Data Sharing Across Osteoarthritis Research Groups and Disciplines: Opportunities and Challenges

Jill Evans
University of Warwick
Paul Biggs
Cardiff University
Mark Elliott (m.t.elliott@warwick.ac.uk)
University of Warwick https://orcid.org/0000-0003-4000-0198

Research article

Keywords: Osteoarthritis, data sharing, data harmonisation

Posted Date: October 29th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-97815/v1

Version of Record: A version of this preprint was published at Osteoarthritis and Cartilage Open on January 1st, 2022. See the published version at https://doi.org/10.1016/j.ocarto.2022.100236.

Abstract

Background: Osteoarthritis is a heterogeneous condition characterised by a wide variety of factors and represents a worldwide healthcare challenge. There are multiple clinical and research specialisms involved in the diagnosis, prognosis and treatment of osteoarthritis, and there may be opportunities to share or pool data which are currently not being utilised. However, there are challenges to doing so which require carefully structured solutions and partnership working.

Methods: Interviews were conducted with nine experts from various fields within osteoarthritis research. A semi-structured approach was used, and thematic analysis applied to the results.

Results: Generally, osteoarthritis researchers were supportive of data sharing, provided it is done responsibly and without impacting data integrity. Benefits identified included increasing typically low-powered data, the potential for machine learning opportunities, and the potential for improved patient outcomes. However, a number of challenges were identified, in particular related to; data security, data harmonisation, storage costs, ethical considerations and governance.

Conclusions: There is clear support for increased data sharing and partnership working in osteoarthritis research. Further investigation will be required to navigate the complex issues identified; however, it is clear that collaborative opportunities should be better facilitated and there may be innovative ways to do this. It is also clear that nomenclature within different disciplines could be better streamlined, to improve existing opportunities to harmonise data.

Introduction

Osteoarthritis (OA) is a heterogeneous condition characterised by a wide variety of clinical factors and is a significant health challenge worldwide (1). Subsequently, OA research covers a broad range of disciplines, ranging from cell-based studies through to population level, epidemiological research. The resulting spectrum of datasets is heterogeneous but tends to remain siloed within disciplines and often within research groups. This could be detrimental to the application of modern data-driven analysis methods that allow the analysis of large datasets. Such approaches can be used to identify patterns that could assist with stratification into disease subsets which may then be targeted differentially and potentially provide improved scope for successful treatments, as well as predictive models that allow for earlier detection and interventions (2). Hence, the facilitation and promotion of sharing and combining datasets within the OA research community is likely to be advantageous to future research into this condition.

In addition to facilitating new analysis methods, data sharing is beneficial in reducing cost to researchers in terms of data collection, and hence also reduces the need for patients as participants (3). However, there is a notable lack of action in adopting this approach by researchers (4). There are many challenges that remain before the sharing of clinical study data is fully adopted, including ethical and privacy requirements along with the loss of control of who will use the data and for what purpose, outside of the

original study. At the extremes there is a fear that data sharing will result in “research parasites”; researchers who will purely use “other groups’ data for their own ends” (5).

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’. Additionally, there are multiple datasets being collected in siloes by researchers, which may have potential to be collated. 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.

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 heterogeneous 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).

Despite these challenges, the opportunities for the analysis of large datasets, generated through sharing, are clear. Recent years have seen major advances in imaging techniques and machine learning, improving upon traditional data collection and analysis methods in OA. Magnetic Resonance Imaging (MRI) has been demonstrated to provide a detailed examination of whole joints and also offers several benefits as a research tool (6) including expediting research participant screening and selection, consequently reducing timescales for clinical trials. 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 (7).

Machine learning is also increasingly of interest in health research due to advances in data generation, and specifically OA, 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 (8). Currently, there are no standardised guidelines or frameworks in terms of data collection taxonomies in OA, and thus sharing across these disciplines are limited. An example of the potential achievements from large, multi-variate datasets can be noted by the outputs of the Osteoarthritis Initiative project (OAI) (9), however. This large 10-year longitudinal study, collected data on nearly 5,000 participants with knee OA, collating imaging, activity and patient recorded outcome measures amongst other variables. Analyses on this dataset have generated over 400 publications to date (10), which is still one of the most comprehensive in the OA research field.

In this study, we investigate the opportunities and barriers to data sharing and the implementation of a specific OA research data repository. We utilised the expertise within the OATech + network (12) to

interview eight senior researchers across disciplines to explore the current thinking on this topic. We subsequently discuss the key themes that emerged from the interviews and make recommendations around how OA research data sharing may be implemented in the future.

Methods

Study design

A qualitative design was used for this study, using one-to-one interviews with active researchers who are focussed on osteoarthritis. The study was granted favourable review by the University of Warwick Biomedical & Scientific Research Ethics Committee (BSREC) and the Health Research Authority prior to any data collection taking place.

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. The 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, 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.

Recruitment

Interview participants were purposively sampled based on their professional experience and roles. Communications were circulated round the OAtech + Network asking for experts across different areas of research related to OA. Other academic and commercial representatives outside of the network were also invited to participate in a broader study, not reported here.

Analysis

Qualitative data from interviews and focus groups were 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.

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 were several areas of overlap.

Table 1
Research areas and expertise of each participant

Participant ID	Specialism(s)/Research area(s)
IP01	Physical function from a clinical perspective
IP02	Biomarkers, biomedical engineering, activity monitoring, gait analysis
IP03	Genomics, proteomics
IP04	Pathogenesis, bio markers, cell therapy
IP05	Rheumatology, imaging
IP06	Genomics
IP07	Rheumatology, epidemiology, lifestyle interventions
IP09	Biomarkers, population studies

Results

The following sections describe the findings from the interviews, based on the themes derived from the analysis.

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 test hypotheses in small samples was seen as a 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.

IP04: "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."

IP06: "If it does work it could be quite transformative. […] Because what tended to happen in osteoarthritis is people progress slowly, companies don't want to do clinical trials of three, four, five years."

IP09: "I think there's probably more around diagnosis than there is maybe around prognostic modelling."

One aspect of machine learning on which participants strongly agreed, was that collaboration with people with expertise in the area is essential since the field is relatively new and also extremely complex. Participants were agreed that specific knowledge is required 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. IP02 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.

IP02: "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."

IP07: "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."

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.

IP05: "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."

IP09: "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."

Size of datasets

Given the need for larger datasets to facilitate the application of machine learning approaches, participants were questioned on their current, typical data collection, in terms of size, format and minimum data requirements. There was a large amount of variance in the sizes of the datasets collected or used by participants, depending on the nature of the study. 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. However, it was felt that smaller sample sizes are an inevitability with this type of research. IP02, 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 to collect multiple measures from large cohorts.

IP02: "We've got a database of [...] I think it's about 200 people, some have imaging as well some don't. […] 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, […] 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. 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.

IP07: "For electronic health records, if it's primary care data, you can get thousands of participants. […] Secondary care data is less widely available for research. We're accessing our electronic health records from a single local hospital."

IP03: "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."

Minimum data collection requirements

Participants were asked whether there are any existing guidelines regarding 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.

IP02: "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."

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 that 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.

IP09: "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."

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 motion capture marker placement was also a consideration which may make comparing datasets difficult.

IP02: "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 values 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.

IP04: "The MRC have set up the Biobank, haven't they, which is a good exemplar of what can be done, […] You could have a common core that different centres could use, that might be a way to improve it."

Alongside discussing the idea of standardising data collection, it was also commented that even when using clinical data there are inconsistencies which make data pooling difficult. In particular, one participant felt that the nomenclature used across OA is poorly defined, and the standard International

Classification of Diseases (ICD-10) 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.

IP09: "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. […] 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. […] 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."

IP07 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.

IP07: "We're looking to try and structure the data collection in clinical care so that it's then also more useful for research."

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.

IP01: "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 heterogeneous data may be too time- and resource-consuming and may risk diluting or invalidating findings, particularly when resources such as the Osteoarthritis Initiative (OAI) (9) exist and provide large scale data collected in a robust manner.

IP05: "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?"

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.

IP06: "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."

IP09: "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?".

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.

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 research and as such is not yet governed by any best practice guidelines.

IP01: "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.

IP05: "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."

IP09: "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."

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. This 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. 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.

IP05: "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. […] 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."

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. It was also noted that since the introduction of the General Data Protection Regulation (GDPR) in the UK (13), researchers are held to more stringent ethical guidelines.

IP05: "When we set things up 10 years ago, [we] didn't think we [would] want to come back and dip into things again. […] [The] ability to re-contact people, it has to be in your consent forms. […] 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?"

IP05 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.

IP05: "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'.

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.

IP06: "We've noticed that OA patients are delighted that somebody is investigating their disease and wants to know something about it."

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.

IP01: "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."

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. IP02 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.

IP02: "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. "

It was however 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.

IP07: "There would be a data access application process; there would be a review of that application; […] we have data sharing agreements that people have to sign and sort of terms and conditions of use as well in terms of how to acknowledge the data source."

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) (9), the Clinical Practice Research Datalink (CPRD) (13), the Imperial Tissue Bank (14), REDCap (15), OpenClinica (16) and the UK Biobank (17). 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.

IP03: "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."

IP09: "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."

IP04: "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.

IP06: "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 […]. 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."

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.

IP03: "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."

IP01: "So, it's not a simple thing where anyone can just say, "I want this data and I am going to run these tests.""

IP07: "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. IP07 described taking a collaborative approach and sharing their data preparation programmes on open-source platforms to help other researchers do this more efficiently in future.

IP07: "There's a whole data preparation step that is complicated. […] 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."

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. IP03 in particular felt that when collecting tissue, researchers tend to be collaborative due to the difficulty of obtaining samples.

IP03: "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."

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. 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.

IP01: "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 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."

IP07: "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 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.

IP02: "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."

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.

IP09: "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."

Discussion

Large, integrated datasets are of significant benefit to data-driven analyses which can lead to advanced approaches such as precision medicine, where interventions are targeted at the specific characteristics of a patient's condition (19). This is particularly relevant for Osteoarthritis research (20) which covers a broad range of sub-disciplines but typically consist of datasets which suffer from small sample sizes. Here, we have investigated the opportunities and challenges for sharing and combining datasets within the OA research domain. We have used the UK-based OATech + Network to access experts across disciplines in order to gain a broad view of opinions.

The insight we have gathered from OA researchers has provided an overview of the current approaches to data sharing, data harmonisation and collaborative working within the field in the UK. On the whole, our results suggest that the ability to have access to datasets that facilitate the application of machine learning (ML) methods is likely to transform OA research through the development of new algorithms and pattern identification previously not possible due to time or resource constraints. The application of ML methods is being applied across numerous disciplines (21) related to OA research. Therefore, pooling of data within as well as across disciplines is likely to be advantageous the progress of data-driven research.

It is clear from the discussions, however, that there are numerous challenges to pooling and sharing datasets. This included data storage, whereby the strict governance, ethical and data protection requirements were highlighted. A recent study of digital health data governance in low- and middle-income countries suggested a four-domain framework for helping stakeholders achieve an appropriate level of data protection (22). Salient points raised include the avoidance of person-centric gatekeeping – instead using a committee-based approach for access management and long-term storage strategies and the need to implement a well-defined, documented data structure. This corroborates points raised by participants in our interviews, who had a similar viewpoint in the context of OA data sharing. There are also examples of this approach being successful in the UK in different areas, such as The National Joint Registry (23) and the Cerebral Palsy Integrated Pathway (24).

Participants in our study also noted that variance in nomenclature and medical coding can make searching and/or sharing existing clinical and research data challenging and may mean that comparable datasets are missed. Similarly, there are multiple clinical IT systems in place in the UK, and that even within these systems, there are inconsistencies in clinical classifications (25). OA is a condition with many routes to diagnosis and this can complicate the pattern of clinical coding – this, in turn, can make searching clinical data more difficult than other conditions. It may not be practical to fully standardise the way OA is coded in research and clinical care, but a potential opportunity is to create and maintain a training or learning structure for researchers. With a system in place, it may be possible to raise awareness among researchers of the various codes and search terms they can use to identify data and/or patients for trials, and potentially to increase datasets.

Even when a dedicated effort is made to harmonise 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-consuming and may still not yield robust results. We observed concerns from the interviewees about standardising data retroactively and how this might impact validity and reliability. Some level of data pooling was seen as possible where appropriate and where measures align, but where significant effort is required to anonymise or homogenise the data, this was not seen as useful. The European Project on Osteoarthritis (EPOSA) experienced such challenges when attempting to combine data from five multinational longitudinal studies (26). The EPOSA study found that the lack of agreement on data collection instruments and procedures between OA researchers was a key factor in th 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.

It is clear that there are some easily achievable steps which could increase future sharing and integration of datasets. This includes the use of standard wording on ethical applications and consent forms to ensure participants have the option to consent to their data being used for further research in the future. Not including such wording, instantly removes any opportunity for sharing and future use of that

particular dataset and therefore in the past it has been suggested that ethics committees should raise the issue in applications not containing this information (27).

Conclusion And Recommendations

In conclusion, there is consensus amongst the community that using data-driven approaches, such as machine learning, is an increasingly important method to be used in current and future OA research. To maximise the opportunities around these methods, changes need to be made from an individual researcher level through to the broader research community level.

There are several practical considerations for sharing OA research 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. 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 (28). 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 (29).

Another potential solution to some of the problems with data sharing and pooling in OA research is for a governed framework aimed at facilitating communication and collaboration between researchers and other specialist teams. For example, it was noted by interviewees that having access to computer scientists meant that they could call upon the required expertise to set up complex data sharing protocols and processes, or have a dedicated team member developing machine learning algorithms while maintaining research goals. This allows researchers to use their time efficiently and have appropriate input into data science efforts, but also to recognise where additional help from experts is needed to adhere to data management regulations. Such a framework could allow for increased collaboration but still give researchers the autonomy to decide when and where collaboration is needed or viable.

To conclude, we list our key recommendations resulting from this study that, if acted upon, will help facilitate sharing and integration of OA research data in the future.

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). Ensure there are collaborative opportunities between OA and data science researchers. 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.

4). 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.

Declarations

Ethics Approval and consent to participate:

The study was granted favourable review by the University of Warwick Biomedical & Scientific Research Ethics Committee (BSREC; Ref: REGO-2019-2360) and the Health Research Authority prior to any data collection taking place.

Consent for publication:

Not applicable.

Availability of data and materials:

Data sharing is not applicable to this article as no quantitative datasets were generated or analysed during the current study. Transcripts from the interviews are not publicly available due to them containing potentially identifiable information.

Competing Interests:

The authors declare that they have no competing interests.

Funding:

This study was funded by the Engineering and Physical Sciences Research Council (EPSRC) via the OATech Network Plus (Grant Ref: EP/N027264/1). The funders played no role in the design, collection, analysis, and interpretation of results or in writing of the manuscript.

Authors' contributions:

All authors were involved in the design, set up and recruitment stages of the study. JE collected and analysed the data. JE and MTE wrote the manuscript. All authors read, commented on and approved the final manuscript.

Acknowledgements:

Not Applicable.

References

1. Palazzo C, Nguyen C, Lefevre-Colau MM, Rannou F, Poiraudeau S. Risk factors and burden of osteoarthritis. Annals of physical and rehabilitation medicine. 2016 Jun 1;59(3):134-8.
2. Driban JB, Sitler MR, Barbe MF, Balasubramanian E. Is osteoarthritis a heterogeneous disease that can be stratified into subsets?. Clinical rheumatology. 2010 Feb 1;29(2):123.
3. Ross JS, Krumholz HM. Ushering in a new era of open science through data sharing: the wall must come down. JAMA. 2013 Apr 3;309(13):1355-6.
4. Lo B, Goodman SN. Sharing clinical research data—finding the right balance. JAMA Internal Medicine. 2017 Sep 1;177(9):1241-2.
5. Longo DL, Drazen JM. Data Sharing. The New England journal of medicine. 2016 Jan 21;374(3):276.
6. Ding C, Zhang Y, Hunter D. Use of imaging techniques to predict progression in osteoarthritis. Current opinion in rheumatology. 2013 Jan 1;25(1):127-35.
7. Hunter, D. (2009). Risk stratification for knee osteoarthritis progression: a narrative review. Osteoarthritis and Cartilage, 17(11), pp.1402-1407.
8. Ren G, Krawetz R. Applying computation biology and “big data” to develop multiplex diagnostics for complex chronic diseases such as osteoarthritis. Biomarkers. 2015 Nov 17;20(8):533-9.
9. Nevitt M, Felson D, Lester G. The osteoarthritis initiative. Protocol for the Cohort Study. 2006;1.
10. Eckstein F, Kwoh CK, Link TM, OAI investigators. Imaging research results from the Osteoarthritis Initiative (OAI): a review and lessons learned 10 years after start of enrolment. Annals of the rheumatic diseases. 2014 Jul 1;73(7):1289-300.
11. OATechNetwork+ [Internet]. 2020 [cited 29 September 2020]. Available from: https://www.oatechnetwork.org/
12. Data Protection Act 2018, c. 12 Available at http://www.legislation.gov.uk/ukpga/2018/12/contents/enacted (Accessed: 29 September 2020).
13. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Van Staa T, Smeeth L. Data resource profile: clinical practice research datalink (CPRD). International journal of epidemiology. 2015 Jun 1;44(3):827-36.
14. 2020-21 C, London I, London I, Campus S. Imperial College Healthcare Tissue Bank [Internet]. Imperial College London. 2020 [cited 29 September 2020]. Available from: https://www.imperial.ac.uk/imperial-college-healthcare-tissue-bank/
15. Software – REDCap [Internet]. Projectredcap.org. 2020 [cited 29 September 2020]. Available from: https://projectredcap.org/software/
16. OpenClinica [Internet]. Openclinica.com. 2020 [cited 29 September 2020]. Available from: https://www.openclinica.com/

17. UK Biobank [Internet]. Ukbiobank.ac.uk. 2020 [cited 29 September 2020]. Available from: https://www.ukbiobank.ac.uk/
18. National Research Council. (2011). Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. National Academies Press
19. Veillette, C. J., & Jurisica, I. (2015). Precision medicine for osteoarthritis. In Osteoarthritis (pp. 257-270). Adis, Cham.
20. Kokkotis, C., Moustakidis, S., Papageorgiou, E., Giakas, G., & Tsaopoulos, D. E. (2020). Machine Learning in Knee Osteoarthritis: A Review. Osteoarthritis and Cartilage Open, 100069.
21. Tiffin, N., George, A., & LeFevre, A. E. (2019). How to use relevant data for maximal benefit with minimal risk: digital health data governance to protect vulnerable populations in low-income and middle-income countries. BMJ Global Health, 4(2), e001395.
22. Tiffin N, George A, LeFevre AE. How to use relevant data for maximal benefit with minimal risk: digital health data governance to protect vulnerable populations in low-income and middle-income countries. BMJ Global Health. 2019 Apr 1;4(2):e001395.
23. National Joint Registry [Internet]. Njrcentre.org.uk. 2020 [cited 30 September 2020]. Available from: https://www.njrcentre.org.uk/njrcentre/default.aspx
24. Tillmann R, Maizen C, Bijlsma P, Firth G. Cerebral palsy integrated pathway (CPIP) of hip surveillance, for children with non- cerebral palsy diagnosis?. Physiotherapy. 2020;107:e208-e209.
25. Zhang, J., Sood, H., Harrison, O. T., Horner, B., Sharma, N., & Budhdeo, S. (2020). Interoperability in NHS hospitals must be improved: the Care Quality Commission should be a key actor in this process. Journal of the Royal Society of Medicine, 113(3), 101-104.
26. 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).
27. Van Den Eynden, V. (2008). Sharing research data and confidentiality: Restrictions caused by deficient consent forms. Research Ethics, 4(1), 37-38.
28. 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.
29. 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.