't even know about. […] We were training it to use a scoring system which took the radiologist about 35 minutes and of course once you've got the machine learning algorithm sorted, it can be done in a few minutes or seconds."

"What's happened in the past is as part of this machine learning project we've used expanded data, so data from […] thousands of individuals, to see if we can work out what are the best biomarkers for predicting progression, and on the basis of that we've then selected a cohort of 300 patients who we're now following over two years and taking a variety of different measures. […] In this machine learning project, on the basis of analysing several thousand using

databases that are publicly available, we're then able to work out that we can probably pick up effects in 300 and that's what we're doing at the moment. […] What you can do is you can use the data from those to them refine the individuals who you think are most worth following up and then seeing if you're correct, does their disease progress or not?"

"If it does work it could be quite transformative. But it is principally designed to help companies test the efficacy of treatments in a fairly small window of time. Because what tended to happen in osteoarthritis is people progress slowly, companies don't want to do clinical trials of three, four, five years."

"I think there's probably more around diagnosis than there is maybe around prognostic modelling in terms of how solid things look because the system can learn and so on."

"Personalised medicine of course is a fabulously exciting area. You've heard of examples of cancers being effectively cured where medication wasn't working, they've taken immune samples and they have done genetic evaluation of the actual cancer in question and they've pulled out personal factors that will attack that cancer, and they have built them up in culture and then re-inject them back into the patient and that is having a huge effect where that wouldn't on somebody else. You can see how there are potential offshoots to big data approaches, it is just still in its infancy. Where it is all going is of course very interesting"

One aspect of machine learning on which participants strongly agreed, was that since the field is relatively new and also extremely complex, expert help is required. Participants were agreed that specific knowledge is essential in order to develop the algorithms and approaches needed to tackle large datasets and extract meaningful insights. Those who had already explored machine learning in their work spoke positively about collaborators with specialist knowledge, but also acknowledged that due to the infancy of the field there are few analysts with the correct set of skills currently. It was explained that it is crucial not only to have someone who understands coding and the appropriate computer programming languages, but also to have someone who can understand the research aims and what exactly is being sought within the data. This was seen as extremely important in order to ensure that the results are meaningful. For large scale projects, a consortium of collaborators was seen as a way of approaching such a huge task.

"I don't do it, I get someone else in. I think specialist knowledge I think that's the thing, and it's having the data in the right format for them to use. I think the other thing is when we started doing this there weren't many people that knew about it, we had to train the computer scientists to understand where our data came from otherwise they didn't use it in the right way. So it's about speaking the same languages rather than necessarily specialist… I'm sure there is some specialist software but a lot of [it] is their knowledge of what you can do and how you can use that data. […] I had a computer analytics person for a couple of years and she did some papers looking at machine code learning and vector analysis. […] So we've used different analytical

approaches and tried not to just do the standard just trying to compare time points but use more intuitive machine learning and work with the right computer people, there's not so many of them around."

"The increasing prevalence of data scientists, which is something that didn't really exist in academic circles certainly ten years ago, that kind of, the recognition of the need for data management, so it's like 90 percent of the work that we do as researchers, getting the data ready to do the analysis, the analysis part is actually the easy part, but I think as people with those sort of skills are increasingly employed in our sort of health data research departments, then they will bring that knowledge of, and sort of, insight, and the ethos of needing to share these things more openly and widely."

"It does take a bit of time but it also takes a lot of training to be able to do that. So that's not something I can do but it's something that [a data scientist] as a collaborator was able to do with my dataset, combine it with freely available datasets on the internet."

"I'm involved [in] a European Union grant to use machine learning to try and predict who will progress with regards to osteoarthritis and who will not. That's a consortium across the EU, it also has commercial partners, so there's three pharma companies involved in that. […] Machine learning is a core component of that, being used to predict based on biomarker changes in OA patients, whether we can realistically predict who will rapidly progress in the disease and who will not. So yes, we do. […] The work that's been done so far is looking really promising."

Though participants were generally positive about machine learning, there were some words of caution. One important observation was that whilst machine learning can facilitate large scale analyses, large scale datasets are required in the first instance. As discussed in previous sections, datasets in OA are typically much smaller in scale than the numbers required for this approach, and as such it is vital that either data pooling is achieved first, or that the algorithms are trained on existing large datasets. There was a concern from some participants that if not applied carefully and cautiously, machine learning studies would be underpowered and therefore the reliability and validity of the outcomes could be compromised. Access to large data repositories, for instance with imaging, was not seen as something which is routinely available through clinical data but could perhaps be in the future. One participant also urged caution in pooling datasets as this may lead to a risk of homogenising data which may have been better kept separate.

"The numbers in most studies would not be big enough by far. And the problem is, if you look at most x-ray studies of OA, we now know that if you were doing a, even with an enriched cohort, you'd probably need about 600 patients per arm in an x-ray study with a 12-month outcome. And, when you look at most studies, they're 100 patient per arm or 50 patients per arm. […] They're all markedly underpowered. With MR, with cartilage thickness for example, you'd probably need maybe 150 patients, 160 patients in the 12-

month period. With bone shape, you might need 100 patients so the more sensitive tools demonstrate structural progressions with smaller numbers."

"That is the risk, that you're just doing multiple testing and then you find things by chance, or if you're coming up with a model to explain your data, it's just horribly overfitted, so i.e. it works perfectly with your little set of data by chance, but it isn't in any way generalisable to anyone else's, and that's the risk. […] Making it bigger is good, but […] sometimes things are different enough, that it doesn't actually help you, that they're better to be dealt with separately. So, one good example, on the injury cohort that we have, which is really interesting because it was recruited in private healthcare, so immediately you just have a slightly different group of people to if you're recruiting in the NHS, they are 70 percent professional sportspeople, so that's really unusual and I appreciate, again, is not the same as if I go down the clinic and sort of recruit 150 people there with knee injury. […] I would think you would probably lose some of the validity of the overall cohort because there would be very big differences with the data sets that you brought in."

"It's hard to imagine how you would change, kind of, clinical practice around that, but I think there are emerging opportunities of automated image analysis, so were you to be able to access the raw images and be able to analyse that through what would be a, kind of, a standardised protocol, then you could, kind of, consistently define people as having OA or not. Radiographic OA or not, but that doesn't currently exist, you know, CPRD don't provide you with the raw images, but the… I think the infrastructure around health data research is changing nationally and it may be, in time, that there are image repositories of things that are collected as part of routine care, but that's not close yet."

3.1.8 Attitudes towards partnership working and collaboration in OA research

Collaboration in OA research was viewed as potentially useful and important, but not necessarily a widespread approach currently. One participant in particular felt that when collecting tissue, researchers tend to be collaborative due to the difficulty of obtaining samples.

"I think in OA people are pretty collaborative to be honest, because we know how difficult it is to get tissues in the first place, to be able to actually do any experiments on. So within the OA field, I found it pretty collaborative."

"I think it is changing. I do feel that change, there is definitely more willingness to collaborate, definitely. […] I think there's concerns but they are sensible concerns."

Several participants felt that collaboration was difficult in such a small field, as competition for funding is high and researchers can be protective of their data. This was seen as slowly changing, with many researchers being open to collaboration and data sharing if certain

barriers are considered. However, there was a feeling that if researchers spend a lot of time and resources on collecting datasets, there may be a reluctance to share the data, particularly if due credit is not given. It was noted that sharing data beyond the scope of the original study can require significant additional effort in order to prepare it, including data storage solutions and potential costs. Despite this, participants felt that collaborating and working together could potentially improve research outcomes, if effective ways of doing so could be determined.

> *"I think the problem is when you've done that everyone wants you data and nobody wants to give you credit, because it's a huge amount of work but you don't get that much out of that huge amount of work, if that makes sense. [You don't] necessarily get the credit for how hard it is to clean the data and make sure it's high quality and all that post-processing."*

> *" I think there are lots of good, especially trial data sets out there, and the thing about the randomised trial, someone has taken it and done an enormous amount of work, it's taken five, if not ten, years to get the final paper out and there's a fantastic resource and they've written one good paper with it, and that data needs to be packaged and if people could access it, not just for metanalysis purposes but to draw different data and you see that more and more. There are some groups in England are now pooling different data sets to look at certain questions. […] I think the collaboration side of it is still a relatively new thing. So there is obvious resistance to, "Well, I did all of this work and I'm not quite sure how much to package it.""*

> *"Sharing experiences, sharing codes would be useful and we know people don't really do that, and there are reasons why they don't, but also to do with kind of the academic treadmill to an extent, you know, you need to be in competition with others, and so giving away stuff that's taken you a year to do… But then that is in conflict with the transparency that we should have with research, and the, you know, spending the money from research councils and charities efficiently, we should very much should be sharing."*

Some participants felt that there is potentially work being duplicated within OA research, with very similar studies happening and little communication between research centres. This was seen as being due to several reasons, but ultimately communication was seen as a key contributing factor. Whilst participants acknowledged that sometimes similar studies are needed, there was also an acceptance that with improved communication and collaboration, data sharing might be a positive step forward.

> *"I think the problem is as a group of people interested in arthritis if we all pull together a lot of the time we're saying the same messages but we use a different language or different way, and if we could come forward with a better dialogue that shows that we are all saying the same thing we could be much more effective as a community."*

> *"My observation there is that there are lots of people doing similar, yet slightly different solutions, some of which are, you know, you can buy, some of which are academically licensed, all of which look a little bit different, maybe have*

different stage of validity and testing. […] the idea of us all going off and spending our £10,000 developing our app to ask a person X, particularly for patient-reported stuff is challenging."

It was mentioned that journals and funders have the potential to enforce some level of data sharing, by accepting papers and applications only if data are to be made available to other researchers at the conclusion of the study. This participant felt that particularly when a study is charitably funded, there is an increased responsibility to share the data with other researchers in order to maximise its impact. Additionally, this may improve transparency and study quality by allowing post-hoc verification of findings. This enforced sharing approach may also provide solutions for researchers who may wish to share their data but not know how or have the means to do so.

"I think it depends on the person. I review a lot of papers with omics data and people haven't put it in. This is for various levels of journals, you know, from what I would call high impact OA journals, to the less high impact, then people have a... I always make sure that they deposit their data. I basically say I'm not going to accept that paper for publication unless they do that. But some people are very... they're protective over their data, but I see it as... I'm not funding any of my work personally, so in my case it's the Wellcome Trust that are funding it. […] You could say it's their data. But it's everyone's data and so I don't see any reason why I shouldn't be publishing my data out there unless you've got... you're really protective over it because you think it's got some IP or it's just your personality, you don't want to do that, or you just don't know how to do it."

Though many participants felt that collaboration within OA research is possible, and potentially a positive approach, it was clear that this is not something to be forced. Participants preferred to allow collaboration to happen naturally, and where appropriate, rather than being mandated by frameworks. However, it was generally agreed that there may be space for the introduction of resources in order to connect researchers with each other, with expert collaborators/advisors, and to disseminate information about what research is being conducted.

"We can't force people into a model of collaboration, but I think we can provide platforms that help make it easier for people if they want to engage. I think I would probably approach it that way."

3.1.9 Attitudes towards post-hoc data harmonisation and pooling

A number of data sharing approaches were discussed and evaluated by the interview participants, and barriers and enablers of each were identified. Participants did see benefits to having access to larger datasets.

"Coming into it there is a lot of new stuff to learn but I think once we get over these sort of teething processes, access to bigger data sets will obviously mean for better studies."

It was generally agreed that harmonising heterogenous data may be too time- and resource-consuming and may risk diluting or invalidating findings, particularly when resources such as the Osteoarthritis Initiative (OAI) exist and provide large scale data collected in a robust manner.

> *"Now there's also never enough studies going on that are collecting things in a systematic way that may make it worthwhile. And are people collecting data better than was collected in the nine-year follow-up of the osteoarthritis initiative, which is freely available now for anybody to use? […] My concern would be if we take a whole lot of disparate data bases around the UK, with different non-well controlled imaging acquisition sequences on different magnets, with probably no good quality control, locally about standards of images and pull them together, I think we'd be lucky to come up with a few hundred people. And OAI has got 5,000 people in it and it's all well-controlled."*

Rather than homogenising data, participants instead felt that combining already similar datasets would be more appropriate, but only if there is a sufficiently persuasive argument for adding impact to the findings. There were other advantages seen to combining datasets, including the potential for acceptance into higher impact journals.

> *"Well, it's not so much it's homogenising it, it's in science the more evidence you have for something, the more compelling it is. So you tend to combine as much data as you can to show that something genuinely is happening. And the positive side of that is it can get in a more prestigious journal as well. To work in that way you kind of combine data, but you're not homogenising as such, it's providing additional support for a hypothesis."*

> *"You have to have a really good question and have a persuasive reason for people actually putting loads of effort in, because it's quite a pain, the sort of legal side of data sharing. […] You know, is this in the interests of the research that we set out to do, rather than just, "Oh, let's just chuck this data together and let's hope that something good comes out of it." "*

3.1.10 Barriers to sharing data

Though participants felt that harmonisation of different datasets may not be the right choice for OA research, they did agree that in some situations data sharing and pooling may be possible. In discussing how this might work practically, participants were first asked to identify the barriers which might be currently preventing data sharing from happening.

3.1.10.1 Data management and storage

Participants were in strong agreement that the logistics of sharing data are the biggest barrier, and felt that storage was a considerable challenge. There were two main strands to this challenge – determining an appropriate and capable data storage solution and generating the funding for it. For non-physical data, a cloud database was seen as the most appropriate solution, but setting this up was not considered to be simple. The main difficulties within the issue of storage were seen as data security, and being able to accommodate the size of the datasets. It was clear that the issue of mass data storage, particularly in readiness for sharing,

is a very new concept in the field of OA and as such is not yet governed by any best practice guidelines.

> *“Because I’ve got some long term cohorts I understand the need for… sort of long term data management, that was never an agenda when I was doing things ten years ago and I think it’s only experienced researchers are probably coming to this now, people are all starting to twig this is an important thing.”*

Concerns were raised about the responsibility of ensuring that data sharing can be conducted securely, including preparation of the data and also users downloading it safely. Participants mentioned using so-called ‘safe havens’; secure portals designed to allow access to sensitive data without the need to transfer it or download it. The suggestion of cloud storage was seen as viable for the management of such large data, but there remained questions about who would take responsibility for this, how it would be funded and the protocols and processes by which it might be managed.

> *“So remember these sets are massive so really you're thinking you're not going to be able to download it onto your computer. You either need to download it into a cloud or into a server at your university. So when the biotechnicians are analysing this data for us, or if we had analysed our data ourselves, you need a big memory in your computer to be able to do that sort of thing, because the datasets are so large. So when [the analyst] did it she would have downloaded it to a server at her university. Because she was within the bioinformatics department, that's what they did, they just dealt with big data so it had massive memory capacity to do that sort of thing. But she still had to say, “I've got all this data to come in...” because obviously it's using up some of that memory, so she had to get permission to do that.”*

> *"You have to have a safe site to download things to, and you have to prove that you’ve got all the data security on your sites before you can get downloads. It’s quite laborious and complex. And it takes many months after you take a download before you can clean all the data up and start to do anything with it […] so you need big servers set up to deal with this, and then appropriate software for dealing with big data.”*

> *“It has to be carefully approached and thought through and there have to be clear analysis plans and data management plans, so you can’t do it in a kind of half-baked way. […] you still have to harmonise your data and have the right field names, you can’t just do it after, drop everything in, and then sort it out afterwards.”*

> *"I think that the kind of concept of safe havens is one that’s evolving and I think universities are sort of slowly working out what they’ve got to do and how they support them. Sometimes, I know of instances where there are multiple safe havens in single institutions, so they working out what they do about that as well, is it a physical local storage, or is it stored in the cloud, and again I think this is, kind of, a moving target.”*

With digital data, there was also a concern about the size and format of imaging files, which are not only very large files but also often stored on NHS systems where anonymisation is not necessary. Should these files be required to be downloaded or stored elsewhere, agreed upon anonymisation protocols would be essential, which would present new challenges. The process of removing patient identifiers from imaging is problematic; this can result in either reducing the usefulness of the data by also removing key information, or conversely can miss elements which would make patient identification possible. For example, researchers may need to know the date of an MRI for analysis, but in certain datasets having access to the date and the patient's diagnosis may reveal the patient's identity. This may compromise ethical boundaries in some cases and would need to be considered if and when images are transferred from NHS secure systems to local research systems.

> *"In terms of can you construct such databases, well the next issue is it's relatively easy for patient reported outcomes and demographics. Much harder once you get to imaging data and being able to pool and share such data. The big issues are where do you store, DICOM files which are very large. […] And I'm talking that because the majority of the OA research around the country would be using MR, not ultrasound. And MR images, DICOM images are large, stored on people's routine hospital PACS systems, where they don't have to be anonymised, because only relevant clinicians can access them. But, for research purpose, they would have to be anonymised in a very good system before they could be shared. […] You need special software that strips all identifiers off it. And the problem is, if you strip off all the identifiers, it may adversely affect the image analysis that's done later where certain types of image analysis need to know some things about the sequences. So, it would have to be set up very carefully from the start."*

In addition to the challenges associated with digital data, participants who routinely worked with physical specimens felt that storage and transfer would be difficult. Due to the nature of specimen storage, concerns raised were around maintaining the integrity of the samples (particularly those which require temperature controlled storage), and ethical transfer agreements.

> *"I think there may be some issues where if you've collected certain biological specimens, for example serum or plasma, then you may use them for a study and then they need to be frozen down and stored, there are sometimes issues over long-term storage, who's going to pay for that, how retrievable will the samples be and things like that.[…] Say you worked on 200 serum samples from osteoarthritis patients, you would publish the data, make that data available to others but if somebody then wanted to work from those serum samples that's slightly more arduous because it's a physical thing. […] And that would be harder to resolve because then you will have issues over ethics and material transfer agreements and all of that kind of stuff, but more difficult."*

3.1.10.2 Ethical considerations

Another major consideration when discussing data sharing was ethical clearance to do so. Participants noted that as studies often span several years, the ethical applications for studies ending now were written prior to the idea of data sharing becoming more common. Therefore, many ethical documents make no mention of data sharing, or perhaps explicitly state that this will not happen. In these cases, seeking consent from participants retroactively can be problematic – if the ethical approval is based on documents which state that participants will not be contacted after their participation in the study is over, then it is not possible for the researcher to contact them in order to request permission to share their data. It was also noted that since the introduction of the General Data Protection Regulation (GDPR) in the UK, researchers are held to more stringent ethical guidelines.

> *"I think is an issue within sites and between sites. So, within sites because we often, especially when we set things up 10 years ago, didn't think we want to come back and dip into things again. We didn't think about those issues. And sometimes we did, so what are the issues? Ability to re-contact people, it has to be in your consent forms. Ability for the data to be shared with other databases. […] We've slowly got better and it's a bit of our corporate memory about making sure that your in-house patient information sheets are updated and contain some broad statements about further contact, and sharing of data and using broad terms as much as possible. […] But, of course, you have to identify your patients if you're going to go back, and how did you keep a record of them, why did you keep a record of them when you shouldn't have after the finish of the study?"*

> *P002: "If we could go back to the normative date we've collected in the past ten years and follow those people up. It's getting the permission and the ethics to do that along with those things that with data protection to be allowed to go back and follow people up, even if it's just an email saying 'can you tell me if you've now got knee problems? […] But having that access to what happens later and to a degree the ability to reassess them would be really valuable, and that's something we all forget to do or don't think about."*

> *"GDPR has changed everything. So, what you might have said previously, you know, "We're going to use your data, anonymised data for research purposes," you know, tick the box kind of thing, that wouldn't be enough anymore. So, data that has previously consented under the old framework would be fine, whereas now, the transparency agenda that's come in really in the last year, the formalisation of that, I think everyone is changing how they tell patients what's going to happen to their data."*

It was also highlighted that not only must consent be taken for future sharing of data, but that researchers hold a responsibility to be clear with participants about what their data may be used for. This was seen as important not only from the perspective of informing the participants and obtaining true informed consent, but also at the later stage of determining whether to grant access to other researchers, and whether their proposed use would meet the description given at the consent stage.

"Apart from GDPR, it's the issue of what did people give consent for? And most people in their studies weren't thinking five years ahead, or 10 years ahead, or pooling their data with other people. […] This to me is main issue number one, it's how do you get the community to include certain phrases, like you should be providing phrases and we'd say, 'Put these, make sure these are in your ethics'. Because our problem was, for example, [another University] contacted me recently about a shoulder study we did a decade ago, and we were very happy to share data with them, and they could have had all our patient data. It was an investigator-initiated study and we just went back and were really worried that our ethics would not work and that it was not inclusive enough."

It was however, generally agreed that people who participate in OA research are happy for their data to be shared providing there is a well established rationale for doing so. Participants reported more recent ethics applications having been updated to include the option for data sharing at a later stage, and felt that their participants showed no changes in willingness to consent since introducing these clauses. Changing consent forms and ethical applications was not seen as burdensome, although some participants felt that researchers do not always remember to include this option. Several participants reported that they now include keeping data for as long as possible in their ethics applications, to allow them some flexibility.

"We write an ethics that means that we have this specific question in mind but it's possible that we may have alternative questions in the future and are they agreeable to us using their data and their samples to answer those questions as well. So we write the ethics so that it's kind of all encompassing."

"We've noticed that OA patients are delighted that somebody is investigating their disease and wants to know something about it, and when it's put to them that these samples with basically be burnt or they can go to the lab and people could discover more about your disease but the data that is discovered won't necessarily come back to you and help you in your treatment, they're fine, they're absolutely fine."

"I think it really varies and I think it's about getting your PPI involved early on, and most of them if you explain why they get it. […] it's just having the conversation with people and making sure you express it the right way in your consent things, but so far we haven't had actually any issue."

"It all depends what's in your ethics. We try and go for longstanding ethics; most of ours are for 20 years."

Furthermore, it was felt that some academics are still unwilling to consider future data sharing, for various reasons. In some cases, it was felt that OA researchers are protective of their data and do not wish to share it. In others, there was a perception that researchers who are close to retirement or who may move on at the end of their projects do not recognise the benefit of sharing data after their involvement in the project is concluded.

"Having that access to what happens later and to a degree the ability to reassess them would be really valuable, and that's something we all forget to do or don't think about. Because you could argue that in ten years' time I'll be retired so why would I put that in my ethics? But I think we've got to be more about setting the next generation up and about the research that needs to be done rather than a lot of people just think of their own careers."

3.1.10.3 Governance

A further challenge identified when discussing data sharing was the management of the process itself, and governing appropriate and ethical use of the data. Participants agreed that the responsibility for this must rest with the original data custodian, and as such there need to be robust processes in place to maintain data security. This was seen as time consuming and costly, and potentially outside of the expertise of the researchers depending on the set-up required. There was a cautious attitude towards the practicalities of data sharing, and a recognition that the impact of doing so improperly would be serious. Participants also felt that where secondary analysis has been completed, the original researchers should be properly credited, and there could be guidelines for doing this appropriately.

"Above all you have to be sure that appropriate data is being safely released, or safely used. I know that in an organisation it is paramount. [...] So, you have got to be very careful, we have got to have processes."

"We need to know that they're using the data appropriately. I think you don't necessarily expect to be an author on a paper, but you would expect to see 'this was…' the work that collected the data referenced and the acknowledgement that they have shared the data with you, which you don't always see."

Participants who had experience of setting up (or beginning to set up) data sharing processes felt that there were different ways to navigate these challenges. One described a panel approach, whereby researchers wishing to access the data would write an application, which would then be assessed against predetermined guidelines for data use. This proposed approach was also seen as appropriate in terms of anonymisation, and could provide clear guidance for anonymisation standards and procedures. Having clearly defined processes and responsibilities would also mitigate legacy issues such as researchers leaving or retiring, or proprietary systems which become defunct once the responsible party is no longer in place.

"I think there would have to be some sort of panel [...]. So they can apply to a panel that makes sure that you adhere and you can then say 'yes, you can have this data on these grounds' and you sign up to it, and there's all the governance that goes with it. […] Within that you can say, 'you need to recognise this work and that work with how it was collected' or whatever. I think there's a way of doing it administratively that wouldn't be too burdensome but would allow people to use the data appropriately, and whatever you do with it maybe you have to run it past some committee to make sure you have done that properly."

It was acknowledged that the resources required to facilitate this process would be significant, and may require full-time administrative responsibility from someone outside the research team.

> *"I think we actually need someone whose responsibility it is to manage it and administer it because I think it's actually quite hard to do, because everyone comes in and thinks their programme for cleaning the data or processing the data's better. […] I think at the time we did that study we had a proviso that the data could be shared in future studies as long as it's anonymised. So it's making sure we set up future studies in such a way that data can be shared in an anonymised form, or that we can go back to people, follow them up in five or ten years' time, so that's one of the things we're trying to do now. Then we're looking at the best way to store and back-up data and get our heads round the new Data Protection Act."*

One participant was actively working on a shareable database, which was in the early stages of development at the time of the interview. The participant described the administrative processes planned for their data sharing safe haven, and felt that the complexity of the task meant that time must be taken to ensure that each step is carefully designed.

> *"We are working out how to share it. We took consent so that we could share it with others, but we haven't done so yet, but we're, as I say, working out where it would sit. We want to put it in a… rather than sending out data sets to people, we want to put it within a safe haven and enable access to other people. […] We said that they consented to us sharing it with people where we saw that as being… you know, where that was in our control, rather than taking consent for it to be freely available.[…] Conceptually, in terms of data sharing, we plan to store the data in a safe haven, or a trustworthy research environment, where we have control over it, there's an audit trail of who touches the data and what they do to it, there's a process whereby you can only take things out of the safe haven once it's been checked by the staff of the safe haven. There would be a data access application process; there would be a review of that application; there'd then be... we have data sharing agreements that people have to sign and sort of terms and conditions of use as well in terms of how the acknowledge the data source."*

Another described using an externally developed open-access database software and adapting it to their research needs, with support from their organisation's IT department. This was viewed positively and working well, however this solution was being used internally only, for audit trail and data management purposes.

> *"We've used the software to build a database, basically we have IT people here who help us to do that. We had various solutions over the years, but I'm hoping this is a more sustainable one, because it's sort of an open access platform that universities can subscribe to. I'm not sure if there's even an actual cost involved, and for academics you can just use them freely, so it's quite good. So, it means you can have a secure database for your data that's hosted in an appropriate place, and it's compliant with the various things we*

need to do. So, for example, if we edit things, that that is all tracked and so on."

3.1.11 Use of data from databanks and databases

Alongside the potential to share study data between researchers, there exist a number of databases and databanks with purposively collected datasets, or curated and collated data from primary sources. Participants were aware of several of these data repositories, and had varied experience in accessing them and using the data. Some of the examples mentioned were the Osteoarthritis Initiative (OAI), the Clinical Practice Research Datalink (CPRD), the Imperial Tissue Bank, REDCap, OpenClinica and the UK Biobank. There were also institution-specific databases run by Universities and research centres.

The concept of using data repositories in itself was viewed positively, either for increasing sample size and thus improving statistical power, and for reducing replication of others' previous work. One participant felt that using these datasets was potentially a viable alternative to costly and time consuming randomised controlled trials (RCTs), should the relevant data be available. Those working with tissue samples were able to access specimens which would otherwise go to waste, by applying for them via a databank.

"I think there's people we could learn from in different contexts that have… so we don't reinvent a wheel or find problems that other people have already solved."

"If there is data available for... so for my group with ten samples, someone else's group with ten samples, clever people can combine those datasets and then you increase the power of your analysis."

"We have a musculoskeletal tissue bank here and we will tend to use that to acquire tissue samples that would be essentially waste tissue for [joint] replacements."

"Registries are increasingly important now for studying lots of things. I think they're getting more respect as well. Indeed, they are sometimes suggested as an alternative to clinical trials, because RCTs are incredibly difficult and costly to do."

Other uses of databanks included using existing data to answer a specific question and generate/support a hypothesis, and then following this up with a more bespoke original study in the lab. The use of databanks in this context was seen as a cost-effective method to prove a concept, which could then be developed further.

"So if you're looking for risk factors for osteoarthritis and you're doing a genetic screen you can use subs databases. So the most informative one so far is the UK Biobank.[…] So within the UK Biobank there are measures relating to the musculoskeletal system, so it's possible to identify individuals that do have osteoarthritis and then do a genetic analysis of those patients, and that's already been done and lots of new osteoarthritis risk genes have been identified through that study. But once that's done, that just tells you the

genetic signal. The next thing is to go in the lab and try and work out what that genetic signal is doing to gene function."

Other participants had successfully used databanks to conduct their own analyses and publish new insights from the data. One described a study which had benefitted from the addition of data and had identified insights which would not have been possible otherwise.

> *"We've published, our first paper out of it was a pharmaco epidemiology paper looking at drug use, of what Americans were using for osteoarthritis therapy. And subsequently our papers have all been around imaging studies on proportions of the OAI and we're now working on much larger datasets from the OAI."*

> *"We will take specific research questions and we'll draw information from the database to answer these questions. […] So I'll conceive a project and I'll go to the data custodian and the data administrator and say, "Right, I want to run this project over the next year or two, I want to collect this data," and we'll get that approved to take part in the database and they'll collect the data for me and then we'll take the data out and we'll run that through analysis."*

> *" I had a […] computational biologist who managed to take other datasets online and combine it with my dataset. So my dataset was... it was a small sample size and because of the difficulty in getting different tissues, it was a mixture of sexes. So you can imagine sex probably has a role in musculoskeletal disease, well, we know it does. We know that it has a role in osteoarthritis and we now believe it to have a role in tendon disease. So when [they] combined my dataset with other datasets online […], [they] found that the change... age related changes are also dependent on sex. If she hadn't had access to those other papers [we] couldn't have come to those conclusions."*

The application and access processes when using these datasets were generally viewed as appropriate on a governance level, but there were varied experiences in terms of ease of access. Some participants reported positive experiences, whereas others felt that the process was a steep learning curve and somewhat bureaucratic. It was agreed however, that stringent protocols are appropriate to protect the data.

> *"It's normally pretty easy. You do have key words. It might take about half an hour to get your head around it but it's pretty simple to do. […] You just press the buttons and download it."*

> *"It has to be there to comply with the Human Tissue Act and all of those things too, so there's all of those things and the ethical [processes] are all set up in such a way that they don't make it too bureaucratic that you have to apply for 15 things before you can take a blood sample and do something with it. So it's been thought out quite well and the governance of it's been thought out quite well."*

"So, it's not a simple thing where anyone can just say, "I want this data and I am going to run these tests." It is still… there is still quite a big process and you have to be working for a public agency or a university as research, you know, to have the infrastructure in the first place to even try and apply for it. So, everything makes a lot of sense in hindsight, it's just like the first time you try and do something there is a lot of new things to learn, but subsequent applications will be a lot more straightforward."

"It's a large, complex data set that you have to, kind of, slowly come to understand. […] It is a learning curve with it, like there is with most things and once you understand it, then it becomes that bit more easy to use."

Once in possession of the raw data, processing and preparation can still be a considerable task. It was described that taking a collaborative approach and sharing data preparation programmes on open-source platforms can help other researchers do this more efficiently in future.

"There's a whole data preparation step that is complicated. I mean, we've done a lot of medication safety research using CPRD, and when we first did it, it took us, like, over a year to go from the receipt of the raw data to the data ready for analysis, and we've written, kind of, programmes and scripts to make that more efficient, and we have shared those on GitHub and Zenodo repositories so that other people can do that more efficiently than we did, to begin with."

One difficulty which was noted was once again the inconsistency in coding OA, and the impact this can have when querying the CPRD. This echoed the feedback in section 3.1.5 about the challenges associated with coding OA and the different pathways to diagnosis that can lead to varied disease classifications.

"It's not so much about what the codes mean within the system, because you often have, kind of, guidance about what they do mean, but actually how a GP practices. […] If they saw a patient in front of them with this particular problem, how would they go about coding it? Because we're working with the output of what they do in their electronic health records system."

One participant had extensive experience using the OAI and felt that it is a very easy system to use. This participant compared using the OAI with using the CPRD and found the OAI more user-friendly, as well as having the benefit of being free to use.

"You can sign-on and register very quickly, and then if you want lots of images, you have to send them a terabyte disk and they'll just download all the images onto a terabyte disk and send it back to you. […] Well they used to do that, but it might be you now download it from an FTP site, from a distant site. So, rather than the physical sending, I suspect you download it. And the people I know who work with it are all reasonably happy. It's like CPRD or any other big database, you have to get used to all the variables, and you know the naming of all the variables, but it's pretty good. It's been pretty good to

work with. […] OAI is free; CPRD has a cost and it takes a lot of expertise to play with it, it's not something anybody can just take a download. You have to write an application to get in and they look for experience of using stuff. So CPRD covers I think maybe 10% of the UK population currently and has GP life records just about for those."

3.1.12 Contribution to databanks and databases

Typically, large databanks and databases do not accept contributions form researchers, and instead collect their own data to ensure consistency (for example, the OAI). Alternatively, some databanks collate data from primary sources such as the NHS. Therefore, participants had not submitted their data to any of the larger databanks or databased.

However, participants were open to the idea of submitting data to other repositories – either those which already exist, or in principle should an appropriate repository become available in the future. It was felt that this would improve collaboration, completeness of datasets, and the potential to discover new insights more efficiently.

"Within the OA field it could be MRI datasets, x-ray datasets, all that sort of thing, where there isn't necessarily a repository for you to stick that data. So if that was available and then other people in other fields would be able to collaborate, they'd be able to increase their sample size. […] So if for instance you had the information on the MRI scans of the patients that you'd then taken cartilage from and then done the sequencing, if you could collate all that together, then that would be brilliant."

"We ourselves are perfectly happy to do it so long as there are means to do it, and that is facilitated principally by the journals."

"If it was more flexible and the outcome measure that we have always used for our knee patients was a key dataset, then we could put ours on."

Despite a willingness to do so, participants reported that sharing data in this way is not standard practice for the majority of researchers, for various reasons. One reason given was a lack of awareness, either of the existence of such repositories or of any requirement to submit to them. In particular, one felt that in their field of research there is an expectation that datasets are shared at the point of publication. This participant felt however, that not all researchers would know this and therefore may not be submitting their data. The participant recalled not knowing this earlier in their career and thereby missing the opportunity to share data.

"I think people need to be aware and I'm not sure how many people are aware. If you're doing these sorts of studies all the time then you're obviously aware. The first time I did a paper I didn't realise that you're supposed to put... the day somebody said to me... and the reviewer said, oh you need... you should put your data on a repository. So from then on I did. But until that point... it was my first study, I didn't realise that you were supposed to do that."

They went on to feed back that it is becoming not only best practice to share data, but in some cases a requirement from funders or journals. Others also reported experiencing this approach and felt that it was a positive solution.

> *"The solution is that if you want to publish in a good journal they say you have to publish your data on a repository and they tell you which repository and those repositories produce the highest standards, so would include their end values. You make people know it basically. […] so if you want to put your data in a... your paper in a good journal then the journal says you have to deposit your data on X, Y or Z repository and those repositories say you have to include this information about your data, that would be the best solution, I think."*

> *"Journals are becoming quite insistent on things like that, and so are the funding agencies. So they will fund a project [...] then they do more or less insist that that data is then made available for others if they want to do secondary analysis. So yes, we do routinely make available our data for third parties. […] It's normally in a format that you've created for your own usage anyway, so it's just a case of uploading it in that format and making it available."*

Some of the participants however, felt that creating new databases for researchers to upload to would be time- and resource-consuming, and may not yield sufficient results to be worthwhile. The participant who had worked with the OAI extensively, again reiterated that they felt that anything new would also need to provide a different angle to existing robust databases. This was seen as particularly important given the resources required to run a new database, and the robustness of the OAI. Similarly, it was noted by another participant that in the field of OA, much of the data can be considered international and therefore the UK does not necessarily require its own duplicates of databases elsewhere.

> *"It's interesting in a way. The NIH in America, the equivalent of our MRC, has driven a lot of that, so they've created a lot of databases. So science is international, so you can often say to a funder or a journal 'We'll deposit it on this database.'"*

> *"We tried early on in the piece to get some sort of national registry of studies for osteoarthritis to understand what people had and it was just on a no funding basis; it was just too hard to develop that dataset. And also, the problem we had when we started to look below the surface of sharing of data and providing combined datasets was ethics. […] If we're going to do something in the UK, it must be different from OAI, it must provide something that differentiates."*

Other participants, whilst supportive of using large databases in principle, felt that there must be a solid basis for doing so and clear funding plans in place to manage the data. There was some caution from participants in terms of whether the need for the datasets outweighs the resources required. Those who had access to and regularly used clinical data appeared to voice these concerns, perhaps reflecting a variance in need for supplementary data across

disciplines. These participants also noted that currently there is not necessarily a suitable repository for their data, due to differences in the format of the data or requirements from the databases that specific measures are collected which currently are not.

> *"Ultimately, I think there's two questions' is well what are you asking, what are you going to find, and how are you going to find it? You know, what's the point? And also, who's going to pay for it? Because that sort of level of databasing and banking in a non-hypothesis driven way is really expensive. […] I can see in some ways the attraction of it and you could put some resource at it, but I don't think anyone, us or anyone else has sort of the current resource to do that unfortunately."*

> *"We would put [our data] onto the [International Cartilage Regeneration & Joint Preservation Society] registry already if it was more compatible. […] If it was more flexible and the outcome measure that we have always used for our knee patients was a key dataset, then we could put ours on. […] It is happening, it's not the best database and it's always difficult getting everybody's data in the same form to collate something in a common way, but it is there and it is working to some extent. We haven't entered our data into it yet because it would mean a lot of reorganisation of our data ... there's some glitches in their database that would be difficult for us at the moment to comply with. But, we've created our own database."*

In these cases, researchers tended to have their own internal databases where they sometimes pooled data from different studies, but only for internal use. They were not necessarily opposed to the databases becoming open to external collaborators, but referred back to the comments about ensuring this is done correctly, and requiring a level of funding and resources not currently available.

> *"In theory, it was open to people outside of the university, but I think in practical terms that hasn't really been happening, again, really, just from a resourcing point of view.*

> *"I think we've got such a good database for what we want from our data that I don't know that we would get anything more at the moment. But like all of these things, they develop with time and it's a good thing to support. I'd like to support it, but for example, one of the outcome measures that they insist on having, we haven't been using on most of our patients. So you can't go back and take those outcome measures retrospectively."*

3.2 Interviews – commercial representatives

3.2.1 Participants' professional roles

Invitations to take part in an interview were sent to commercial representatives who were known to be involved with the OATech Network, of which two chose to take part. Both of these participants were from orthopaedic manufacturing companies; therefore, the following feedback is from this perspective.

> *CP001: "I work for an orthopaedic manufacturing company, so we create prosthetic implants, at the moment hips and knees, […] assistive technologies, pre-operative planning, pre and post-op analysis, things like that, to improve surgical outcomes in patients and patient outcomes as well."*
>
> *CP002: "We're a medical device manufacturer. We work in the field of orthopaedics amongst other things so primarily my group looks at developing hip and knee prosthesis and the technology that helps you put those in, so a lot of our work is around surgical workflow and instrumentation and development. […] We are interested in the development of the disease, particularly in osteo formation, because that kind of affects how our treatment path may all go."*

3.2.2 Research involvement

Participants were asked to feedback on how their companies engage with research, and how this fits into the commercial activities of the company. Both participants felt that their companies were very actively involved in research, and that this approach underpinned development of their products. There were strong connections with both clinical and academic research partners, either by funding fellowships and such like, or by working with surgeons to obtain pre- and post-operative data in order to assess the relative success of the devices or products. There was a clear understanding of the importance of scientific research, and accurately measuring the impact of the devices. Working to improve the sensitivity and accuracy of outcomes measures was also seen as important. Participants felt that the attitudes towards research were not limited to their companies, and were considered to be industry standard.

> *"Previously surgeons would have measured success of the operation on revision rate. So as long as they're not revising any of these joint replacements, then that's considered a success. However, in today's society, that's been deemed not sensitive enough."*
>
> *"All the pre-op data, the intra-op data and things like that, is our surgeons' data, so we often publish or present on findings from that data. The company I work for is very, very scientific, clinically focused."*
>
> *"We fund, as many companies do, we fund fellowships. So, you know, we pay for younger surgeons to go and spend six months or a year with a senior*

surgeon to learn their tradecraft. And often an expectation of our funding is that that fellow produce clinical papers or clinical studies. […] if there's imaging or a gait analysis or something like that that's needed to prove a theory or to prove a concept, then we also would pay for that. And then we also provide research assistance and such to collect data and to write up papers and things like that. So, it's a very heavily involvement from the company from that point of view."

"And that's not unique to us, that's industry-wide."

The type of research conducted either by, or on behalf of the companies was varied, and was deemed to be needs-driven. Due to the nature of the products being developed, different types of study are required at different stages of development, and different types of assessment are needed depending on the type of product. For one of the companies, teams and projects are often funded or formed specifically for each product. From a funding perspective, this often results in a mix of self-funding from the company and grant funding from public sector bodies.

"Like a lot of large companies, we have both internal and external research, so we partner with various universities to undertake research. We also undertake some stuff ourselves."

"We run everything from randomised clinical studies in support of products right the way through to fundamental research that's part funded or fully funded by ourselves. I guess we literally cover all the bases. We have very much a clinical focus research which we'd be performing with hospitals, often teaching hospitals which are linked to universities, and then as I say, right the way through to the more lab-based fundamental research."

"I will say that whilst we try and do research that isn't purely connected to our projects, the reality of the way our funding works is that most of our research is linked to a project. So, if, for example, we were working on a new finger implant then we would use that finger implant project to fund research that would support that project."

The needs-driven approach was also reinforced by describing how projects and teams are initially set up; the company assesses what resources or expertise are available internally, and what must be outsourced in order to carry out the research effectively. This approach ensures that the company is not overstretching in terms of resources or expertise, and can run as many or as few research projects as required at any one time. In addition to this, partnerships are often formed with external clinical or academic teams from a data access perspective, as the company does not have sufficient patient data, but this can be more easily accessed via surgeons or universities. Therefore, some projects are run with external partners for this reason.

"I guess when I mean funding I'm talking about the internal resource that's required to do anything, so whether that's people, resources or literal money, a certain amount of that needs to come from the organisation and then, once that funding is available internally, then we would then say might fund a team, a project team, and that project team might then go and either fund research

directly, so that will be wholly funded by us as an organisation, or we might go out and look for grants that we can apply for to support the work."

"I think often it's about capability and expertise, so we identify that we don't have the expertise internally and then we identify the expertise externally and work with them. […] We have worked with a third party, whether it was an academic party or a company, purely because that third party had good access to surgeons and patient data."

"So even if they had no expertise, no capability beyond what we could do ourselves, we would still reach out to those people just because they have that access because it's such a challenge for us."

"Even when we're working doing research internally it tends to be in really close collaboration with academic partners."

3.2.3 Types of data collected

For one company, there were many different data streams of interest. Of particular interest was an imaging project involving each patient completing a pre- and post-operative protocol to generate imaging data. This database has been populated using full resolution imaging data collected in partnership with radiology clinics and will eventually include a battery of patient measures which will inform pre- and post-operative comparisons.

"We are embarking on a telehealth platform at the moment which every orthopaedic company is doing for data collection, for patient engagement type things. […] the big thing for our company, we have quite a unique pre-operative planning programme which requires post-op imaging in the form of, we call them functional X-rays, X-rays showing specific patient positions in flex seated and full extension and contralateral leg step-up positions, paired with a pre-op CT. […] We have a series of over 12,000 pre-op CTs."

"Each patient that goes through our planning protocol, [...] They need to go through a series of X-rays and CTs, so that's our big data pull at the moment."

"We have a direct link to the radiology, so we validate radiology clinics and we have a direct link with the radiology clinics, so we get the raw imaging data."

"The imaging is for the pre, so basically from that imaging, we develop a patient specific plan for their surgery, so the idea here is that we have a more accurate and more personalised surgical plan for individual patients. And the idea there is that we're improving the longevity of the implant and patient outcomes."

"This is sort of the telehealth thing, so you know patient prompts, pain levels, sleep patterns, activity levels, those sort of things, which we haven't fully launched as yet, but this is something every company's doing at the moment,

so the idea is we're sort of saying all these things that we're doing improve patient outcome and now we've got to prove it through these objective and validated prompt scores which is your patient reported outcome measures, like your Oxford knee scores and hip scores, and things like that."

"You can attack it from a few different ways, so it's wearables or […] the ability to pull the data from your smartphone."

The advantages of this approach include being able to obtain patient measures in real time, as well as reducing the need for patients to attend clinics whilst simultaneously increasing the amount of data collected due to the passive nature of wearable technology. Using telehealth also allows for longer follow-up periods, though patient compliance is expected to reduce after 12-18 months post-operation.

"The idea here is to make it mobile health, so instead of patients coming into surgeons' clinics and filling out paper forms and things like this where it's very, very manual. [...] it's very, very dependent on the surgeon's preferences, and then they get that data."

"We haven't implemented it yet, but I would suggest it would be most likely a year, but we may well even want to push out to 18 months, two years, I think. It would probably become more of a compliance thing from that point of view. I think most patients 12 months post-operation have sort of moved on and almost forgotten about it, so I think the compliance rates, with interaction with a rehab tool, basically that far out's probably starting to push the limitations of their compliance capacities."

The other company was more focused on using existing data for secondary analysis and to investigate new ideas before using them to form the basis of new products or interventions. One of the frustrations that this participant had encountered was the difficulty in finding sufficient database records for the conditions needed. Though there are some records of the type required, the order of magnitude would be in the thousands, which is not presently available.

"We've used retrieval data banks, so retrieval centres. I have tried to find data banks with the information I need in the past, but I've struggled to find them."

"At the moment, I'm really interested in large data sets of pre and post op imaging and whether I was looking in the wrong place or ignorant on what's out there but I'm unable to find what I need."

"So for example, if I'm talking about a hip application I would need full pelvis, hip x-rays, AP x-rays and probably lateral x-rays, basically the routine views that you use for applying surgery, but what I've found is I think it's the cancer research has an image database, when they go into that image database in theory it has thousands of patient images but by the time I've subdivided it down into what I actually need, which in my case is patients with osteoarthritis in their hip and then an x-ray which is representative of the type

of x-ray you would take in surgery, the numbers go right down to just a handful."

Again, it was emphasised the importance of working with clinical partners to overcome this difficulty. Whereas clinical studies may yield small sample sizes per study and would take far too long to generate big data even if pooled, having access to clinical data points can achieve this in a fraction of the time and expedite the product development process.

"If you think about us needing thousands of data points and our clinical studies are generally in their hundreds and we're not necessarily running clinical studies continuously, we're probably talking about hundreds of years for us to actually collect that data naturally. […] If we start thinking about CT which we also would like, then again that's not something we would routinely get necessarily in a clinical study, whereas if you look at some of the large orthopaedic centres that are out there that do routinely CT, if we were able to effectively get it from hospitals we could probably collect that data within a year."

3.2.4 Data usage

Similar to the academic researchers, the commercial representatives described a needs-driven approach to using data. Study design is based on the research question, and the datasets are explored accordingly. Unsurprisingly, 'big data' is becoming more and more useful and important in order to train new algorithms or look for patterns. In particular for the manufacturing companies, large repositories of imaging data tend to be among the most useful due to the amount of detail available. Additionally, the ability to compare pre- and post-operative data is crucial to determining the success of devices and surgical techniques.

"It's really dependent on what you're looking at. So, for example, if we just want to look at something that looks at variation between patients from our pre-op CTs, I mean we do that retrospectively, we've got that huge database of CTs, so that's something that we can pull up very, very quickly, and there's no need to wait X amount of time."

"Almost all orthopaedic companies now are looking at the information they collect and then how it could be used for, you know, big data analytics or predictive analytics in the future."

"The idea here is that information from our pre-op imaging, from our intra-op experience, from our pre and post-op prompts collection, all becomes de-siloed and starts to interact together so that we have this wide spanning picture of individual patients and a better understanding of what their outcome has been, or will be, and so on."

Another benefit of larger datasets, in particular those rich in detail such as imaging, is the ability to test theoretical models and concepts on big data, which can then be refined as needed. As well as allowing expedited testing of new concepts, access to pooled data means that commercial companies no longer need to rely on academic partners for every project. This

was seen as useful when the needs of the manufacturer are not necessarily aligned with the interests of researchers.

> *"For example, we have modest CT scans and I think it's in the region of about a hundred, and we use that to explore statistical shape modelling. So, we're really interested in skeletal morphology and how that differs across populations, so we do that kind of morphology studies as well. It's a lot of similar stuff that you will see in the literature, so things like patient specific FE models and then progressing from those into statistical shape model based FE models so you can start thinking about how your implant might perform in a population. Really, our aspiration is a lot of the research we do externally so ten years ago that's exactly the work we were funding externally but it's now got to the point where research partners don't want to just do the same work over and over and over again. As far as academic partners are concerned, it's done, that research is no longer research. It's now a technique that you can apply and so they are very keen for us to then be able to apply those techniques internally and that's what we would try to do. We do that with things like physical testing as well. We develop the test methods with a research partner and then we would bring those methods in-house and scale them up and again we might use patient data say on gait cycles, so feed into those kinds of tests."*

3.2.5 Databases

Due to the need for large datasets, commercial companies are embracing the idea that pooling data into usable databases can provide a valuable internal resource. Creating this, however, is still in its infancy. Companies (non-competitive) are working together to try and pool data, but this is technically challenging and will take time. Collaboration is seen as important, and key to achieving big data in the future.

> *"This is something that every company in the industry is sort of moving towards, but probably isn't quite there enough as yet. So, obviously, when you're talking about predictive analytics and things like that, the datasets that you need are huge. And so, I think probably over the last few years these personalised planning sort of focuses have now been realised. We sort of need to understand, okay, now that we've got all of this data, how do we use it to further patient experience or further improve patient outcomes? So, I think this is something that the industry is definitely looking at, but it's probably too premature to understand how successful it's going to be as yet."*

> *"It's very much each company is doing, you know, we're probably all doing a similar thing, but we don't talk to each other […]. it's not just the implant companies that are active in this space, we've got the university groups and things like that, that have a less of a commercial interest in the joint replacement scene, but they still have these fantastic datasets, so we often*

collaborate with [third parties]. […] So there's a significant collaboration within the space, just not between implant companies."

"We'd love to say that we had a really coherent well-curated database. Unfortunately, that's not the case, so a lot of data gets encapsulated within a project so, as I said, it's funded because it's relative to a project and then the reports, the data associated with that will get documented and captured within the signed documentation for that project. So that's the most formal place where stuff gets captured. Our clinical team are doing a much better job of keeping a curative list and they're actually using a third party company to help manage the patient image data so you can, for example, say could we pull all the x-rays of study one, two, three, four, five and they will be able to pull that information back so that they are building that database."

3.2.6 Ethical processes

Both commercial representatives were very conscious of the ethical responsibilities when collecting and using data, and described robust consent procedures. Due to the way patient data is used, companies are careful to ensure that any commercial venture is not compromised by ethical issues, therefore consent is taken for all intended uses of the data. Retrospective consent has been taken where needed and allowable, when data use changes and is not covered by the initial approval.

"Any time we request something invasive for a patient or something like that, it needs to pass ethical committees, and so on and so forth, so it's a very regulated sort of component of the industry I suppose."

"Consent is a huge thing for us. […] I think we're hypersensitive to making sure that… because we basically use the data to field the commercial products and what we don't want is at the eleventh hour to find we have the inappropriate consent and that we have to pull data because that would affect significantly what we're doing. So, I think a lot of our energy goes into making sure that we actually have the appropriate permissions in place to allow us to use the data in the way we want to use it."

"Everyone's interested in data and it's becoming more of an issue […] so when we're doing a clinical study, clinical trial to support one of our products, through those clinical studies we will routinely collect pre and post op x-rays but what we haven't done routinely is consent participants for anything other than their data being used in that study. So what that means is we have a large amount of data that we can't really use for anything because it doesn't have the appropriate consent, so what we've now started doing is including two levels of consent, number one, do you consent being part of this clinical trial and then, number two, do you consent to your information being used for applications beyond this clinical trial?"

"We've had situations in the past where we've had to retrospectively go out and get patient consent because we've spotted something that we think is interesting, so we'll publish. Maybe if we were doing some video research, so ethnographic research where we were in operating theatres videoing surgery and we had all the appropriate consents to allow us to do that, but then having done the work we actually thought this is really interesting, we can publish on what we've seen but then when we look back at our consent we realise we haven't actually covered that use of the data. We'd only covered the use of the data internally and so we went through the whole process of going back out to all the participants and getting them to consent for this new use of the data, but you can imagine that that is a significant undertaking and we couldn't get that extended consent in quite a few cases from the patient."

3.2.7 Approach to data sharing

Commercial participants described a very collaborative approach to data sharing, in order to work towards common goals and potentially speed up or improve the quality of results. However, as with the ethical and consent issues mentioned above in terms of internal data use, any sharing is also stringently monitored.

"Yes, again, we're very, very protective of our database, because it's obviously very, very valuable for us, but it's also very, very strictly regulated as well. So, any collaboration where data is shared needs to be looked at from a value perspective from our company. You know, we don't want to be sort of giving away all this hard-earned data unless there's a significant upside for us, but then we've got to be very, very careful with GDPR regulations and things like that in equivalent countries."

Both companies were open to allowing publication of findings by their academic partners, with some exceptions where commercial sensitivity may need to be considered. It was seen as a benefit in most cases, for the information to reach the public domain, to increase awareness of product-related issues and to drive innovation. Understandably, data sharing did not extend to competitors or anything which may compromise intellectual property (IP) rights.

"A lot of our research that we do, we're quite open in letting our academic partners publish. We don't tend to prevent the publication of anything. There's also a level of detail that we want to retain but we're quite supportive of publications, so I think a lot of our research ends up in the public domain."

"One restriction we'd definitely put on is if we think there is IP related to it, being able to patent something, then we basically request that we are allowed to process that patent application prior to them publishing that and we'd work hand in hand with them to make sure that that worked. […] If we're talking about fundamental research that relates to the development, we might ask them to postpone publishing those research findings until we've been able to launch the product."

“There is an argument there about we tend to develop products based on what we think is sound science and sometimes we think it’s useful to get the science out ahead of the product to kind of prime the market about what’s coming and let them know that there’s this interesting research topic we found that we believe then our products can address.”

3.3 Interview – case study of cerebral palsy database

3.3.1 Background

In order to find out about the pros and cons of setting up and running a large-scale condition-specific database, an interview was conducted with an orthopaedic surgeon who had been instrumental in setting up a database for children with cerebral palsy (CP). Although this participant was not involved in OA research or treatment, it was felt that their insight would be valuable due to the similarities of the conditions. Cerebral Palsy is, like OA, challenging to identify and treat due to being a varied condition which presents in different ways at different ages. Historically, it has not been well recorded in terms of incidence rates and large-scale data, and IP008 wanted to address this with a new approach.

> *"The variation is enormous, not only in the severity of CP, the disease itself, but also in the way it's treated, the way it's assessed. Within our own region we've got huge variations in practice from physiotherapy, orthopaedics, speech and language, paediatrics."*

> *"There's registries out there everywhere but for some reason there was never a cerebral palsy registry, not in the UK anyway."*

The participant described a doctor in Sweden who had previously set up a system for children with CP (known as CPUP), and that this was the inspiration for the new UK database. Not only was the intention to capture data, but to standardize the treatment pathways within CP, which can be varied and can result in patients being 'lost' in the system, particularly when they reach adulthood. The introduction of the CPUP in Sweden was designed to register all children with CP over time, eventually leading to a much fuller picture of the rates of CP and efficacy of treatment.

> *"So in Sweden, that's going back about 15 years now, a guy called Gunnar Haglund developed something called CPUP [...]. So the idea is that not only do you collect some fairly basic demographic data about patients with cerebral palsy, also their diagnosis, their subtypes and various other different bits of information about their type of disease, but then also standardising a pathway of assessment. So now not only are we collecting children with cerebral palsy, we're also collecting some meaningful data about them. So take physiotherapy, for example, one of the main difficulties in assessing children with cerebral palsy is that different physiotherapists do it in different ways, so the idea is to standardise this so as part of the pathway we would educate physiotherapists. Actually, not so much education, that's the wrong word because they all knew what they were doing, it's just making it consistent so that we were all doing it the same way, so some people would learn things slightly differently and so forth, and I evolved my practice as well."*

3.3.2 Introduction of the UK database

Using Gunnar Hagland's CPUP as inspiration, IP008 worked with a team who introduced the Cerebral Palsy Integrated Pathway (CPIP), first in Scotland and later the West Midlands of England. As of the date of the interview (July 2019), approximately 1,100 children in the West

Midlands and 4,000 in Scotland had been registered on the database. These numbers were seen as promising, particularly as the Scottish database pre-dates the English database by three years.

> *"In the West Midlands we've got this thing called CPIP, so it's the British equivalent, so Cerebral Palsy Integrated Pathway, and that is an online database that's hosted by the Health Informatics Centre in Dundee. You have to be particularly trained to access it and it can only be accessed from an NHS M3 computer, and once you're on the system you can then start entering data. […] It's been running in the West Midlands now for three years, so it's been running in Scotland now for six years, so CPIP Scotland was the first UK one to run and that was based heavily on the CPUP, the Swedish model [...] But we've been doing it for about three years and we've got about 1,100 kids on our database now. Now we're lagging behind the Scots who have got nearly 4,000 on their system but they are kind of three years ahead of us."*

IP008 described how the database approach helps identify and track CP patients over time, and works with rather than against the nature of the way CP presents. Rather than wait for new patients to receive a diagnosis, the database is instead primarily populated by physiotherapists when they see children with CP. A new entry is initiated on the first time a child sees a physiotherapist, even if they are not newly diagnosed. It is at this point that their treatment plan and notes are entered, and they can then be updated on subsequent visits. This approach means that over time, all patients should be on the database and have detailed notes, however there is no conclusive way to know when all children with CP are registered.

> *"It's incidence is relatively low, it's only about two or so per thousand births. So if you do that, if you just have a new patient approach, it will take you ten years before you really got population level data. So we took a slightly different approach and basically said if a physiotherapist is seeing a child with cerebral palsy, they can then enter them onto the system, plus all new diagnosis would automatically get entered, any new diagnosis plus anyone who's currently on the system. So the advantage of that is that we'll get more data more quickly. The disadvantage is that we still don't quite have a… we won't know for a number of years if we've got everybody so to speak, because if there are people that have been lost to the system and never seen a physiotherapist, they won't then get entered onto our system."*

The newly introduced database is intended to provide a more systematic approach than previously used. The hope is that by rolling out the CPIP nationally, the database will gain traction and be able to provide two benefits; a more structured and consistent method of assessing children with CP, and a searchable database which may be used to track progress and treatment efficacy on an individual and cohort level. One problem when assessing, diagnosing and treating CP is that like OA, it does not have clearly defined codes which can lead to difficulty collating data.

> *"It's not correctly coded then you don't get the right data, and cerebral palsy has got a number of different ways it's coded. […] Because cerebral palsy is*

quite a disparate group of conditions that all come under the one umbrella term, it can be quite difficult to find it."

As nothing similar currently exists in the UK, the CPIP provides an opportunity to create a uniquely useful clinical and research resource if it can be introduced in the right way.

"All the core information about the patients are identical, and the idea being when we get UK-wide data we'll have one of the biggest databases in the world."

To do this, buy-in is crucial from ground-level staff, particularly physiotherapists. In order to effectively implement the use of the CPIP without increasing workload, physiotherapists are asked to introduce it into their existing assessments, but it has been important to ensure that there is as little duplication of effort as possible. Including CPIP training as part of core competencies is being considered for the future, and it is hoped that the format of the information in the database will make the patient data more usable without impacting workload.

"The physiotherapists put in the lion share of the workload. They're the people who really do the assessments, so having them on board is critical, more critical than any other unit."

"In order for a physiotherapist to fully assess a child and enter it onto the system it takes around 45 minutes, and it's usually a two or three person job. The way we run it in Coventry, we have two therapists and then a physio assistant who effectively just documents everything. So it's three people for 45 minutes, which is quite a big workload."

"The way it's been sold by the therapists actually centrally is that this isn't an addition to your practice. It's a change in practice, but there are some that see it as an additional workload and additional burden to have to do these assessments, whereas previously they didn't have to write it down in this way and so forth. A lot of people were keeping the information but it wasn't quite the way… it wasn't in a usable fashion really. […] because this has been led by their national group, and in fact they're writing up a piece to say this is part of your core competencies as a physiotherapist to be able to assess children with cerebral palsy, a paediatric physiotherapist that is, and most actually have come on board."

Consequently, the CPIP has been designed to be a clinical tool first, and database second. As part of this, the fields aim to match as closely as possible those already used in clinical assessments and on the proprietary NHS systems. Simplicity of design has also been key, with an intentional compromise between a 'tick-box' interface for speed, consistency and accuracy of data entry, and some resulting loss of granularity. To mitigate any loss of detail, there are also free-text boxes and notes sections for users to add anything not captured by the standardized format. To reduce duplication, the CPIP system is used first, and the data then exported as a PDF to other NHS systems.

"We've tried to make it as seamless as possible. So what we'd like people to do and what most are doing is choosing your CPIP first. So you go to CPIP and you put all of your data in there, and then what the guys did in Dundee was a lovely little bundle just called Get PDF, and effectively then that exports everything you've just done, every measurement, every comment, all the traffic lights, exports it as a PDF and then that can be usually seamlessly uploaded into electronic patient records, or printed out and put in paper records."

"Basically, it just exports it as a PDF, so none of that data is analysable unless someone is actually doing contact recognition and converting it and so forth. [...] we're very conscious of the fact we didn't want to duplicate workload. […] It has been written such that all the fields are in the right order and they've got the right labels so that any other NHS system should be able to then extract the data which minimum fuss."

"Some of the criticisms about it so the therapy side of it is literally tick boxes and it's designed that way so it can have database functionality. The more words you have in there, the harder it is to search for things, but obviously then that does… you miss out some granularity of the information, so there are other information boxes where therapists can type in other things and so forth."

For imaging, it was decided that in order to include X-rays without the need for vast storage to be required, the compromise must be to reduce the quality. As such, in comparison to the full resolution images available on a PACS system, there is data loss. However, it is now possible to include a snapshot of X-rays on the CPIP system. To do this, the user must manually log in and add the pictures which is additional effort and has met with resistance from some users. One solution to this was to create a student project from this task, and have the images uploaded in bulk.

"The quality of the X-ray that's uploaded to CPIP is much, much less… is effectively a jpeg capture of the picture on your screen, so it doesn't have all the data that you have on a PACS machine. […] So you lose all that data, but obviously the benefit is it's not a 20 megabyte file for every picture."

"In order for me to put an X-ray data on I have to log in and physically do it, so that's an additional task on my day, and you get some people who just don't want to do that. So the way they sorted it out in Scotland was effectively they had two medical students who wanted to do it as a summer project, and they got all of the X-rays from the CPIP database in Scotland and uploaded them and wrote it up and published it and so forth, so that worked quite well for them."

3.3.3 IT challenges

Trying to work seamlessly with existing proprietary systems has been an important consideration in setting up the CPIP, since it would not be feasible to replace them, nor would it be possible to integrate the CPIP into the other systems. To best address this, it was decided that the CPIP would be hosted centrally by one organization, and would then be accessed remotely via the various Trusts. An additional benefit to this approach is that the central host organization, the Health Informatics Centre (HIC) is already set up to provide this type of service and as such possesses the necessary expertise and resources to do so, allowing a faster and more robust roll-out. The HIC have championed the CPIP which has helped with uptake.

> *"There are lots and lots of different systems that each individual clinician uses, and God forbid having the IT skills to get all those to talk to one overarching system that will talk to every other NHS system, so I don't think it's an easy job. So that's why CPIP we've very strongly supported having it hosted by one person or one group that would then, rather than have everyone develop their own systems around the country, I think partly because HIC, Health Informatics Centre, are inexpensive and very good that everyone else has been happy to jump on board. Their website is incredible. It's blisteringly quick. There's no waiting around. There's no fancy graphics or animations or anything like that, it just does what it needs to do and it works on any computer no matter how fast or slow. […] It's all hosted by them, yes, so nothing is stored locally, not that I know of anyway."*

> *"Some of their senior leaders at national level have been really strong advocates of CPIP, and driving things from there, cascading down to senior regional paediatric physiotherapists and really driving things forward, so that's been hugely beneficial."*

The HIC have also provided technical support which has greatly enabled the onboarding of users. For example, common requests from the database have been programmed into the system to save time and generate reports quickly.

> *"There are a number of requests that are so common that HIC basically programmed them into the website. So I can go on, for instance, and click a button that says X-ray data missing, and it will give me a list of patients that are outside… that their X-rays are out of date, so I can see I've got six patients who need an X-ray and I can request the X-rays on that. They've got about 11 or 12 of those things where effectively I think they were getting requests so many times for this data they just built them into the database, but that's only for my patients, not for the whole nation, and that's the same for a therapist. […] so I log in and I go to the report section. I click on X-ray missing and it will give me a list of X-rays missing."*

The system may also in future support the NICE guidance that CP care should include strong communication across teams by facilitating the easy recall and sharing of patient information. It is hoped that this will act as in-built auditing and provide important evidence for future clinical pathways and guidelines. The hope from the CPIP team is that the database will be adopted

nationally and become 'too big to fail', thereby becoming embedded in practice and subsequently influential on governance.

> *"So there have been a number of guidelines published about the management of children with cerebral palsy and similar conditions, so NICE have got a spasticity guideline and cerebral palsy guideline, and then the recent NCEPOD report, called Each and Every Need, which was about children with physical disabilities, all of those… in fact, NICE is brilliant because page one, paragraph one, chapter one is all about having a network team with good communication between members of the team, and one of the advantages of CPIP is that if a physiotherapist enters some data I can instantly see… so I see a patient in clinic and I can just look and see what their last CPIP assessment was and vice versa, they can then look at my orthopaedic part of it and see how this kid's had Botox or operation or whatever, so it facilitates that kind of team working. In the NCEPOD report there was a lot of emphasis on good quality assessment, good quality data collection and all that kind of stuff, so CPIP allows you to tick that box. So I can now go to my Trust, if they have a CQ6 inspection, and say, "I've got good communication, here's the evidence for it. I've got a good X-ray screening protocol, here's the evidence for it. I've got good assessment protocols and here's the evidence for it."*

3.3.4 Information governance

Alongside the need to work with the different computer systems within the NHS, a larger challenge has arisen when approaching Trusts, in particular when considering funding. Despite low running costs, each new region must allocate their own funding, so a robust business case is essential in order to justify the additional cost. For the longer term, it is hoped that the database can be centrally funded and the Trusts can simply adopt the system into their own practice, rather than need to cover any cost. With this in mind, it is hoped that the database can become a standard part of practice rather than a cost consideration.

> *"At the moment each region that wants to come on-board is having to effectively find their own sources of funding."*

> *"We've prepared business cases which we're sharing with all our colleagues so we don't have to re-invent the wheel, to a certain extent, and it's one of those projects that everyone I've spoken to have said it's a good idea. […] I think it was about 20,000 at the time, which was enough for me to set up the database and run it for three years in the West Midlands. […] Now the database has been set up the running costs are incredibly small, so it's £3,000 plus VAT per year for the whole of the West Midlands, and that's a population of five million. So the money is tiny, absolutely tiny and Health Informatics Centre in Dundee have been brilliant. It's a very, very good website. It's incredibly basic but very, very quick. You can access it incredibly quickly from anywhere. It's got really, really good database functionality and they run it on a shoestring. They're a not for profit organisation."*

> *"We're trying to say that, "You don't need any money, we've got the money for the database, that's all funded, it's just a change in your practice rather than additional practice." Most have been okay."*

The structure of the NHS, which frequently changes, has also proven challenging to navigate when introducing the CPIP. Governance issues have been problematic, particularly in light of recent changes to data protection legislation in the UK. The perception from IP008 and their team is that there is a willingness to use the CPIP, but that there is a strong risk-averse attitude towards the required approval processes.

> *"I think there is a genuine willingness. [...] It's just that every single Trust wants to have their own approval, and every different individual making that approval has got different queries or different things, they've argued about certain sentences in the information sharing agreement and so forth. Some Trusts seem to be a bit more careful than others about having their patient identifiable data held on an offsite database, even though it's within the UK."*

One concerns which has been raised by some Trusts is the issue of the data being held offsite, and not on secure NHS systems. In addition to this, there were some concerns about patient consent for secondary use of the data, by researchers. This is important to address, and IP008 felt that these concerns are similar to other patient management systems and consent processes already in use in the NHS.

> *"When we went through all this with NHS England, because this was a patient management system the argument was, which we made and they agreed with, was that this is no different to having your patient data on any electronic patient record, and you don't require patients' consent. If they come to you for treatment then they give tacit approval that you can keep their data on as long as it's in a secured NHS system. […] However, we do get them to sign the consent form and that's partly so we've got some approval for using their anonymised data for audit and research purposes, and we make it clear that for audit and service evaluation it would all be anonymised. [...] so if someone wants to then use data from within the CPIP system, they would apply to it. There would be a group of us that would say, yes, that's a good idea or not and then they'd have to go through the normal ethical approvals in order to use that data, and that would be anonymised. So patient level data, identifying data can only be accessed by a therapist or a surgeon within that Trust with access to an NHS computer already."*

Coupled with the effort required to navigate the information governance with each Trust, the number of individual Trusts, each with their own set of approvals, has further complicated the roll-out of the database. IP008 felt frustrated that it is not currently possible to obtain one approval from an umbrella organization, which can then be ratified locally on a smaller scale by individual Trusts.

> *"We've had a real difficulty in England because just in the West Midlands there are 14 different Trusts that look after children with cerebral palsy. Each one of those Trusts has put up various barriers about getting IG approval in order to… they're already paranoid about having their children's data— […]*

every single Trust wants to do things differently and wants to check all the documents."

"It just seems bizarre to me that we can't go to someone in NHS England and say, "Please look at all this and if you approve it then all the other Trusts can then say, 'Yes, this has been approved nationally, we don't have to'.""

3.3.5 Clinical benefits of the CPIP

Key to the success of the CPIP is, and will continue to be, illustrating its benefits to clinicians and decision-makers. Having access to insights from the original Swedish CPUP has been invaluable in being able to demonstrate the possible future impact of the CPIP. There have been encouraging results so far which can be attributed to the introduction of the CPUP – not least because the data itself is now available in a coordinated and consistent format. Beyond the availability of the data itself, significant improvements are visible over the time since the introduction of the system. For example, fewer children are now needing surgery, and the system allows for earlier identification of contraindicators and abnormalities. These positive results are not necessarily directly influenced by the database, but the active monitoring processes are positively associated with the clinical improvements.

"The other thing we've got which is really helpful is a lot of data out of Sweden, like I said who have been doing this for over a decade, are really showing really substantial benefits to the kids. So these kids who are assessed regularly and early are having far less orthopaedic surgery, for example. So when Sweden looked at their data pre-CPUP, they will have that about 40 percent of kids needed orthopaedic surgery during their lifetime, now it's 15. Ten percent of kids have got a dislocated hip during their lifetime, only one percent now they've got proper surveillance."

"It's not randomised trials, it's registry data, pre-registry, post-registry so there are obviously confounders and so forth, but it does seem that good active monitoring of these children influences their outcomes."

"So one of the examples given in CPUP was a lot of these kids, for example, when they get older start getting tight knees. So they get flexion contractures, and when your knees start getting tight it can really affect your ability to walk. So historically what would happen, and this is a bit of a generalisation, but you would get picked up eventually by your GP who then would refer you to your physio, who would then refer you to the surgeon. There's delays at every step of the way and so your contracture can get quite severe before it gets treated and then it's harder to treat, whereas one of the possible advantages of CPIP is that that's picked up early and treated early. So as part of the management system, there's so-called traffic light system, so when you're entering in the data, if you enter knee flexion contracture zero, which is good, it will come up green. If you enter ten, it will come up red, and so it flags up to the therapist that there's an abnormality. They know that anyway but it just sort of highlights it, and also it

will highlight trends so then you can see six months ago it was five, now it's 15, and so those ones then stimulate "All right, this kid needs some attention, needs something doing about knee flexion contracture". So again, I'm not saying this is a definite benefit but what we think is that that is positively affecting their outcome by getting targeted early treatment before they get to a stage where it's too late, or not too late, but harder to do something about it. […] Again, dislocated hips is another one so a lot of children, particularly with the higher level CP, the more effective ones, their hips start to sublux and dislocate, and it's clinically silent in the early stages so you can't pick it up clinically, not easily anyway, and so they need X-rays. So as part of the CPIP system, not only is there physical assessment but also an X-ray assessment, and depending on your level of CP you get X-rays at different intervals. They get assessed by an orthopaedic surgeon and that gets put on the database as well, so again, you're picking that up earlier, and there is definite evidence that early detection of hip subluxation allows much better and easier treatment than picking up a hip that's completely dislocated and having to salvage it later on in life."

3.3.6 Stakeholder engagement

The team behind the CPIP have been working hard to ensure that the database is marketed as a clinical tool which will aid practice rather than work against it. As part of this, it has been important to recognize that clinical decision making is not compromised by using the tool, and to reassure the end users that clinicians have been heavily involved in its development. To do this, the database has been described as a patient management system rather than a research tool, which has improved its reception, particularly with clinicians who themselves have a research interest.

"We're trying to sell it as a patient management system with a database stuck on the back, and we're not trying to take away the professionalism of the people involved. […] We're basically saying, "This is a tool, it allows us to collect good data and it allows you to use that data for your own benefit to develop your practice", and then with the added benefits that we can get pooled data nationwide. [...] It's clinician developed as well, which helps, so all of the measures… they weren't developed by someone who's never treated cerebral palsy, they're developed by therapists and by surgeons to be used by therapists and surgeons."

"If I go to another Trust and say, "I've got this great database", that raises hackles, but if you're talking about patient pathway or patient management system, people are a bit more warm to that, particularly in this day of having integrated pathways and equity of access to healthcare, and trying to make things a bit more consistent across the region and across the country."

"Some of the senior paediatricians are really keen on it, particularly the ones who have got a research interest because they can see the value in having that data available, but you get some that [...] don't want to suddenly have to use yet another electronic system that they have to go to a website, log in

with a new username, a new password and put in new data. No-one really wants to do that."

In addition to stakeholder engagement prior to using the database, processes are in place to maintain this involvement. Once any organization begins using the CPIP, they are invited to send representatives to regular network meetings. This is an opportunity to feed back any problems or ideas, and to ensure that the views of the users as well as the Trusts are captured.

"[It's] called the National Network, so that's hosted by the APCP in London, and the idea is that as soon as CPIP is authorised or funded or whatever in your region, you send someone to National Networks and National Networks will always have at least one person, ideally two, but at least one person from every single region that's live, and that's obviously growing as more and more regions come online. […] Partly just so we can keep track, and also then we might think "Oh that's a good idea, let's have that, implement that nationally"."

Alongside the network, records are being kept to measure and monitor the uptake of the CPIP. This has included ensuring that there are contact details for users, and creating a map of how many Trusts are signed up to the system. Not only is this helpful in terms of visualizing the reach of the system, but it is also a valuable tool in making a case for new sign-ups.

"We've collected email addresses from the Start Working addresses which the Scots didn't, so they said that's a really good idea, so they've gone back and added that to their system."

"So more and more Trusts go online, we've got this map of red, amber and green, so red have got no CPIP, amber want to do it but haven't got funding yet and green are live. Then as more and more places go green, we can then use that as an example to go to other Trusts and say, "This is how many people are using it across England." […] That can be quite helpful, so we're kind of hoping that as more and more people do start using it, it will become too big to fail."

Keeping up to date records of users has also allowed for training opportunities and quality control. As part of the roll out, users are trained by registered CPIP trainers and may only access the system once this training is complete.

"One of the things we did in England was we wanted to keep track of everyone who was a current CPIP user, so the idea is that when you're trained by a registered trainer then you get a CPIP number and that becomes your training number and then that gives you access to the system."

3.3.7 Using the CPIP data

The data stored on the CPIP is available for researchers and clinicians, for uses which are deemed to be appropriate. The intention for the future of the database is to pool all of the data and create a more connected system for partnership working and data sharing. All applications

are assessed on a case-by-case basis to determine whether the data can be used, and the appropriate approvals (e.g. ethics) must be in place as with any other research undertaking.

> *"At the moment it's all kept region by region, Trust by Trust, but it's all within the same system, and the main idea of us all using that same system was that in the future we could pool data."*

> *"What tends to happen is if anyone wants to use the data, it's very brief, it's a small form, a little paragraph plus some demographic data just saying, "What do you want the data for and what data do you want?" Then that group then will decide whether or not they need more information or, no, they can't do that or whether they can do that. If it's audit you have to have your own Trust's audit lead approval in the same way you would if you were running any type of audit. If it's research, you'd have to have ethical approval in the same way any other research project would have to, and then if that's all approved then the data can be made available."*

Currently, the types of data available are somewhat limited sue to the infancy of the database. However, this has been intentional so that the system can be introduced gradually and built up over time. Initially, physiotherapy and orthopaedic surgery datasets form the bulk of the available data, as these were seen to be the specialisms which made sense for the first phase. Basing the work on the Swedish model, it is anticipated that using the limited data to identify and illustrate the benefits of the new system will lead to evidence-based introduction of a broader range of clinicians. IP008 again referenced the idea that should the database become 'too big to fail', and become the standard tool for CP care, then mass adoption will be easier and will happen more quickly. This has also had an effect on funding in Sweden, as the system has demonstrated that it can contribute to a net positive outcome for children with CP and therefore the need to campaign for government subsidy has effectively been removed.

> *"At the moment it's really physiotherapy and orthopaedic surgery that are within CPIP, and there's boxes for them but we want things like occupational therapy, speech and language, paediatrics. The kind of vision is that every subspecialty that looks after kids with CP will have their own little box to click and they will have their minimum data set that they want to collect for their particular thing. [...] Gunnar Haglund tells quite good stories about the Swedish model and he said he started a bit like us, started small and gradually build up, and then it just got too big to fail because there was parent pressure, carer pressure, patient pressure, clinician pressure and so forth. Everyone thought this was a good thing to do and so now it's been adopted as a national priority, or something, so effectively they don't have to make the financial argument anymore because it's been shown to be so effective that the government just paid for it in the same way that they'd pay for chemotherapy or anything else."*

3.3.8 Future funding

The CPIP is currently centrally funded for a limited time, so thoughts have turned to future funding. Ideally, the system would become best practice and therefore NHS funded in the same way as other CQUIN protocols, but alternatives have been considered.

> *"It's almost like a best practice-type thing and that also, Trusts really take notice because if you say CPIP isn't CQUIN, and then they say, "Right, okay, this is something we have to do," because if you don't do it then potentially you can be in all sorts of trouble. So smoking cessation is part of a CQUIN and so all Trusts now have to have smoking cessation as part of their… everything, health improvement and so forth."*

Industry funding from commercial partners is a potentially viable option, however there would be important considerations in terms of ensuring impartiality and an appropriate level of separation from the commercial partner.

> *"There are other sources of funding we thought about, although we haven't tried it yet, so industry. So take Botox and we use quite a lot of Botox in cerebral palsy to relax the tight muscles, and Allergan, who make Botox, are a wealthy company and we have thought about going to ask them to fund it and they can get their logo on the website and so forth, but so far we've kind of resisted that yet."*

> *"There are other models, so talked about industry funding, so I think for the national joint registry, as you may know, every patient in the UK that has a hip or knee replacement, their joint gets entered onto a registry collecting various data about it, and that's industry funding so there's a small amount of the money that the cost of a hip replacement, for example, is part of the registry, so we've got other ways of funding it."*

Because the database is primarily a clinical tool, the data will be securely kept even if the funding is not immediately replaced, meaning that in principle it is possible to regenerate the research side of the system at a later stage should there be a gap in funding.

> *"The data is secured under various NHS regulations. I think it has to be kept for kids for seven years after their 18th birthday. […] So all of that will be kept. It's just the website won't be accessible anymore."*

3.4 Interview - Case Study of the Secure Anonymised Information Linkage (SAIL) database

In order to better understand how a large databank may or may not be helpful to the OA community, a representative was interviewed from the Secure Anonymised Data Linkage (SAIL) databank. Attempts were made to interview representatives from other databanks however these were unsuccessful.

Questions put to the SAIL representative focused mostly on how the system works both from a researcher perspective and from a data management and governance perspective. The researchers were interested in knowing the process of using the system, and what considerations may be needed should an OA researcher wish to use it.

3.4.1 SAIL - overview

Initially, the representative was asked to give an overview of the process in brief. SAIL is a secure environment in which several large datasets, mostly health but more recently including non-health data, are co-located. Health data from around 20 years has been collated, and more recently there are now education and census datasets. SAIL is made up primarily of data from Wales, but wider UK datasets have now been incorporated. The secure environment created for SAIL is now also used within other databanks within a group of orgnisations including the UK Biobank.

> *"We've got nearly 20 years of good quality health data."*

> *"We were able to bring in England, Ireland and Scotland in to look at, to support questions there, we were looking at mental health and cystic fibrosis, so we've done that. The secure framework environment we've actually developed, initially it was for SAIL but my colleague has developed it so it's a product in its own right now, […] so SAIL is just one of those of 14, so you've got the dementia platforms, the UK Bio Bank is using it, Healthwise Wales and there is many more that I can't remember them all."*

The datasets are anonymised and encrypted so that they may be used for research purposes, and this process is handled by a third party company to ensure that no identifiable information is carried over once the datasets reach SAIL. As part of this process, NHS identifier numbers and other such unique person identifiers are replaced with a proprietary ID for each person which is attached to all records from the individual. Therefore, it is possible to track patients through the system at various points of contact (e.g. GP and hospital visits), without being able to identify them.

> *"SAIL is a secure environment where we co-locate different data sets, and so we initially started off looking at the data sets that relate to health, so your primary care, GP appointments and secondary care hospital appointments, outpatients and also the registry data of birth and deaths and different sets like that. […] Over the years we have accumulated many data sets that I have stored within one central database and the part of bringing them into SAIL is that we can use them to link to full research."*

> *"We have built that up, we started off with health but we have got more and more different types of data in now. So we have got education data, we've*

got screening data, we've got more administrative data, we are in the process of getting census data."

"We don't ever receive any of the demographic information about the individuals, but we have a trusted third party that will remove like personal identifiers into an unique ID that regardless of where you are on which data set, if you come through that process you will consistently be given the same ID. […] So it's not a meaningful figure that anybody can look up and be tied into who that person is. […] Whenever a data set comes in we send them the demographic to which they then try and match them to that list and try and identify what their NHS number is and then that's the process of looking for their NHS number. Then once you find it, they encrypt that, it comes to us, we then double encrypt it and then when we issue it to a researcher its then the third encryption there."

Where non-health data are held about an individual, there are algorithmic methods available to continue to map data to the person.

"It works okay when the data isn't coming from health, it doesn't have the NHS numbers so we look, so we've got the education data coming in so that won't have the NHS number in the source file but we've got probabilistic and deterministic methods to specifically see who that person is."

Besides the anonymisation, the data are replicated as faithfully as possible to the initial format, as it was decided that this would be the most appropriate way to maintain the integrity of the data. Attempts were previously made to harmonise data post-hoc, but these were untenable as a long term solution. This was due to a number of reasons included increased workload for the SAIL team, as well as the data owners feeling uncomfortable with the data being changed in this way. Instead, the datasets are kept in their original format and researchers using them can prepare the data prior to using it should this be required.

"What we try and do is we try and do a faithful representation of the original data but in an anonymised form."

"Initially when we first started out with SAIL we thought oh we'd try and harmonise the field names so that people knew that if they looked for a date it would be like this format, but then we realised that that was causing us an additional burden or responsibility because then the data providers, because some data providers do prefer that though they actually we want to carry on and work with you once the data is in there so that we can actually link our data sets with other data sets and they were getting a bit confused 'oh this is our data but you have renamed it and rechanged' and they didn't feel comfortable with that. […] So the data sets that are present in the form that they have been provided to us."

3.4.2 Accessing SAIL as a researcher

The SAIL database is open to any researcher using the data for public good. Though the datasets are available for this broadly defined use, there are checks and balances in place to ensure the integrity of the research and to guide researchers as to the most appropriate use of the data. The process begins with an initial scoping meeting with the SAIL team, in which the project proposal is discussed. The SAIL team then determine which datasets are available and appropriate for the project. These meetings may happen either before or after project funding has been secured. During these meetings, the SAIL team may also request additional information from the researchers such as specific clinical classification codes to allow a more thorough search of the data to take place.

> *"People will come to us and we will do a scoping and discussion to say what their research question is and how feasible it would be done to be helped with by the data that's in SAIL. Sometimes we haven't got all the data they need, and they may need to provide data to come through our matching anonymisation process to augment the data so that they can pull it together."*

> *"Each of those discussions are quite unique and tailored to the researchers because some research projects will have a team of the very experienced people that have used routine data before and they just want to be given access to the data whereas others will want the expertise of the data manipulation."*

Following the initial scoping meeting with SAIL, researchers must then submit a full proposal which is further reviewed by a panel who will determine whether or not access can be granted.

To access the data once permission has been granted, researchers are provided with a secure login to a portal where they will find all of the datasets housed within SAIL. Within the portal they are free to work on the raw data and prepare it as needed for their project. The entire process takes place within the portal and no data are ever removed from the system. In order to obtain results in a format which can be removed from the system, the researcher must once again submit to assessment by a panel from SAIL, who check that the data usage matches the proposal and the methodology is sound.

> *"Once that's been put through and people are satisfied that the research is valid use of the data, then there is a separate step of higher access to the data, so what we have developed […] is a secure remote environment that researchers can actually log in from their own desktop […] where they then can access the data on a database if they are confident to do SQL or within statistical packages like R or Scatter or SPSS when they can then start working with the data. They work with the data at a very raw level, granular level and they have their own project area that they can do their work and pull information together. Once they've got results from working with the data then they have to, to get them out that secure environment, they have to formally request them out and this when myself or someone from my team will scrutinise what the outputs are, so then they'll say, they'll check against the IDRP application that they made to say that what they are using is what they have said they are going to use it for, and then there's that process of*

review and then they can go ahead and publish that work once they are ready to."

The current system has worked well thus far, and allows an appropriate level of rigour in order to protect the data and prevent misuse. However, efforts are now being made to streamline the process by allowing researchers to share and access algorithms and other data preparation methods in order to reduce time spent on this part of the process. There are a number of common data queries which are repeated, and the SAIL team aim to be able to assist researchers and save time. This will be achieved via a combination of SAIL providing the algorithms, and encouraging other researchers to share their work.

"We are trying to identify areas that we can try and get data more research ready so try and find other common pieces of work that people are doing that gets data into normal, reproduceable research ready data so the people aren't having to repeat common data preparation tasks."

"We are trying to encourage more and more people to work with us to try and share their knowledge that they have learnt from using the data."

In addition to sharing data preparation methods and processes, researchers are encourage to share their diagnostic codes and the links with specific conditions. Sharing codes is hoped to be something which will aid data searching and save time for other researchers.

"My colleague here has developed a concept library which is a system which you can store a standard set of codes that you would say, oh this code is what I am going to use to identify depression or these are the codes that I am going to use to identify antibiotics prescriptions and you can store that list of codes in this system and then give that then to run within your code to select the records."

"When you publish your research you can say, well actually I am happy to then share my code sets and people can use them."

Where these codes are not shared or replicable from a previous project, researchers must instead discuss this with the scoping team and try to determine how to identify patients with specific conditions from the data. This can be laborious with complex conditions and involves mapping out the care pathways of patients and pulling data from services they may interact with. This can be challenging for the SAIL team, who are not clinicians and therefore do not have specialist knowledge which may aid these searches.

"We would have to sit down and talk with the person about what service would that person interact with, where would see that, that interaction there?"

"So if we say, okay it sounds like they have quite a lot of their interactions with their GP, can you go back and try and find out what read codes that would relate to and we would ask them to look up either the [clinical codes] for certain things or published papers that have already detailed that, because we are not clinicians in ourselves and so we wouldn't be able to say 'ah we know what medications they are being prescribed'."

3.4.3 Who can access the SAIL data?

The SAIL databank is strictly available only for research which is considered to be for public good, and this is determined throughout the various approval points. Although the team do work with commercial companies, there are rules governing this which prevent the use of the data for profit-driven purposes. Commercial entities must also have an academic partner in order to access the data.

> *"One of the things is making sure that the research is for the public good and it's not something that is going to, its not something that's coming from a private company because they want to pitch a certain drug or, so that sort of commercial motivated profit thing, it's more looking for research and improving our knowledge base of healthcare and other aspects of the data that we've got there."*

> *"It would have to be framed in an appropriate, non-biased research question. And it would have to be led by an academic or clinician. So we do work with pharma companies, but we need to make sure that it's pitched in that we are doing it to understand more, rather than to spin."*

3.4.4 Funding

Projects using SAIL must be fully funded, and must cover the costs incurred by the SAIL team. Though there is no direct cost for the data, the data preparation team are funded by a cost recovery model. Cost recovery is discussed in the initial scoping meeting and is determined on a case-by-case basis depending on how much work is involved. Updated datasets can be accessed at agreed upon time points during long-term projects, but this incurs additional costs to the researcher.

> *"Most of my team is funded by research projects, so the projects that we support that we have to do the data preparation or the actual full project."*

> *"It's quite a discussion on what the research you want to do, is the data there, and then we talk through, well these are the steps we think you need to do and these are the associated costs with that and then you review the document and see if they agree and then that tends to be when people then go off and try and seek funding to support that and once you've worked through those bits. So we try and guide you as well so its like some people might people think, 'oh we'll use this data set to do this' and alright then well, actually you'd be better off using this different data sets. So we try and help be supportive and try and recommend, because we want people's projects to be successful, so we are trying to make sure people are using the best and most out of the system."*

> *"There is a charge associated with work with the data so we don't charge for the data itself, but it's trying to cost recover using the infrastructure technology and everything that's involved in getting the data ready there for you to use. It's variable on the level of support you have, it's variable on the level of data refreshers you have, so you get provisioned a version of data but then if you want, if your project is five years long and you want refreshers*

every three months, then there's going to be a charge associated with that. […] Its all negotiated during the scoping process."

3.4.5 Data governance

The datasets co-located within SAIL are governed by strict consent and

> *"We've got Healthwise Wales project in Wales where it's encouraging people to be more involved in the research communities, so people volunteer to be contacted to do research questions and things like that, and as part of that consent they've agreed that their data we hold, is held about them is shared with that project. […] Because they've got given informed consent and any data that they collect is then joined with the data that we have within SAILS so its augmented together."*

> *"If you want to bring the data in then we need to make sure that the person, the data owner, the person who is responsible for looking after that data set, they've given permission for the data flow to come in. So that it can come into SAIL and be used. If you wanted to do like we've got some studies where they are set up UK wide and they are collecting their data and they want the data from the four nations to come into them, to their central repository then that's when that the ethics and consent needs to be set up at the right the beginning to ensure that the patients are fully, or participants are fully informed that that would happen."*

3.5 Questionnaire

A total of ten complete questionnaire responses were collected, which was substantially below the target number of 140. Though the questionnaire was circulated to the OAtech Network mailing list, there were a number of possible explanations as to why the target sample size was not met. The invitation was sent in the late summer months, which for many academics are a holiday period. Additionally, staffing changes at the network meant that the mailing list was not being as actively monitored therefore the questionnaire link was not re-circulated to increase responses.

Due to these limitations, it was not possible to perform any detailed analysis on the results. However, the responses are presented below in order to provide some additional information to the interviews.

3.5.1 Participants

A total of ten participants completed the survey (in this instance, 'completed' refers to responses with a 'yes' response to the consent statements, and completion of all questions). Of those participants, two were clinicians actively involved in OA research and the remaining two were academic researchers.

Participants were asked to select from a number of options to describe their area of research activity, and choose all which apply. Table 2 shows the frequencies of these areas of interest within the sample.

Area of interest	Frequency
Biomechanics / Mechanobiology	5
Medical devices	4
Implants	3
Orthopaedics	3
Biomarkers	2
Imaging	3
Radiology	2
Activity modelling	2
Machine learning/AI	1
Mathematical modelling	2
Radiology	2
Physiotherapy	2
Clinical/biomedical engineering	3
Proteomics	1
Genetics / Genomics	1

Table 2: Areas of interest as chosen by questionnaire participants

One participant stated that they do not work with clinical patient data, three stated that they work with clinical patient data but do not collect it themselves, and six worked directly with patients.

Participants were also asked to choose from a list of data types which they routinely collect and were again allowed to choose all which apply. Table 3 shows the frequencies of responses.

Data type	Frequency
Body sensing – activity data (e.g. step count)	3
Body sensing – galvanic skin response/electrodermal activity	1
Imaging – X Ray	6
Imaging - CT	6
Imaging – MRI/fMRI	5
Imaging - Ultrasound	1
Biomechanics – Kinematic data	4
Biomechanics – Kinetic data	5
Biomechanics – EMG data	3

Biomachanics – Other	1
PROMS – Oxford Knee Score	4
PROMS – Oxford Hip Score	1
PROMS - KOOS	5
PROMS - WOMAC	4
PROMS – Visual Analogue Scale	5
PROMS – Forgotten Joint Score	3
PROMS - PROMIS	1
PROMS - HAQ	2
PROMS – Knee Society Score	1
PROMS - Other	4
Cells/Bio	3
Other (not listed)	1

Table 3: Types of data routinely collected by survey respondents

3.5.2 Data collection

Participants indicated that the most commonly collected data collection tools within this sample were imaging (X-Ray and CT being the most common, closely followed by MRI/fMRI). Of the biomechanics data types, kinematic was the most common and of the standardised tests the KOOS was most commonly used. Other data types not listed but specified by participants included the EQ5D (frequency: 2), Back pain outcome measures (frequency: 1), the SF12 (frequency: 1) and one participant who gave the following list; Lysholm, ICOAP, Histology, ICRS cartilage repair score, positive and negative effect schedule, brief pain inventory and RNA sequence data.

Participants were also asked about the format of their data, and what they would typically use. Table 4 gives the frequencies of responses. As shown in the table, within this sample the most common data format used was Exce/CSV/Text. Other formats given by participants were binary (frequency: 1) and R (frequency: 1).

Data format	**Frequency**
DICOM	5
Excel/CSV/text	9
C3D	2
Matlab	5
SPSS	2
Other	2

Table 4: Format of data routinely collected by survey respondents

There was a wide range of participant numbers within shareable datasets as given by respondents - estimates ranged from 10 to 500.

When asked whether there are a set of minimum measurements typically collected during their work, five participants answered 'yes' and gave further details. The minimum requirements in participants' own words were; "Tests repeated in triplicate for sensing activities", "Seven different MRI protocols including 1H and 23Na", "Cartiage thickness and bone shape", "Spinal movement" and "Age, sex, height, KL grade, weight".

3.5.3 Data sharing

Statements:	**Frequency of participant agreement levels**				
Thinking about the opportunity combining datasets in your research area might provide for data science and machine learning methods, please rate your level of agreement with the following:	**Strongly Agree**	**Somewhat Agree**	**Neutral**	**Somewhat Disagree**	**Strongly Disagree**
My own research would benefit from having access to large datasets	6	4	0	0	0
My research area in general would benefit from having access to large datasets	9	1	0	0	0
I am willing to spend the time necessary to make my data suitable for sharing	4	5	1	0	0
It would be relatively simple to share my datasets	1	2	5	2	0
My datasets are too complex to make accessible	0	0	5	3	2
I am concerned sharing datasets will breach ethical requirements in place	1	4	2	3	0
I am concerned sharing datasets will breach	1	4	2	3	0

GDPR requirements in place					
My data would benefit from being combined with other datasets of the same measures (e.g. combining biomechanics data with other biomechanics data from different labs)	4	5	1	0	0
My data would benefit from being combined with other datasets of different measures (e.g. biomechanics combined with genomics data)*	5	4	1	0	0
Thinking again about the opportunity combining datasets might provide for data science and machine learning methods, please state agreement to which of the following areas could benefit:	**Strongly Agree**	**Somewhat Agree**	**Neutral**	**Somewhat Disagree**	**Strongly Disagree**
Better prediction of OA risk	3	6	1	0	0
Earlier diagnosis of OA	3	6	1	0	0
Stratification of OA (defining/classification of OA types)	4	6	0	0	0
Clinical Decision Support - assisted diagnosis	2	5	2	1	0
Clinical Decision Support - assisted intervention/care	3	3	4	0	0

Table 5: Agreement levels with statements related to data sharing

* Participants were asked to give examples of the types of data which they saw as being beneficial for combination with their own. Responses were; Biomechanics, X-Ray and CT, Genomics, MR 3D for volume measurements, Kinetic data during gait, High resolution imaging collected during total joint replacement and other gene express work.

The responses in Table 5 suggest that in this sample, participants felt that larger datasets would be beneficial to their general areas of research, and only slightly less agreed that their specific area of research would benefit. The responses suggest that there are some concerns that sharing data would present ethical and GDPR challenges but these were not strongly held. Generally, participants agreed that combining datasets could help with a range of areas of OA, such as stratification.

Only one of the ten respondents had actually submitted data to a larger databank (APPROACH EU consortium). However, six had accessed data from larger databanks – the Osteoarthritis Initiative, National Joint Registry and UK Biobank were cited as examples. The reasons given for doing so included gathering comparative data, investigating trends and obtaining imaging data.

3.5.4 Ethics and Governance

Just two of the ten respondents stated that their participants are routinely asked to give consent for data to be shared outside of their organisation, with three answering 'no' to this question and five stating that this is sometimes the case but not always. Six respondents stated that their work is subject to specific internal ethical or governance processes, with four participants also routinely completing external processes (IRAS, REC and GDPR).

3.5.5 Barriers to data sharing

Finally, participants were asked to explain in their own words what they felt to be the main barriers to data sharing within OA research. There were insufficient responses to conduct a thematic analysis, however a small number of common themes were evident within the comments. Ethical issues were seen to be a concern, as well as researchers sometimes being protective of their data which may cause difficulties with collaboration. Consistency and comparability of data were also mentioned, as well as challenges related to the time and resources required to facilitate data sharing.

> *"Ethics concerns; willingness of researchers to collaborate"*
>
> *"Ethics – if patients have not consented we cannot do it"*
>
> *"Lots of regulations, that aren't immediately clear, so sharing data takes a lot of time and effort to research the correct way to do it, if it is actually allowed. There needs to be clear, user friendly documentation on data sharing and modern technology to improve the correct sharing of data."*

One participant felt however, that with some dedication and effort, a solution is possible. This participant felt that a coordinated effort to set up a multi-centre collaboration of some kind would be beneficial to OA research and could improve the quality of data available.

> *"A co-ordinated multi-centre effort to establish a network of OA researchers (along the lines of multi-centre trials) is achievable (with some stats work and a lot of co-ordination) and would allow a step-change in to quality and scope of UK OA research".*

4 References

Ader, D. (2007). Developing the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(Suppl 1), pp.S1-S2.

Ashinsky, B., Bouhrara, M., Coletta, C., Lehallier, B., Urish, K., Lin, P., Goldberg, I. and Spencer, R. (2017). Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. *Journal of Orthopaedic Research*, 35(10), pp.2243-2250.

Azzi, E., Thienpont, E., Avaux, M., Houssiau, F. and Durez, P. (2014). AB1121 The Forgotten Joint Score, A New Questionnaire to Evaluate Patient's Perception of Total Knee and Hip Arthroplasty in Patients with Established Rheumatoid Arthritis. *Annals of the Rheumatic Diseases*, 73(Suppl 2), pp.1173.1-1173.

Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988; **15**: 1833–40.

Beukenhorst, A., Parkes, M., Cook, L., Barnard, R., van der Veer, S., Little, M., Howells, K., Sanders, C., Sergeant, J., O'Neill, T., McBeth, J. and Dixon, W. (2019). Collecting Symptoms and Sensor Data With Consumer Smartwatches (the Knee OsteoArthritis, Linking Activity and Pain Study): Protocol for a Longitudinal, Observational Feasibility Study. *JMIR Research Protocols*, 8(1), p.e10238.

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. and Rong, X. (2015). Data Mining for the Internet of Things: Literature Review and Challenges. *International Journal of Distributed Sensor Networks*, 11(8), p.431047.

Dawson, J., Fitzpatrick, R., Murray, D. and Carr, A. (1998). Questionnaire on the perceptions of patients about total knee replacement. *The Journal of Bone and Joint Surgery*, 80(1), pp.63-69.

Ding, C., Zhang, Y. and Hunter, D. (2013). Use of imaging techniques to predict progression in osteoarthritis. *Current Opinion in Rheumatology*, 25(1), pp.127-135.

Driban, J., Sitler, M., Barbe, M. and Balasubramanian, E. (2009). Is osteoarthritis a heterogeneous disease that can be stratified into subsets?. *Clinical Rheumatology*, 29(2), pp.123-131.

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J., Hastie, T. and Delp, S. (2018). Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of Biomechanics*, 81, pp.1-11.

Harvey, N., Matthews, P., Collins, R. and Cooper, C. (2013). Osteoporosis epidemiology in UK Biobank: a unique opportunity for international researchers. *Osteoporosis International*, 24(12), pp.2903-2905.

Hunter, D. (2009). Risk stratification for knee osteoarthritis progression: a narrative review. *Osteoarthritis and Cartilage*, 17(11), pp.1402-1407.

Kingsbury, S., Corp, N., Watt, F., Felson, D., O'Neill, T., Holt, C., Jones, R., Conaghan, P. and Arden, N. (2016). Harmonising data collection from osteoarthritis studies to enable

stratification: recommendations on core data collection from an Arthritis Research UK clinical studies group. *Rheumatology*, 55(8), pp.1394-1402.

Kittelson, A., Stevens-Lapsley, J. and Schmiege, S. (2016). Determination of Pain Phenotypes in Knee Osteoarthritis: A Latent Class Analysis Using Data From the Osteoarthritis Initiative. *Arthritis Care & Research*, 68(5), pp.612-620.

Kraus, V., Blanco, F., Englund, M., Karsdal, M. and Lohmander, L. (2015). Call for standardized definitions of osteoarthritis and risk stratification for clinical trials and clinical use. *Osteoarthritis and Cartilage*, 23(8), pp.1233-1241.

Liebl, H., Joseph, G., Nevitt, M., Singh, N., Heilmeier, U., Subburaj, K., Jungmann, P., McCulloch, C., Lynch, J., Lane, N. and Link, T. (2014). Early T2 changes predict onset of radiographic knee osteoarthritis: data from the osteoarthritis initiative. *Annals of the Rheumatic Diseases*, 74(7), pp.1353-1359.

Maska, L., Anderson, J. and Michaud, K. (2011). Measures of functional status and quality of life in rheumatoid arthritis: Health Assessment Questionnaire Disability Index (HAQ), Modified Health Assessment Questionnaire (MHAQ), Multidimensional Health Assessment Questionnaire (MDHAQ), Health Assessment. *Arthritis Care & Research*, 63(S11), pp.S4-S13.

Nuesch, E., Trelle, S., Reichenbach, S., Rutjes, A., Tschannen, B., Altman, D., Egger, M. and Juni, P. (2010). Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ*, 341(jul16 1), pp.c3515-c3515.

Peat, G., Riley, R., Croft, P., Morley, K., Kyzas, P., Moons, K., Perel, P., Steyerberg, E., Schroter, S., Altman, D. and Hemingway, H. (2014). Improving the Transparency of Prognosis Research: The Role of Reporting, Data Sharing, Registration, and Protocols. *PLoS Medicine*, 11(7), p.e1001671.

Palazzo, C., Nguyen, C., Lefevre-Colau, M., Rannou, F. and Poiraudeau, S. (2016). Risk factors and burden of osteoarthritis. Annals of Physical and Rehabilitation Medicine, 59(3), pp.134-138.

Petersen, I., Welch, C., Nazareth, I., Walters, K., Marston, L., Morris, R., Carpenter, J., Morris, T. and Pham, T. (2019). Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clinical Epidemiology*, 11, pp.157-167.

Ren, G. and Krawetz, R. (2015). Applying computation biology and “big data” to develop multiplex diagnostics for complex chronic diseases such as osteoarthritis. *Biomarkers*, 20(8), pp.533-539.

Roos, E., Roos, H., Lohmander, L., Ekdahl, C. and Beynnon, B. (1998). Knee Injury and Osteoarthritis Outcome Score (KOOS)—Development of a Self-Administered Outcome Measure. *Journal of Orthopaedic & Sports Physical Therapy*, 28(2), pp.88-96.

Schaap, L., Peeters, G., Dennison, E., Zambon, S., Nikolaus, T., Sanchez-Martinez, M., Musacchio, E., van Schoor, N. and Deeg, D. (2011). European Project on Osteoarthritis (EPOSA): methodological challenges in harmonization of existing data from five European population-based cohorts on aging. *BMC Musculoskeletal Disorders*, 12(1).

Swan, A., Stekel, D., Hodgman, C., Allaway, D., Alqahtani, M., Mobasheri, A. and Bacardit, J. (2015). A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC Genomics*, 16(Suppl 1), p.S2.

Taichman, D., Backus, J., Baethge, C., Bauchner, H., de Leeuw, P., Drazen, J., Fletcher, J., Frizelle, F., Groves, T., Haileamlak, A., James, A., Laine, C., Peiperl, L., Pinborg, A., Sahni, P. and Wu, S. (2016). Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors. *The Lancet*, 387(10016), pp.e9-e11.

Tegner, Y. and Lysholm, J. (1985). Rating Systems in the Evaluation of Knee Ligament Injuries. *Clinical Orthopaedics and Related Research*, &NA;(198), pp.42-49.

Waarsing, J., Bierma-Zeinstra, S. and Weinans, H. (2015). Distinct subtypes of knee osteoarthritis: data from the Osteoarthritis Initiative. *Rheumatology*, 54(9), pp.1650-1658.

Wroblewski, B. (1996). Questionnaire on the perceptions of patients about total hip replacement. *The Journal of Bone and Joint Surgery. British volume*, 78-B(5), pp.856-856.

Yan, P., Suzuki, K., Wang, F. and Shen, D. (2013). Machine learning in medical imaging. *Machine Vision and Applications*, 24(7), pp.1327-1329.

Zengini, E., Hatzikotoulas, K., Tachmazidou, I., Steinberg, J., Hartwig, F., Southam, L., Hackinger, S., Boer, C., Styrkarsdottir, U., Gilly, A., Suveges, D., Killian, B., Ingvarsson, T., Jonsson, H., Babis, G., McCaskie, A., Uitterlinden, A., van Meurs, J., Thorsteinsdottir, U., Stefansson, K., Davey Smith, G., Wilkinson, J. and Zeggini, E. (2018). Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nature Genetics*, 50(4), pp.549-558.