Edinburgh Research Explorer

The long and winding road

The long and winding road: archiving and re-using qualitative data from twelve research projects spanning sixteen years

Dodds, Catherine; Keogh, Peter; Bourne, Adam; McDaid, Lisa; Squire, Corinne; Weatherburn, Peter; Young, Ingrid

This is the Author Accepted Manuscript of the Article. Full publication details are above, and a link to the Version of Record/DOI will be added as soon as it is available. The only differences between this version and the VoR are very minor changes that occurred in the copy-editing stages.

Abstract:
We describe a pilot project designed to assess the feasibility of re-use across 12 diverse qualitative datasets related to HIV in the UK, from research projects undertaken between 1997 and 2013 – an approach which is chronically under-used. First, we consider the sweeping biomedical changes and imperatives relating to HIV in this timeframe, offering a rationale for data re-use at this point in the epidemic. We then reflexively situate the processes and procedures we devised for this study with reference to relevant methodological literature. Hammersley's (2010) and Leonelli's (2016) contributions have been particularly instructive through this process, and following their lead, we conclude with further considerations for those undertaking qualitative data re-use, reflecting on the extent to which qualitative data re-use as a practice requires attention to both the *given* and the *constructed* aspects of data when assembled as evidence.

Keywords: data re-use; secondary data analysis; methods; qualitative; HIV; biomedicalization

Introduction:
The past 25 years have given rise to considerable discussion among social scientists regarding the re-use of qualitative research datasets to generate new insights during the exploration of new questions (Heaton 2004, Mason 2007, Moore 2007, Hammersley 2010, Slavnic 2013, Tarrant 2016, Davidson et al. 2019). Some of the tantalising possibilities afforded by this way of working are indicated by Walters' statement that:

> "the ability to revisit qualitative data in light of social change may allow the future researcher to attribute those participants with a degree of prescience about future social conditions that the original researcher was in no position to understand." (2009: 313)

While such literature reflects the growing attention that is being paid to this emergent methodology, a recent review of actual qualitative data re-use since 1997 verified that *only* 347 published research papers that had used this approach (Bishop and Kuula Lummi, 2017). Given the proliferation of Open Data imperatives emerging from funders, publishers, ethics committees and governments (Bishop and Kuula-Lummi, 2017; Corti et al., 2014) - the identification of this data re-use 'practice gap' in the social sciences provides evidence of a considerable lag in willingness, readiness and capacity to re-use qualitative data, with comparatively few researchers pursuing this approach. As Slavnic (2013) and others have stressed, there has scant investment in the skilled preparation, documentation, curation and archiving of qualitative social science data for re-use. So perhaps it should not be surprising that as a result, we don't yet have a culture of qualitative data re-use, particularly when we compare this scenario to the well-funded international and interdisciplinary infrastructures organised to support the exchange and re-use of biological and biomedical quantitative data (Leonelli 2016).

In addition, the relevant methods literature tends to be dominated by extensive debates about feasibility and validity, with few critical methodological reflections from actual practitioners. Perhaps this is one of the reasons why Davidson and colleagues describe many social research peers asking, "Why would you want to do that?" (2019: 365). This paper therefore adds to a small but emergent body of academic writing that has started to offer answers to that question, and many others. Using social aspects of Human Immunodeficiency Virus (HIV) in the UK as a case study, we outline the motivations driving our collaborative project on the collation and re-use of qualitative datasets. Furthermore, this work offers a response to Hammersley's (2010) call to consider re-used data as being both given *and* constructed. Ultimately, we describe how and why a network of social scientists working in the HIV field explored the feasibility of re-using data from numerous datasets spanning sixteen years and involving nearly 600 participants. We pay particular attention to the specific demands involved with working across projects and across institutions, and with sensitive data frequently collected from those in marginalised groups.

Why re-use qualitative data on HIV now?
More than thirty years into the epidemic, HIV continues to be scrutinised from a wide array of disciplinary perspectives, across its local and global socio-economic, political and biomedical contexts. Our knowledge about HIV is inevitably contingent upon, and shaped by, biotechnological development, in addition to: the range and variability of geographic locations where HIV is concentrated, the demographic sub-groups most affected in each of these locations, the extent to which those groups are strongly or weakly networked, and the impact of successive biotechnological developments which themselves are impacted by material and social change. In particular we draw attention to some key moments in the epidemic, such as: the introduction of HIV tests (1985); introduction of highly successful antiretroviral (ARV) treatment (1996); use of ARVs in the prevention of transmission to children (2000); and ARVs for adult prevention (2008 onwards) - both through the use of ARVs to render people with HIV un-infectious, as well as their prophylactic use among those who may be exposed to the virus. It is in response to this latest critical juncture in the biotechnological history of HIV that we initiated our data re-use project in 2015, during a time when the success and re-purposing of HIV ARVs is promoted as the key to ending AIDS within a generation (UNAIDS, 2014). Those with a critical eye for the social science of medicine have rightly regarded such claims with caution (Kippax and Stephenson, 2016), given the unequal terrain of ARV access and informational bio-citizenship (Fassin, 2007; Rose 2007). It is out of concern for the likelihood of particular groups and classes of people being left behind in this technological race for the 'AIDS' finish line, that our work seeks to engage directly with a considerable volume of social science evidence regarding people's embodied experience of biomedical change, and the ways in which this experience has always been socially stratified, complex, uncertain and slow-paced (Davis 2010; Keogh and Dodds, 2015; Kippax and Stephenson 2016; Mykhalovskiy et al., 2004; Nguyen 2010;

Paparini and Rhodes, 2016; Persson et al., 2016; Squire, 2013; Young et al., 2019). We want to find new ways to explore how these experiences have already unfolded alongside the deeply engrained moral attitudes about sex and HIV which have inscribed political, economic, personal and medical responses to the HIV epidemic. We therefore believe it to be not only reasonable, but essential, to raise strategic questions with the support of qualitative data re-use, in order to critique a set of highly technological global and local HIV policy propositions featuring pharmaceuticals which promise rapid-fire change, while demonstrating insufficient regard both for the social context within which HIV is played out, and the bodies upon which these events continue to trace diverse biosocial histories.

Datasets collected as a part of qualitative HIV studies are often chronically underutilised and rarely re-examined, particularly where the drive for applied research has maintained an explicit focus on the design of treatment, prevention and care services. This vast evidence base (like so many others) is generally underpinned by underlying epistemological presumptions that new understandings will be exclusively driven by the collection of new data. However, there are other ways to develop knowledge, including those which ask new questions of old data in order to purposively disrupt embedded epistemologies. This paper offers an account of this disruptive journey into data re-use undertaken by UK social scientists working on HIV throughout a period of intense biotechnological change. We wanted to see if it was possible to ready a large set of qualitative UK social science datasets for re-analysis, enabling a deeper exploration of the changing nature of engagements with ARVs over time. We felt this should enable the development and application of research questions that simply would not have been available to the researchers who were working *in situ*. We surmised that this process would help to develop rich insights, rather than collecting even more new data in the here-and-now.

From the outset, we had a shared interest in exploring how biomedicalisation of HIV was, and is, unfolding in the UK context (Clarke et al., 2003; Nguyen, 2010; Squire 2013), specifically through the repurposing of ARVs for prevention and attentive to the ways in which these global claims and trends might contribute to broader understandings of the pharmaceuticalisation of public health (Bell and Figert, 2015). Therefore, we were also interested to explore these data for any anticipations of biomedicalisation and its effects, with a keen eye on the 'history of the present'. Our intention has thus been not only to inform, but to help re-situate the HIV social science research agenda going forward, given the enormous strategic and economic attention being directed toward the exclusive use of ARVs to end the epidemic. As Mason points out, "some forms of interpretation are only possible from a distance" (Mason, 2007: 3.2), and the global policy focus on ending the epidemic by 2030 with biotechnical solutions meant the time was right for us to return to the rich qualitative material collected from the frontlines of the HIV epidemic in the UK over the past two decades.

Project aims and procedures
Our aim was to assess the feasibility of data re-use through the merging, sharing, archiving and pilot analysis of a considerable volume of qualitative data spanning the past two decades. This work was supported by the Wellcome Trust (grant 110452/Z/15/Z). The first two authors (CD and PK) undertook the majority of the work on the project with support from a research administrator, alongside periodic engagement and reflection from steering group members. As the project proceeded, a range of sizeable demands emerged at different stages, including considerable time and effort required to locate and anonymise samples, and also to ensure all bureaucratic processes including the completion of data-sharing forms were undertaken in accordance with different universities' procedures. Emergent practical issues as well as key themes were considered by the project steering group, and the lead authors kept detailed records of these developments in order to assess feasibility.

Methodological sources of inspiration

As a network of social scientists working on HIV in the UK, our starting point on this project was our shared knowledge of, and immediate access to a considerable volume of unarchived qualitative data. As novices in qualitative data re-use, we turned to the methodological literature for guidance. It was immediately evident that this literature was filled with debate about whether data re-use is feasible or useful (see Slavnic, 2013 for an overview). Rather than becoming enmeshed in these debates, we found a good ideological fit with the work of Mason (2007), Hammersley (2010), and Tarrant (2016). These researchers speak of qualitative "data re-use" to refer to a wide set of practices that involve a *return to* or a *repurposing* of data – sometimes in isolation, but more frequently in conversation across datasets. Their work variously presents qualitative data re-use both discovering and applying new research questions that have emerged subsequent to the original period of fieldwork and analysis. This approach enables the exploration of elements that were originally ignored, invisible, or not yet in existence, and it can support researchers in devising new strategic directions for their research. While Heaton (2004; 2008) and others refer to this type of work as "secondary analysis", this term can infer that there is a singular type of analytical practice that is attached to qualitative data-reuse, which is not at all the case. Instead, as Davidson and colleagues have recently asserted, the selected format of final in-depth analysis is generally "of the type that is familiar to most qualitative researchers" (2019: 363), while others have pointed out that there is a long history of social scientists returning to their own data (Moore, 2007).

We found particular inspiration in Mason's (2007) championing of qualitative data re-use as a part of what she calls an "investigative epistemology"; characterised by creativity, purpose, and energy, while still being underpinned by critical reflectiveness. Furthermore, she points out that when we start to acknowledge that the politics of reflexivity and interpretation is complex by its very nature, then it can be quite useful to consider data both from near and far perspectives – both embedded within and extracted from their original contexts – as a means of undertaking more fruitful investigations (Mason, 2007: 3.3). She reminds us about the value that curiosity and freedom can afford to our research efforts, and that data re-use, while not without its challenges, offers just such an opportunity for the expansion of knowledge. Thus, researchers are prompted to consider the potential benefits of temporal and/or biographical distance from the processes of production, which can help re-confirm the validity of the original researchers' findings (Haynes and Jones 2012). We can bring new reflexivity, critical distance and rigour to qualitative data re-use in ways that can facilitate fresh insights, and which can serve to strengthen or challenge existing arguments.

Our team set out to establish the feasibility of re-using a considerable expanse of data: an endeavour somewhat akin to the scope and ambition of the ESRC Timescapes project (Irwin et al., 2012). As such, it has been essential to expand our traditional ways of working, acting on Moore's injunction (2007) to breach boundaries. When doing so, Moore stresses that "eschewing our comfort zones, and developing a more creative, and even messy, approach may be the key to opening up the full potential of qualitative data reuse" (2007: 4.8). This is an approach which felt suited to our project, given that we were assessing how to apply completely novel research questions about the steady emergence of novel applications of *ARVs for prevention* to data collected before such uses of ARVs were fully conceived, evidenced and recommended. In embracing Moore's call for creative and energised approaches to qualitative data re-use, we were hopeful that these artefacts from the past (Zimbra et al., 2010) might help us to better understand our shared present.

It was ultimately Hammersley's (2010) thoughtful reflection on the nature of data and of its re-use, that helped us to sit more comfortably with the complex task we had set ourselves.

> "…these two meanings of 'data', as *constructed* and as *given*, are both essential: they relate to different but equally important aspects of the material we use as grounds for inference in

research…we collect data as a resource and then use some of it, in particular ways, as evidence, in order to draw inferences relevant to our research focus; and we discover how to do this in the course of our work. So, in using data, we necessarily reconstitute it as evidence." (Hammersley, 2010: 4.6, our emphasis)

Hammersley goes on to describe how, in every qualitative social research project, researchers variously codify some materials as data – depending on the extent to which they conform with our research questions. The status of these materials will vary throughout the changing course of the project, alongside the shifts in research questions that frequently occur. Therefore, we organise, compare, select and re-form these data to generate evidence, and we are "reforming something that already exists, not making it up" (Hammersley, 2010: 4.7). This account chimes closely with the experiences described by researchers in our network, while each of us collected, curated and used the data originally, and also while considering the feasibility of data re-use.

Parallel reflections have also been made by Leonelli (2016) about the essentially human task of packaging and mobilising data in order to enable it to 'travel' on what she refers to as 'data journeys' across time, space and research teams. In her philosophical reflections on these practices as they relate to the construction of online quantitative data-sharing databases in the biological sciences, we can see considerable synchronicity with the thoughts that Hammersley offers up to social scientists regarding data re-use. In entirely distinct research spheres, Hammersley and Leonelli implore us to maintain awareness of the inevitable social situatedness of our practices of data collection, preparation, curation and re-use (across varied personal, institutional, disciplinary, political and economic terrains) – and to use these reflections to enrich our explorations. Of particular salience is Leonelli's discomfort with the metaphoric use of the term 'data flows' - as applied in data-centric scientific endeavours. As she states:

> "Not only do data not 'flow' toward discovery, but it is the lack of smoothness and pre-defined direction that makes their travel epistemologically interesting and useful." (Leonelli, 2016: 41)

Similar to Hammersley, then, Leonelli exhorts us to see data not simply as inert objects that hold meaning independently, but instead to also see their capacity to collect and build evidential value through the very act of mobilisation on a data journey which is influenced by every person, system and contextual situation that has touched those data along the way.

Hammersley's and Leonelli's contributions therefore informed our thinking considerably, as they helped to enliven the possibilities of what qualitative data re-use enables. Their work encourages us to anticipate epistemological divergence in the ways that researchers engage with data as we use what is given *and also* as we create something anew. At times it was easy to feel like we might be trying to engage with an unmanageable volume of 'given' material, contributing to a sense of alienation and disconnection. And yet, we also remained attuned to Hammersley's reassurance that there are always periods within qualitative analysis when there is a feeling of unsteadiness, whether or not we are re-using data. These considerations endowed us with a useful balance, even when we were feeling slightly at sea in a mass of data, with a host of emergent questions.

The practicalities of getting started

At the outset, we assembled eight UK social scientists who had worked in the UK across the past two decades, leading a wide variety of qualitative projects focussed on the social, policy and behavioural aspects of HIV in different parts of the country. All were interested in exploring the feasibility of data re-use, so we identified projects from between 1997-2013 led by members of this collaboration which were most likely to offer insights into the emergence and development of HIV technologies following the introduction of ARVs. Most of the studies utilised semi-structured or narrative

interviews with individuals, though some used focus groups, or a blend of interviews and focus groups, and one employed structured questionnaires via telephone.

Each researcher took responsibility for checking the quality and coherence of their materials, in order to assess what format they were held in and to consider how to make changes to file formats if necessary; and also to identify any files that were missing, corrupted, inaccessible or incomplete.

VIGNETTE #1

We needed to share data across four (current) UK institutions of higher education, and in addition, many of the datasets were collected during earlier employment at different universities. This raised a curious set of questions about institutional ownership of data and completion of data-sharing proforma when the data in question had travelled with a researcher. Our discussions about institutional attitudes toward the data that travels with us when we start new jobs were never fully resolved. We recommend that other researchers initiating formalised data-sharing discussions (which are frequently framed as simply being institution-to-institution) will want to take time to consider in detail how institutional attitudes towards such data are likely to impact on 'sharing' procedures.

All of this activity was tracked so that the group could maintain a record of the status of potential data materials. Quite quickly, we were able to identify that we might potentially work with data emerging from 741 diverse participants, with just over half this sample being gay, bisexual or men who have sex with men of a range of ethnicities and HIV statuses, and a fifth of the sample being black African people with HIV from a range of sexualities (together these groups account for the majority of HIV infections in the UK). The remainder included those working and providing care in the HIV sector, and others with and without diagnosed HIV. Study topic guides, questionnaires, outputs and basic participant demographics were collated and shared, as well as contextual details regarding study design and data collection. One outcome of this stage of appraisal was the realisation that the original material from one of the candidate datasets had been destroyed, and so would not be available for re-use.

We devised a comprehensive data management plan to support the completion of data-sharing agreements. The plan articulated our core principles of data governance and curation. Most helpfully, focusing on these matters at an early stage led us on to the creation and testing of a tailor-made anonymisation protocol (discussed in further detail below). We emphasise these somewhat day-to-day details here, because any team who is considering simultaneous data-sharing, archival depositing and data re-use across multiple projects that span diverse institutions need to be apprised of the considerable amount of administrative labour that is required to prepare the data for re-use. As our team discovered at all stages of this work, we routinely underestimated the required time needed for the more mundane aspects of data assemblage and collaborator communications.

The Anonymisation Protocol
We were strongly committed to minimising the potential for harm to come to these historical research participants as a result of our archiving and re-use of this data. In reviewing the quality and format of the datasets, the researchers recognised that although some of the data had been considered to already be fully or partially anonymised, approaches to anonymisation had varied dramatically both between and within datasets. For this reason, it was necessary to develop a uniform anonymisation protocol to be applied to all data being considered (see supplementary file).

The development of this protocol served a range of unanticipated purposes. Firstly, each member of the steering group was compelled to reflect on what was meant by the notion of 'anonymisation' in

concrete terms, and the extent which such practices were feasible prior to sharing data within the group, and beyond it, via data archiving. It also helped to make some of the challenges of data sharing immediately apparent, as set out in vignette #2.

VIGNETTE #2

Following a delay in sharing two datasets from one source, it was during a phone call that a member of the network explained they had been reflecting for a while and had decided not to share their data (and not to work toward deposit with the UK Data Archive) due to concerns about the risks of participants' identities being unwittingly disclosed. This decision was respected by the remaining network members, and serves as an important learning point, as such endeavours may not be deemed appropriate by all colleagues, for all datasets, or for all research participants. Thus, sometimes a point is reached in a data-reuse project when it becomes clear that either some or all of the data will not be made available for this purpose. We still had a sizeable sample to work with, consisting of 12 datasets, with 589 individual participants.

Following advice from the UK Data Archive (UKDA) our approach to anonymisation required flexible and tailored solutions, enabling each study to be considered within its own right. The protocol stipulated that all direct identifiers (peoples' names, addresses etc.) would be removed and replaced with standardised text in order to indicate that a specific type of identifier had been removed (ie. [name of friend]). When it came to indirect identifiers (such as geographic location and/or place of employment), we resolved to remove as little information as possible, because doing so might strip out important contextual details for our own and others' future analysis. The protocol gives examples where it would be necessary to remove *a series* of indirect identifiers when it was judged that in combination, they could potentially identify an individual, as well as instances where an entire transcript may not be deposited or shared in the same way as other materials in the set, due to risk of identification from detailed narratives that are embedded across the document.

We devised procedures that paid close attention to the following three issues: consent, de-identification practices (anonymising), and regulation of access to data. We were encouraged by the UKDA to think of the way these three issues should be maintained in mutual balance and suspension (Steering Group meeting notes, June 7th 2016), so that where one element was low for a particular project, the others could be strengthened. For instance, we were working with a number of historic projects with consenting procedures that did not mention future re-use. As a direct result, for those projects we decided to considerably heighten our de-identification practices (if needed, removing all possible indirect identifiers), placing restrictions on future access, and at times removing an entire transcript from the dataset if the overall narrative ran a particularly high risk of identification.

Once the drafted protocol was discussed and agreed by the seven collaborating researchers, we randomly selected three data encounters from each of the 12 data sets, and each researcher applied the principles of anonymisation set out in the protocol to this sample of our own work, while keeping a separate record of changes. This enabled us to test and strengthen the protocol, and it also helped all co-authors to recognise that this flexible and context-driven approach to anonymising was a complex, time-consuming and skilled process. It also became clear that the process of anonymisation required contextual knowledge and experience which was gained either through: proximity to the original research project; and/or familiarity with the social contexts and experiences of particular sub-groups involved in these studies. Such insights supported the complex decisions that those undertaking anonymisation needed to reach about assortments of highly contextualised and interrelated information that could potentially identify an individual.

VIGNETTE #3

While applying and testing the anonymisation protocol, one member of the network experienced an unexpected and sustained emotional response to their randomly selected transcripts. This researcher's personal connections with the interviewers and the research participants had come flooding back. Some had died of AIDS not long afterwards, while others who had spoken of being very close to the end of their lives had not died, and are still playing key roles in our HIV research and policy infrastructure.

Although it had not been anticipated, the development and use of this anonymising protocol was a hinge point for the work of this pilot feasibility project. Creating, revising and applying the three-page tool elicited the reflection that was needed to consolidate collaborators' shared norms and ethos, enabling the project to proceed. It also (again) forced the research team to concede that the task we had taken on was considerably more involved and laborious than we had anticipated.

Analytical procedures

The testing of the anonymisation tool on three randomly selected transcripts from each dataset served a further purpose: it meant that the team now had a randomly selected sample of 36 transcripts/notes from across 12 different datasets that we could start to work with for our pilot analysis. Given the scope and scale of the available data for this feasibility study, this smaller sample that was originally compiled to test our anonymisation protocol now presented itself as the most obvious material on which to perform our Stage 1 analysis to determine whether data re-use which incorporated so many diverse sources was likely to result in meaningful outcomes. The two lead authors read and became familiar with this smaller sample, alongside a set of metadata, including the original project descriptions, outputs and question guides.

Using NVivo 10, we then coded material that related to HIV antiretrovirals (ARVs) in any way (both prompted and unprompted), making note of any sections where we may have expected such discussion to arise, yet it did not, given that it is important to listen for silences when re-using data (Irwin et al., 2012). This enabled us to identify for the first time, which datasets might be most fruitful based on how the topic of ARVs was (or was not) situated within these different data sources. Given that this pilot coding elicited a considerable volume of material to work with in its own right, we decided to remain focussed on this pilot sample of 36 data encounters.

For the Stage 2 level of deeper analytical exploration, we used a framework approach (Ritchie and Lewis 2003), focusing upon discourses of treatment literacy and engagement with / or rejection of expert knowledges related to ARVs. This approach enabled us to apply our research questions to varied materials, collected at different points in time, in different places, and with the involvement of diverse participants and researchers (Irwin et al. 2012). We used the themes identified at Stage 2 to generate a more focused inquiry for Stage 3, which will be reported in greater detail in subsequent publications.

Colleagues working on Timescapes have recently described using similar approaches in their re-use of qualitative datasets including around 700 text files, using CAQDAS in order to undertake word and proximity searching based on keywords (Davidson et al., 2019). In our project no computerised searching of this type was required, but could have been considered had we selected a larger sub-sample. However, on a range of other analytical choices, there is considerable parity, including a series of processes that Davidson and colleagues (2019) refer to as:

- Undertaking an overview survey and constructing a corpus
- Recursive surface 'thematic' mapping
- Preliminary analysis
- In-depth interpretive analysis

In particular, we have found their sustained use of archaeological metaphors to describe the work of qualitative data re-use to be particularly helpful, and far more useful than the concept of 'mining', to communicate analytic approaches to re-using multiple qualitative datasets. This is because the former "evoke[s] the idea of moving between breadth and depth while retaining the integrity of a contextualised and detailed qualitative approach" (Davidson et al. 2019: 369). The notion of 'data mining' may have relevance when re-using quantitative data that is potentially quite standardised, with a high degree of coherence between data-sets. However in contrast, we have found that archaeological metaphor encourages us to generate practices that are based on locating diverse forms of data *in-situ*, and on utilising this context as a strength for theory-building, rather than as a weakness. At a purely pragmatic level, due to the scale and the scope of the task that we had set ourselves, using 'test-pits' as mechanisms to undertake surface mapping was highly beneficial, as it fitted in well around multiple research obligations, allowing time for reflection, discussion with other network members, and subsequent returns for further thematic analysis.

VIGNETTE #4

One of the studies was completed in 1998, and as researchers, we broadly recalled that period immediately following the roll out of ARVs as one of heady optimism, of 'Lazarus-type' recoveries, and emptying hospital wards. Yet what stood out for us as we re-engaged with these data were experiences of the new treatments that were often painful, long-term, uncertain and dangerous. Some participants spoke of their struggles to stay on the treatments, of side effects and crippling guilt and anxiety about missing doses. We heard from people who could not believe what they were reading in the press about the transformative moment they were supposed to be at the heart of, because it did not reflect their own embodied treatment experiences. Our team reflected upon how the extent to which we, as researchers active throughout this period had forgotten what these data portray.

Steering Group Meetings
Central to the drive and focus of this project was the support of its Steering Group comprising all original principal investigators on the projects being considered for re-use, a public health historian and a policy expert from a national HIV thinktank. The first two Steering Group meetings helped to meet a range of bureaucratic and strategic functions, while also enabling consolidation and agreement of study procedures (including the Data Management Plan and the Anonymisation Protocol). In the second of these meetings, two members of UKDA staff joined us to discuss key principles related to archiving and to answer questions about the practicalities and ethics of making deposits into national data repositories. Therefore, a considerable amount of group/individual labour was focussed on process and procedure, before the work of analysis could begin in earnest.

Our final Steering Group meeting focussed entirely on the emergent findings that the lead researchers had revealed during Stage 2 analysis. At this meeting, the group tested and explored these themes and considered next steps. We structured that meeting along the lines of what Tarrant (2016) has described as a 'data sharing workshop' to explore common themes between projects by bringing the data, the researchers involved in its original production, and its re-users, into conversation. We utilised the following key questions set by Franz (2013) to elicit colleagues' iterative responses to the emergent themes and patterns:

- What surprised you about the data?
- What was confirmed by the data that you already knew?
- What was missing in the data that you thought you would see?
- What other meanings do you see in the data that we haven't already discussed?
- What other comments do you have about the data?

This structure enabled us to discuss any overlaps, silences, surprises, recollections and differences, as well as exploring the social, political and emotional contexts of data production in much greater detail. As a group we discussed the challenges involved in returning to this material – including the challenge of 'hindsight' as a researcher, and being somewhat overwhelmed by just how much we had allowed ourselves to 'forget' - which many agreed was a job requirement, but also tended to come as a surprise when confronted with transcripts of our own interviews about which we could recollect nothing at all. We also discussed the emotional labour involved in returning to difficult data, and when making personal reflections about working in this field across the decades. Given the richness that was uncovered through our initial process of analysis, the possible directions that re-use could take us in could have overwhelmed us, but we found instead that this meeting was pivotal in helping to centre our thinking, contextualise the research, and consider what was both feasible and desirable in terms of how we might work with these data moving forward, both in the latter stages of analysis, as described above, but also beyond the life of this project. What had started as a feasibility project was very quickly opening up a considerable number of potential lines of enquiry, each of which were intensively considered and critiqued by members of the Steering Group. Not only had the group helped to confirm the feasibility of this approach through our culminating discussion in the data sharing workshop, it was also very clear that it was bound to bear fruit during a time that was characterised by uncertainty and considerable ambivalence among many social scientists working in our field of study. Ultimately, it was not possible to ensure that all data from all 12 studies could be fully anonymised and deposited into the UK Data Archive within the lifetime of this project, as had been originally hoped. However, Steering Group members reaffirmed their commitment to identifying the resources to enable anonymisation and archiving of these data (and more) over time, due to their burgeoning acknowledgement of the potential that they held for re-use, not only by members of this collaboration, but also for others.

Discussion
In preparing this article, we hoped to convey a number of the practical lessons that were learned throughout this feasibility study, as well as providing broader methodological reflections on the benefits of qualitative data re-use across a diverse range of source materials. A considerable strength of this approach has been the reflexivity it offers the practitioner in relation to aspects of knowledge production that are present in all forms of social research:

> "It is true that data are constructed by researchers. What is and is not relevant evidence, what it means, what should be taken as strong and what as weak evidence, and so on, depends upon the research focus, the data available, and the line of argument that develops in the course of the research. However, if we think of data as resources that we subsequently come to use as evidence, then there is also an important sense in which data exist independently of the research project." (Hammersley, 2010: 5.2)

While revisiting research texts (research notes, outputs, and transcripts) which capture the researchers' original descriptions of *what was there*, we continually found ourselves more enlightened about *what is here*, now, and what may yet come to pass. As Moore suggested, we found that "through recontextualization, the order of the data has been transformed" (2007: 2.3). At the same time, one of the potential limitations was our selectivity in tracing the historical and contemporary narratives of ARV re-purposing, directing us toward the identification of particular projects and sections of transcripts that we examined with great care. In addition, we did not (and

could not) include all qualitative datasets on HIV that were collected in the UK in this period, and as detailed above, some in our group decided to opt out from sharing their data. As such, we acknowledge that this approach involves a range of highly selective processes which are acutely dependent upon the curation of specific data sources. It is our view that this selectivity is ultimately outweighed by the benefits generated through re-use.

Where those undertaking analysis had not been present during data construction (or if it happened some time ago), we needed to be considerably mindful about any assumptions (or gaps in memory) that could obscure the social context of data production. As others have noted (Haynes and Jones, 2012), we too came to the conclusion that our temporal and attitudinal distances from the data we were re-using ultimately demonstrated one of the strengths of this approach, because it encouraged us to interrogate and re-examine assumptions we may have held about what the data 'should' contain, or how original researchers' questions 'should' have been framed. Some data had emerged from the shadows in ways that had taken us off guard and reminded us of what it was to work in different moments and places of an unfolding epidemic, offering snapshots of both mundane and challenging aspects of life in close proximity to this disease. We also found that this process brought data and researchers together across time and place in ways that enabled us to hear silences, which in turn had the potential to open up their own interpretive avenues (Irwin et al., 2012). We also became alert to the 'given' aspects of data from the past which cannot be expanded or extended. Quite unlike researchers who may take a grounded theory approach, through iterative analysis, which can lead to the immediate amending of topic guides or the expansion of a sampling frame, when re-using data we simply need to acknowledge that such existing gaps can only be noted with interest, which can at times place limits on the bounds of exploration.

Ultimately, the materials we prepared for re-use contained considerable internal vitality and coherence, while at the same time, they were also reconstituted, perhaps 're-coloured' under our contemporary spotlights, with new questions helping to uncover new hues, depths and variations when regarded within the context of changing imperatives of ARV use in the HIV global landscape. In turn, some of the most productive moments in this process where when data fragments functioned like mirrors that offered flashes of insight into the future. It was at these moments in particular, that a flurry of further questions tended to emerge, and it is this characteristic of qualitative data re-use that we find the most exciting and generative.

Our foray in qualitative data-reuse also held its own highly personalised and sometimes quite emotional surprises. Certainly the extent to which this was experienced depends on the duration and depth of individual researchers' involvement in the field, and the shape of their personal/community/activist ties within the HIV sector. As such, some members of our collaboration were themselves taken aback when this process triggered a series of recollections about undertaking data-collection right at the moment when effective anti-retroviral medications had proved to be efficacious, but not soon enough to save colleagues, partners, co-researchers and research participants who had already died. It also became clear from reviewing the data that some experiences had become transplanted by the grand-narratives that succeeded them. For instance, while many of us who had been working in the field since the late 1990s had a tendency to reflect back on the emergence of ARVs as bringing about a sudden 'resurrection' impact on people with HIV and AIDS at that time, the data themselves revealed a different and much more mundane truth than some of use recalled. In fact, we could see through the transcripts from that period that frequently, people's recovery journeys on new treatments were often mediated, painful, long-term, uncertain and dangerous. It was therefore a striking realisation that the *post hoc* cultural narratives about the sudden change brought about by a pharmaceutical revolution in HIV had served to obscure our own personal and professional experiences of the period, and that despite remaining embedded in the HIV sector, many of us had glossed over just how complex life with HIV continued/s to be after 1996,

and how diverse these experiences were depending on the material and geographic locations of people living with HIV during this period. What these data served to remind us of are the many messy, halting, ambivalent experiences and accounts of illness and of living with HIV. They also remind us that it is often less a question of responding to, adhering to or even accepting treatment advances or pharmaceutical prevention technologies, but more a question of growing with and around them as they slowly emerge, a theme that has been singularly emphasised in Squire's (2013) work.

It is this realisation that enables us to consider the power of data re-use within the contemporary context where these same pharmaceuticals are being positioned as the mechanism for eradication of HIV in the next decade. For instance, this process helped us to powerfully re-acknowledge that often, the experience of technological change is slow, rather than fast, and this is even more true for those who live on the margins, as is the case for so many people living with and at risk of acquiring HIV. Re-use of these particular data have helped to re-ground members of our collaborative group in experiences of adaptation to ARVs that happen over time, along with the accommodation of new realities, which encourages us to question the temptation to look back, or indeed forward, and impose a sense of 'suddenness' on developments that have been and will be experienced by individuals through an often slower pace of change that is entirely reliant upon social, political and economic context. We anticipate that similar re-discoveries will be likely for those who re-use complex and long-range sets of qualitative data covering other fields of social enquiry.

In practical terms, this project helped us to clarify and reaffirm a range of strong rationales for data re-use. Firstly, it helps to avoid the systematic over-researching of small and potentially vulnerable groups of people. The findings above also demonstrate that qualitative data re-use helps to make better use of scant resources in environments where financial constraints have led to considerably increased competition for funded empirical data-collection. Although anonymisation procedures, data analysis and write up still demand resource, it consumes far less time, energy and expenditure than the collection of new data. The qualitative data re-use practice gap identified by Bishop and Kuula-Lummi (2017) is one that needs to be remedied, given that research funding councils in the UK and elsewhere have started to request that a case is made for the collection of new empirical materials rather than re-using existing data – encouraging re-use to be increasingly regarded as a sustainable and efficient practice when so much existing data is under-utilised. At the same time, the development of these infrastructures requires diverse engagement across the social science spectrum in order to encourage broad critical input on the ethical issues that such developments will inevitably raise. As Leonelli (2016) warns, the Open Data movement is deeply embedded in the logics and structures of globalised market imperatives. Therefore, inasmuch as data re-use may be a mechanism to support sustainability, we must all be vigilant about power dynamics in order to ensure that data infrastructures are not simply left to the dicates of powerful market forces.

Conclusion
Despite a flourishing literature about the re-use of qualitative data, social scientists lack a wide array of sources that offer insights into the practicalities of preparing data for re-use, and describing the work that is entailed in discovering resonances and dissonances among considerably diverse and contrasting datasets. In sharing our methodological inspirations, rationales and day-to-day practices, we seek to inspire other qualitative researchers to consider the value of bringing contrasting bodies of data 'into conversation' (Tarrant, 2016). When working with datasets that have not yet been readied for sharing, researchers will do well to ensure the allocation of considerable resource for anonymisation; indeed, a key learning from this project has been to ensure that in future we will plan ahead for anonymisation and archiving before a study is put to rest. Ultimately, within our collaboration, we found that the generation of the anonymisation protocol alerted all potential data-sharers to the realities of Open Data in a qualitative social science context, and this too was a key

moment for learning. The use of 'test-pits' and 'deeper digs' within context were extremely fruitful endeavours that enabled the clear identification of emergent themes, silences, and forgotten experiences (Davidson et al., 2019). It is clear that in undertaking this feasibility study, we have only just scratched the surface of what is possible with these 12 datasets, and in time the goal is for them to all be made fully available for further research through the UK Data Archive.

In the meantime, we have established that drawing together discrete sets of data, collected by different teams, for different reasons and at contrasting temporal and geographic points is a feasible and considerably rewarding undertaking. It was an ambitious and rare undertaking which has not only had tangible results in terms of theory-building and methodological development, but it has also considerably affected the researchers involved – encouraging us to be tuned in to the history of the present and the future in ways that would not have been possible had we not embarked on this project. Furthermore, we believe that qualitative data re-use strategies and processes will go on to be further developed, refined and built upon by social scientists and interdisciplinary researchers. In particular, we hope to see those working in diverse health fields that are characterised by technological change consider qualitative data re-use more often as a means of exploring processes of biomedicalisation. In drawing upon Leonelli's work on biological 'data journeys' (2016), we also hope that this article opens up space for further exchange between those working on data re-use from a wide range of starting points, including: those interested in the philosophy of science, biologists, medics, medical ethicists, science and technology scholars as well as the full array of social scientists.

References

Bell S and Figert AE (2015) Moving sideways and forging ahead: reimagining "-izations" in the twenty-first century. In Bell S, Figert AE (eds) *Reimagining (Bio)Medicalization, Pharmaceuticals and Genetics: old critiques and new engagements*. London: Routledge, pp.19-40.

Bishop L and Kuula-Lummi A (2017) Revisiting Qualitative Data Reuse: a decade on. *Sage Open* 7(1).

Clarke AE, Shim JK, Mamo L, Fosket JR and Fishman JR et al. (2003) Biomedicalization: technoscientific transformations of health, illness and US biomedicine. *American Sociological Review* 68:161-94.

Corti L, Van den Eynden V, Bishop L and Woollard M (2014) *Managing and sharing research data: a guide to good practice*. London: Sage.

Davidson E, Edwards R, Jamieson L, Weller S (2019) Big data, qualitative style: a breadth-and-depth method for working with large amounts of secondary qualitative data. 53: 363-376.

Davis M (2010) 'Antiretroviral treatment and HIV Prevention: perspectives from qualitative research with gay men with HIV in the UK' in *HIV Treatment and Prevention Technologies in International Perspective*. (eds. Davis M and Squire C). Houndmills, Basingstoke: Palgrave Macmillan

Fassin D (2007) *When Bodies Remember: experience and politics of AIDS in South Africa*. Berkeley: University of California Press.

Franz N (2013) The data party: involving stakeholders in meaningful data analysis. *Journal of Extension* 51(1).

Hammersley M (2010) Can we re-use qualitative data via secondary analysis? Note on some terminological and substantive issues. *Sociological Research Online* 15(1): 1-7. http://www.socresonline.org.uk/15/1/5.html

Haynes J and Jones D (2012) A tale of two analyses: the use of archived qualitative data. *Sociological Research Online*. 17(2): 1-9.

Heaton J (2004) *Reworking Qualitative Data: the possibility of secondary analysis*. London: Sage.

Heaton J (2008) Secondary analysis of qualitative data: an overview. *Historical Social Research*. 33(3): 33-45.

Irwin S, Bornat J and Winterton M (2012) Timescapes Secondary Analysis: Comparison, Context and Working Across Data Sets. *Qualitative Research* 12(1): 66–80.

Keogh P and Dodds C (2015) Pharmaceutical HIV prevention technologies in the UK: s ix domains for social science research. *AIDS Care* (27): 796-803.

Kippax S and Stephenson N (2016) *Socialising the Biomedical Turn in HIV Prevention*. Cambridge: Anthem Press.

Leonelli S (2016) *Data-centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

Mason J (2007) 'Re-Using' Qualitative Data: on the Merits of an Investigative Epistemology *Sociological Research Online* 12(3): 1-4.

Moore N (2007) Re-Using Qualitative Data? *Sociological Research Online* 12(3): 1-13.

Mykhalovskiy E, McCoy L, Bresalier M (2004) Compliance/Adherence, HIV and the critique of medical power. *Social Theory and Health* 2(4): 315-340.

Nguyen VK (2010) *The Republic of Therapy: triage and sovereignty in West Africa's time of AIDS*. Durham and London: Duke University Press.

Paparini S and Rhodes T (2016) The biopolitics of engagement and the HIV cascade of care: a synthesis of the literature on patient citizenship and antiretroviral therapy. *Critical Public Health* 26(5):501-517.

Persson A, Newman C, Mao L, de Wit J (2016) On the Margins of Pharmaceutical Citizenship: Not Taking HIV Medication in the "Treatment Revolution" Era *Medical Anthropology Quarterly* 30:359-377.

Ritchie J and Lewis J (2003) *Qualitative research practice: a guide for social science students and researchers*. London: Sage.

Rose N (2007) *The Politics of Life Itself: biomedicine, power and subjectivity in the twenty-first century*. Princeton, New Jersey: Princeton University Press.

Slavnic Z (2013) Towards Qualitative Data Preservation and Re-Use—Policy Trends and Academic Controversies in UK and Sweden. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* 14(2) Art 10.

Squire C (2013) *Living with HIV and ARVs: three-letter lives*. Basingstoke: Palgrave-Macmillan.

Tarrant A (2016) *Qualitative secondary analysis and research design: reflection on a methodological framework for data re-use*. Timescapes Working Paper, University of Leeds, UK. http://hjh711zvmtv455cla15ikb71.wpengine.netdna-cdn.com/files/2015/07/Working-Paper-QSA.pdf

UNAIDS (2014) *Fast Track: ending the AIDS epidemic by 2030*. Report, Geneva: UNAIDS. http://www.unaids.org/en/resources/documents/2014/JC2686_WAD2014report

Walters P (2009) Qualitative archiving: engaging with epistemological misgivings. *Australian Journal of Social Issues* 44(3): 309-320.

Young I, Davis M, Flowers P & Mcdaid LM (2019) Navigating HIV citizenship: identities, risks and biological citizenship in the treatment as prevention era. *Health, Risk & Society. 1-16.* https://doi.org/10.1080/13698575.2019.1572869

Zimbra D, Abbasi A and Chen H (2010) A cyber-archaeology approach to social movement research: framework and case study. *Journal of Computer Mediated Communication* 16(1): 48-70.

Supplementary File

From Treatment Possibilities to Treatment as Prevention: HIV and biomedicalisation in UK qualitative datasets

Protocol to safeguard disclosure via datasharing

There are three considerations through which research data anonymity is to be protected. Each of them are to be considered in relation to one another (like a graphic equaliser). Each PI will be making their own determination on their datasets. There one consideration may be low, others may need to be raised. These are:

> **Consent** – to what degree did participants want to/expect to be identified as a result of their participation

> **De-Identification** – the post-hoc removal of information from transcripts

> **Regulating Access to Data** – additional access controls that can be set on each individual dataset within the Archive

This document is not designed to manage consent issues. They are to be considered on a study by study basis. What this document does cover is the matter of de-identifying and regulating access to data. Where a decision is taken to remove/replace data, the following core principles should be applied. We do not advise that any existing work to anonymise data should be reversed, although a note in study meta-data on this would be helpful.

Core Principle for De-Identification

With the exception of direct identifiers (see below), no data will be anonymised/removed unless you consider it to be disclosive of an individual's identity.

Our duty of care is usually to the individual, not to any organisations or entities that they may have named. A key exception to this is in service audit research with service providers, where we made a promise to participants that their organisation could not be identified.

1. **Consent**

In most cases, PIs have been able to identify whether research participants were asked for (and granted) consent for data to be shared with other qualified researchers outside of the immediate study team. To date, we are clear that the contributing studies from UEL and University of Glasgow have sought this consent from participants. If data depositing is to take place with these datasets,

some remaining work would need to be done to determine if any participants declined to consent to data-sharing, and those cases would need to be removed prior to ingestion.

In many other cases, including Sigma's datasets consent for datasharing was not sought. On this basis, it will be necessary for the PIs to consider strengthening provisions in de-identification of data and access restrictions in order to compensate.

2. De-identification procedures

DIRECT IDENTIFIERS

All personal names are to be considered disclosive of identity. Remove and replace with [name], and do not use pseudonyms.

Wherever possible, include information about the named person's relation to the participant, such as friend, family member, doctor, fuckbuddy.

> Example:
> 'My mate Pete' becomes 'My mate [friend's name]', *and hereafter if Pete were mentioned again, he would be changed to* [friend's name].
> *If more than one friend appears in transcript, then these all become numbered, ie [friend's name 1] [friends' name 2] etc.*
>
> 'My cousin, Agnes' becomes 'My cousin, [name]'. *If Agnes is mentioned again, consider adding the relationship that a later segment would be changed thus:*
> 'So then Agnes called me later that week' to 'So then [cousin's name] called me later that week'.

If interviewer or participant uses the participant's name in the transcript, it should be replaced with [participant].

Residential or workplace addresses, postcodes and telephone numbers are all to be removed, with descriptors such as, [home address] etc.

INDIRECT IDENTIFIERS

Not all indirect identifiers will absolutely need to be removed. It depends on the extent to which they could be used to identify someone in the context of all the other information in the interview.

Geographic location

We are not stipulating a level of detail of location which 'should' be removed, as this should be considered in light of the core principle.

> If you must remove a city name, replace it with: [city]
> Same for [country], [region], [town], [village], [address].

Avoid any use of inaccurate nouns, such as 'place' or 'location', as the descriptors above help to identify the type of location with more accuracy.

Consider that an exact geographical location may be replaced with a meaningful descriptive term that typifies the location [southern part of town, near the local river, a moorland farm, his native village].

Organisations (employer, service, business, political or social organisation, university, hospital)

We are not stipulating a level of detail of organisation which 'should' be removed, as this should be considered in light of the core principle, and the extent to which it is disclosive when considered in relation other information in the interview.

'I met him in Central Station'   to 'I met him in [name of gay bar]' - As this is a large central London gay business.

'When I was working at The Bell'. This could be changed to [name of gay bar], but he was working in the small town of Stroud which has only one gay bar, The Bell. If we are to keep the name of the town, Stroud, in the data, to name that town AND and describe its one gay bar, even if anonymised, could be highly disclosive, especially as the participant worked there. The options here may include:
- Remove the town name from entire transcript, depending on other disclosures
- Call The Bell his [workplace], but not describe what it is
- Which is more important, knowing he worked in a gay bar, or he lived in Stroud, or neither? Or, would you decide to restrict access to this transcript on this basis, particularly if much more sensitive data was described.

Situations, activities and contexts

We are not stipulating a level of detail of circumstance which 'should' be removed, as this should be considered in light of the core principle, and the extent to which it is disclosive when considered in relation other information in the interview. For example, when a person living in a small town describes their public 'outing' as a person living with HIV who is also seeking asylum, care would need to be taken to determine which piece of information is most relevant to being removed that would sacrifice the least amount of data, while protecting the identity of the individual. In such a case, removing the name of the small town might be the best solution.

Another example of such a situation might be the interview data given by a haemophiliac who acquired HIV in the late 1980s, going on to lead ACT UP London. In such a case, removing any of these 4 pieces of data (health status x2, era, activism) could do great damage to the dataset. In such a case, perhaps it is best to increase control over access rather than anonymise.

For further information see: http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation

3. Regulating access to data

The UK Data Archive affords gradations of data access in cases where sensitivity and potential disclosiveness of data are increased. As in the example given directly above, it may be regarded as disruptive of data integrity to remove information about an activist, his HIV and its acquisition and when that happened, even if all of that information taken together could identify him. In such cases,

it may then be required that access to that particular transcript or group of transcripts is more heavily restricted.

The following access restrictions are afforded by the Data Archive.

- needing specific authorisation from the data owner to access data
- placing confidential data under embargo for a given period of time until confidentiality is no longer pertinent
- providing access to approved researchers only
- providing secure access to data by enabling remote analysis of confidential data but excluding the ability to download data

Where further foreseeable harm could come to a research participant due to a specific disclosure, such as the revelation of serious criminal activity, then simply restricting access to data would not afford them enough protection. In that sort of case, then strong efforts to anonymise all indirect identifiers should be considered. In serious cases, such a transcript may be considered ineligible for archiving.

For more information see: http://www.data-archive.ac.uk/create-manage/consent-ethics/access-control

Blog on The 5 Safes

http://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/