Data sharing in qualitative research: opportunities and concerns

Cormac McGrath[1], Gustav Nilsonne[2]

Corresponding author: Dr Cormac McGrath cormac@mcgrath.se
Institution: 1. Department of Education, Stockholm University, 2. Karolinska Institutet
Categories: Educational Theory, Research in Health Professions Education

Received: 31/10/2018
Published: 13/11/2018

Abstract

Data sharing is increasingly practiced by researchers and mandated by research funders as well as scientific journals. However, data sharing within qualitative research paradigms is less common, and sharing interview data has particular challenges. Earlier debate has pointed to the value of data sharing for discouraging research fraud and permitting critical scrutiny. We elaborate on this discussion by highlighting the value of data sharing for cumulative science, for re-use, and to maximise the value of the participants’ contribution. We review methods and possibilities for sharing interview data, and give concrete recommendations for mitigating risks to the participants. In conclusion, we find that sharing of interview data is possible, valuable, and ethical, and serves a purpose for both journals and researchers.

Keywords: Open data; data sharing; qualitative oriented research

Introduction

Concerns about the rigour, quality and transparency of qualitative research methods have been discussed across the field of medical education for more than a decade, but the topic has not received much attention in broader medical education. In 2006, DiCicco-Bloom and Crabtree drew attention to value of using qualitative methodologies in medical education research, citing among other things the benefits of getting close to respondents experiences of different phenomena (DiCicco-Bloom and Crabtree, 2006). Reeves et al (2006) argued however that qualitative research in medical education, where many new researchers may come from other research traditions, runs the risk of becoming sloppy, and called for increased rigour through closer attention to stringent methodological approaches(Reeves, Lewin and Zwarenstein, 2006). More recently, Cate et al (2013) discussed the risk of fraudulent data and called for a number of measures, among them strengthening the cultural integrity in research teams, confirmation of the completeness and accuracy of data sets by multiple authors, routinely providing reviewers the opportunity to see data, requiring independent review of all original data, and encouraging journals to publish studies that do not report significant effects (ten Cate et al., 2013). In commenting Cate et al, Peeraer and Stalmeijer (2014) asked more specifically what can be done to combat research fraud in qualitative research, and advocated for data
auditing (Peeraer and Stalmeijer, 2014). Peeraer and Stalmeijer argued that researchers using qualitative methodologies must develop high standards for data auditing and sharing (Peeraer and Stalmeijer, 2014). At the same time, the same authors supported having multiple authors verify datasets and member checking procedures, and advocated these activities within qualitative research. Further, the authors also suggested that data audit should be done at the request of reviewers (Peeraer and Stalmeijer, 2014). Pusic (2014), put forward that it is high time for scientific journals to publish research data openly, while also acknowledging that privacy concerns must be addressed (Pusic, 2014). Pusic argued that data sharing can facilitate research replication and synthesis, and further that re-analysis of shared data could allow new conclusions and interpretations which were not addressed by the initial authors (Pusic, 2014).

In medical education research, interviews are a near ubiquitous data collection tool (McGrath, Palmgren and Liljedahl, 2018). Interview techniques include structured, unstructured and semi-structured interviews, and they may be combined with other observational data collection methodologies (DiCicco-Bloom and Crabtree, 2006; Jamshed, 2014; McGrath, Palmgren and Liljedahl, 2018). Interviews can address a broad range of quantitative or qualitative oriented research questions, and may be used on their own or as part of a mixed methods approach (Banner, 2010). Interviews generate data, usually in the form of recorded audio files, transcripts, and notes and other observational data.

In this paper, we elaborate on previous discussions on the extent to which is it possible, valuable and ethical to share qualitatively generated data. We aim to move the debate beyond research integrity, and to highlight the value of data sharing for re-use and synthesis. We will specifically address benefits and challenges pertaining to the sharing of interview data. While principles for sharing interview-data are not unique to the field of medical education, we believe it is valuable that data sharing is addressed within the context of the existing debate in our field.

Data sharing landscape

Currently, there is a worldwide push towards data sharing in many fields, among them; engineering, learning analytics, and the broader field of biomedical research (Kalampokis, Tambouris and Tarabanis, 2011; Siemens and Baker, 2012). Data sharing is increasingly mandated by scientific journals and funders; e.g. the International Committee of Medical Journal Editors (ICMJE), the Wellcome Trust, and the Gates foundation (Walport and Brest, 2011; Taichman et al., 2016).

In a similar vein, the European Commission is pushing for a set of norms and behaviours encapsulated in the concept of Responsible Research and Innovation (RRI). One main point in this effort is to ensure free and open access to research outputs, including data. In RRI, three general principles are identified in relation to open data sharing: Availability and access, re-use and redistribution, and universal participation (Von Schomberg, 2011). The principle of availability and access means that data must be available as a whole and at no more than a reasonable reproduction cost, preferably over the internet, in a convenient and modifiable form. The principle of reusability and redistribution means that data must be provided under terms that permit re-use and redistribution including synthesis with other datasets, and the principle of universal participation entails that others must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups.

Many data sharing policies now advocate the FAIR principles, which state that data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016). These principles highlight the need to share data in a format that is machine-readable, with sufficient metadata to be understandable by humans and machines, and with a licence that makes it clear how data may be re-used. It is recommended to use a data repository, which provides long-term storage and a citable digital handle.
Benefits of data sharing

Data sharing may be valuable for three main purposes: permitting critical scrutiny, facilitating cumulative science, and enabling re-use. Critical scrutiny is an essential part of the scientific process. The possibility of scrutiny of open data is likely to reduce the risk of falsification and fabrication, and can therefore strengthen scientific integrity, as argued by Pusic and Cate et al., (ten Cate et al., 2013; Pusic, 2014). Cumulative science refers to the process by which evidence is gathered and increasingly provides better estimates and clearer support for certain theories over others. A first requirement for cumulative science is that results are reported. However, new theories and methods may prompt a re-evaluation of earlier data. Access to existing data is therefore vital in order to be able to assess and interpret these data in the light of the current state of knowledge. Re-use of existing data may serve to investigate new questions, to investigate old questions in a new way, or to validate findings made in other datasets, among other things. Resources can be saved, and risks to participants reduced, when existing data are used instead of gathering new data.

To conclude this section, we wish to highlight that data sharing can help researchers fulfil the ethical obligation towards research participants to realise the most value from their data. Research participants undertake risks and harms, e.g. with respect to privacy, because the expected knowledge benefit exceeds these risks and harms. If data are lost and never reported on, there is no knowledge benefit, and the research can hardly be justified from an ethical point of view. By contrast, if data are shared, then the greatest gain may be achieved (Poldrack and Gorgolewski, 2014). Value gained from data re-use may be seen as a benefit to participants, particularly when such value can lead to improvements for the population from which participants were drawn, e.g. in terms of improved health care or teaching practices.

Data collection and sharing in research in medical education

Data collection among qualitatively oriented researchers is, in interview contexts, often conducted using a dictaphone and a notebook, and the data sharing extends seldom beyond the immediate research team. Further, there are, to our knowledge, few systematic ways of electronically documenting the research process from design to data collection to analysis and interpretation. As a result, there is no means for external audit, for checking fraudulent data, for considering other interpretations, leaving qualitatively oriented researchers vulnerable to claims of sloppy research (Reeves, Lewin and Zwarenstein, 2006; ten Cate et al., 2013; Peeraer and Stalmeijer, 2014; Pusic, 2014; Wilkinson et al., 2016). This situation is confounded by journals lacking standards, practices, and means to share data via supplementary files available open access or upon request. The results means that the data rarely aligns with the FAIR principles (Wilkinson et al., 2016), and are rarely findable, accessible, interoperable, or reusable. While data transcripts may run to many hundreds of pages, these data sets require little digital storage space (Pusic, 2014). As such, compliance with the principles outlined above is seemingly achievable; the data is possible to make available as it is documented and labelled it is also discernible and can be intelligible to others. Further it is possible to re-use and re-distribute provided there is thorough documentation. It is also possible for others to engage in universal participation provided there is informed consent and ethical approval for such practice.

Consequently, it seems clear that the different principles and recommendations outlined above make a strong case for sharing research data, and do not make unwarranted demands on researchers engaged with qualitative-oriented research but instead could be seen as a way of enhancing a more stringent and rigorous and traceable process. However, challenges remain.
Challenges to open data sharing in the medical education context

A number of concerns are raised in relation to sharing qualitatively oriented interview data, they include; difficulties in anonymizing the participants, concerns that critical voices with a view on workplace and clinical environments or similar may not be shared for fear of reprimands, that people will not willingly offer their information if they know the data will be shared in the public domain (Saunders, Kitzinger and Kitzinger, 2015). Further concerns may relate to disciplinary and paradigmatic boundaries, for example in STEM disciplines, where many research studies involve large data sets, there may exist different views and arguments in relation to the possibility and value of aggregating data than, for example, in research in medical education conducted using interviews (Braver, Thoemmes and Rosenthal, 2014). Moreover, our experience suggests that in qualitatively driven work in research in medical education, it is not uncommon that the specific and idiosyncratic value of the study context is evoked above arguments for sharing data.

In this paper, we argue that many of the concerns raised above could perhaps be easily addressed; interview subjects could be asked if they wished to share their data publicly through the informed consent. Member checking could be used whereby the interviewee approves the audio or transcribed product of the interview. Sensitive or private sections of data from an interview could be redacted on the specific instructions of the respondent as part of member checking (Angen, 2000; Morse et al., 2002).

Taking steps towards sharing data generated in qualitatively oriented research may, however, require even more rigorous approaches to research, and would perhaps require integrating meta-analytical modes throughout the work of study-design, data collection, analysis and interpretation (Reeves, Lewin and Zwarenstein, 2006). Engaging in such a process could also lead to increased demands on transparency of data collection and data documentation. The sharing of open data could also act as a safeguard against scientific malfeasance and sloppy science by creating new opportunities to peer review (Pusic, 2014).

Practical recommendations for ethical sharing of interview data

Below we outline a number of measures which can facilitate sharing interview data.

1. **Informed consent**

   Informed consent is established practice. The principal reason why participants must be informed about how exactly data will be managed and shared is to respect the principle of autonomy, i.e. allowing the participant to control whether to contribute data in the full knowledge of how these data may be used in the future (Department of Health, Education, and Welfare and National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 2014). A further reason involves the principle of beneficence: if the participant is aware of any risks associated with their information, they may adapt their expectations and behaviour accordingly, and reduce the risk of harm. Further, information to participants should make it explicitly clear that the data may be shared with other researchers and in which shape the sharing may take place.

2. **Data minimization**

   To reduce risks, it is best to collect as little information as possible that may be either potentially identifying or that may cause harm when shared. Concrete actions can include greeting the participant by name before turning on the recorder, instructing the participant to avoid mentioning potentially identifying information unless pertinent to the question at hand, and to not ask about potentially sensitive information unless this is clearly motivated by the research question.
3. **Pseudonymization**

In most cases, it is easiest to pseudonymize data during primary transcription. If necessary, two parallel transcripts may be generated, and only the pseudonymized transcript shared. Direct and indirect potential identifiers should be removed from data and metadata (Hrynaszkiewicz et al. 2010). Direct identifiers, such as name or personal identification numbers, can allow deterministic identification, i.e. identification with certainty (Dusetzina et al., 2014).

Probabilistic re-identification is possible when participants can be matched by indirect identifiers. Interview data may contain indirect identifiers such as the general location of a person’s home or the type of workplace where they work. It is recommended to perform a privacy impact assessment for each research project, in order to determine the possible risks that may occur due to re-identification (Clarke, 2009).

Pseudonymization of interview data usually involves replacing names of persons and entities with pseudonymous names. Care should be taken to not automatically find and replace names of persons, as such names may be revealed e.g. in idiomatic phrases such as "taking the Mick", if a person in the data set is named Mick.

4. **Stepped access**

A further risk-mitigating practice for individual participant data is to use controlled/stepped access, where sensitive or riskier data may be shared on request following a structured procedure (Tudur Smith et al., 2015). A list of repositories providing a stepped access service is maintained by the Center for Open Science at [https://osf.io/tvyxz/wiki/8.%20Approved%20Protected%20Access%20Repositories/](https://osf.io/tvyxz/wiki/8.%20Approved%20Protected%20Access%20Repositories/).

Here, journals need to establish guidelines for facilitating review and ensuring data are available as specified in the article reporting on the data. One example of such guidelines are the Transparency and Openness Promotion (TOP) guidelines (Nosek et al., 2015), now adopted by more than a thousand scientific journals. Ethical guidelines will be necessary also to ensure data integrity is maintained.

5. **Member checking**

Member checking refers to having the transcript checked by the interviewee, who may then consent explicitly to sharing of the transcript. This method places the locus of control firmly with the research participant and allows for a fully autonomous and informed decision. Member checking may be done for a number of reasons; to review transcripts to consider if their words match their intended meaning, to check the accuracy of the interview transcript or to validate the researchers findings. (Varpio et al., 2017; McGrath, Palmgren and Liljedahl, 2018). One possible challenge with this method is how to handle the situation if the participant wishes that the record be altered such that the content becomes substantively different. In this case, we argue that member checking be allowed for the purpose of checking coherence of transcript with intended meaning only, and do not recommend member checking the final interpretation.

**Conclusion**

The question raised here is whether data sharing is possible, valuable and ethical. In the paper, we have demonstrated that is it possible to share through data repositories or upon request, we have demonstrated that there are sound arguments for sharing open data and we have demonstrated that there are ethical ways to safeguard the identity of the respondents. Our reasoning also suggests that it may be conducive to good practice for journals to consider requesting authors to make their data available using a stepped access model. Our hope is to initiate a discussion in the medical education context on the sharing of data generated through qualitative data collection methods.
Take Home Messages

- Qualitative oriented data can be shared through data repositories or upon request
- Data sharing can help fulfil the ethical obligation towards research participants to realise the most value from their data
- Data sharing may enable more rigorous and comprehensive peer review
- Planning for data sharing may lead to higher quality in data management and annotation
- Journals publishing qualitative research may request data sharing, e.g. using stepped access.

Notes On Contributors

Cormac McGrath, Orcid: 0000-0002-8215-3646 PhD is a senior lecturer and academic developer at Stockholm University and Karolinska Institutet, with main research interests on curriculum design, education interventions, MOOCs and leadership.

Gustav Nilsonne, Orcid: 0000-0001-5273-0150 (MD, PhD) is a researcher at Stockholm University and Karolinska Institutet in neuroscience, working largely with sleep and diurnal rhythms. Main methods are magnetic resonance imaging and meta-analyses. Gustav also takes a strong interest in open science and reproducible research.

Acknowledgements

We thank Stefan Ekman of the Swedish National Data Service for valuable comments on a draft version of this paper.

Bibliography/References


### Appendices

None.

### Declarations

The author has declared that there are no conflicts of interest.

This has been published under Creative Commons "CC BY 4.0" ([https://creativecommons.org/licenses/by-sa/4.0/](https://creativecommons.org/licenses/by-sa/4.0/))

### Ethics Statement

This paper presents a critical discussion of data sharing in the context of research utilizing predominantly qualitative oriented data collection. No empirical data is presented herein.

### External Funding

This paper has not had any External Funding

AMEE MedEdPublish: rapid, post-publication, peer-reviewed papers on healthcare professions’ education. For more information please visit [www.menedpublish.org](http://www.menedpublish.org) or contact [mededpublish@dundee.ac.uk](mailto:mededpublish@dundee.ac.uk).