HHS Public Access

Author manuscript

Qual Health Res. Author manuscript; available in PMC 2020 July 27.

Published in final edited form as:

Qual Health Res. 2018 August ; 28(10): 1629–1639. doi:10.1177/1049732318759658.

Enabling analysis of big, thick, long and wide data: Data management for the analysis of a large longitudinal and cross-national narrative dataset

Kate Winskell[a], **Robyn Singleton**[a], **Gaelle Sabben**[a]

Hubert Department of Global Health, Rollins School of Public Health, 1518 Clifton Road, NE, Atlanta GA 30322

Abstract

Distinctive longitudinal narrative data, collected during a critical 18-year period in the history of the HIV epidemic, offer a unique opportunity to examine how young Africans are making sense of evolving developments in HIV prevention and treatment. More than 200,000 young people from across sub-Saharan Africa took part in HIV-themed scriptwriting contests held at 8 discreet time points between 1997 and 2014, creating over 75,000 narratives. This article describes the data reduction and management strategies developed for our cross-national and longitudinal study of these qualitative data. The study aims to inform HIV communication practice by identifying cultural meanings and contextual factors that inform sexual behaviors and social practices, and also to help increase understanding of processes of sociocultural change. We describe our sampling strategies and our triangulating methodologies, combining in-depth narrative analysis, thematic qualitative analysis, and quantitative analysis, which are designed to enable systematic comparison without sacrificing ethnographic richness.

Introduction

The increasing availability of large digital data sets has focused attention on the opportunities and challenges of secondary analysis of “Big Data” and on the need for appropriate data management and analysis tools. Some have argued for the need for greater emphasis on thick data (drawing on Geertz (1973)), especially stories, as a potential counterbalance or integrative complement to quantitative metrics derived from Big Data algorithms (Wang, 2013). This article addresses data management for the analysis of a qualitative archive that, while not machine-readable, is large, and is also ethnographically rich, longitudinal, and geographically diverse: that is to say, one that comprises data that is big, thick, long and wide.

The Global Dialogues contests invite young people under the age of 25 to come up with ideas for short films about HIV and related topics. More than 200,000 young people from across sub-Saharan Africa took part in contests held at 8 discreet time points between 1997 and 2014, creating an archive of over 75,000 narratives. Writing on the value of archived

Correspondence details: Kate Winskell, swinske@emory.edu.

[a]Hubert Department of Global Health, Rollins School of Public Health, Emory University, USA

qualitative datasets in research on socio-cultural change, sociologist Michael Savage asks, “How do we deal with this amount of data which by virtue of its qualitative nature does not lend itself to easy summary or codification?” (Savage, 2011, p. 175). This article seeks to answer this question. We describe data management strategies for a study that seeks, without sacrificing ethnographic richness, to allow systematic cross-national and longitudinal comparison of a huge narrative dataset.

Savage differentiates between the – albeit overly simplified – methodological approaches of the historian (which he describes as qualitative, inductive, particularizing, messy) and the social scientist (which he describes as quantitative, deductive, generalizing, standardized), pointing to a possible third path in archived qualitative research. To these disciplinary categories, our own study adds that of literary scholar, as our data are creative narratives submitted to a scriptwriting competition, the particularity of which must be respected.

In addition, our study situates itself explicitly within the field of public health. Our data extend from 1997, when antiretroviral therapy (ART) was transforming life with HIV in the West, through 2005 when ART roll-out was becoming a reality in sub-Saharan Africa, to 2014 when ART coverage had reached 80% in at least ten African countries. We seek to assess to what extent opportunities afforded by the unprecedented scientific, technical and programmatic advances of the past decade to reframe HIV, reduce related stigma, and improve access to services are being integrated into young people’s sense-making. The primary goal is to contribute to efforts to improve HIV prevention and treatment and stigma-reduction outcomes among young Africans by identifying communication needs and best practices at country levels and disseminating country-specific recommendations for improved messaging.

Through the narratives’ unique combination of ethnographic richness and amenability to systematic cross-national, longitudinal comparison, our research is well-positioned to generate findings of importance for HIV policy and for programmatic practice; it also has potential to shed light on processes of sociocultural change. In addition, it is distinctive in the attention it pays to the direct application of findings in intervention practice. The data provide critical insight into youth sexual culture, decision-making, and sociocultural context. As narratives, they can also inspire culturally-grounded and engaging approaches to address the needs they identify.

This applied potential must also be reflected in our data management strategies for this vast and distinctive data source, which must satisfy historical, social scientific, literary, and applied public health needs. We share these strategies in the hope that they may be useful for those designing management approaches to support the analysis of Big Data qualitative sources, such as blogs (Germain, Harris, Mackay, & Maxwell, 2017), and also for more traditional large-scale qualitative datasets, namely oral history archives that may or may not be digitally searchable.

Contest Elicitation of Data

The young participants in the Global Dialogues contests are mobilized by non-governmental and community-based organizations and local, national and international media across sub-Saharan Africa. The winning ideas in each contest are selected – first at national, then at international level – by local juries and, following adaptation, transformed into short fiction films by leading African directors (Winskell & Enger, 2005). A leaflet, identical in all countries and available in several major languages, is used continent-wide to provide young people up to the age of 24 with instructions on how to participate in the contest, inviting them to come up with a creative idea for a short film. From 2005 through 2011, *Scenarios from Africa* contests invited participants to "help other people learn about HIV"; in 2013 and 14, the contest, under the name *Global Dialogues*, was framed in global terms, included a broader array of themes (sexuality, violence against women, alcohol and drugs, in addition to HIV), and encouraged participants to speak out and "participate in creating a better world". The shift in framing and elicitation did not result in noticeably different narratives on the theme of HIV; hence, we treat this data as comparable, all the while remaining mindful of any changes in emphasis that may have resulted. Scenarios were ineligible for inclusion in the study sample if they did not mention HIV.

The contest leaflet also provides a list of suggested topics, or story starters which participants can, if they choose, use as a starting point for their stories. The list is developed through a consensus survey with Global Dialogues team members (Winskell & Enger, 2005). These story starters have changed over time, becoming progressively less leading. For example, in 1997 they took the form of scenarios, such as "He/she wants to speak with his/her parents about HIV/AIDS and finds some clever ways of starting up a discussion." By 2014, these story-starters had been replaced by broad "Suggested Topics", such as "Taking or avoiding sexual risks", "Living with HIV", and "Sex and drugs or alcohol". Participants always have the option not to respond to a story starter and to write about any topic of their choosing.

The data included in our large sample are drawn from these "free choice" narratives, in accordance with the inclusion/exclusion criteria detailed below, with the exception of 2014, when submissions in response to the broad suggested topics related to sexuality and HIV were also eligible as they did not differ from the "free choice" narratives. With the caveats outlined above and to the extent possible in light of the programmatic realities of the scriptwriting competitions, the collection of this secondary data is consistent across sites and over time, allowing for meaningful systematic cross-cultural and longitudinal comparison of a large number of narratives. It is doubtful if data of this volume and scope could have been collected by any other means.

Contest participants were informed on the contest leaflet that their submissions may be made public and may be used for research purposes. The narratives sampled for this research project were de-identified prior to transcription and analysis. This study constitutes secondary analysis of existing data collected for programmatic purposes and was approved by Emory Unviversity Institutional Review Board.

Theoretical Framework: Social Representations and Narrative

Traditionally, the tools used to study and measure cultural phenomena have been surveys, interviews, focus group discussions, and ethnographic observation. In public health, in line with dominant theoretical frameworks, methodologies have tended to focus on data related to individual-level conscious cognitive processes, like attitudes and behavioral intentions. In their creative writing, young people draw on their own lived or imagined experience and on culturally determined sources of social understanding to create narratives imbued with context, meaning, and values. Although fictional stories are invented depictions of social fact, they are still culturally-determined social facts in and of themselves and are a source of insight into how people make sense of the world, and how they communicate those understandings to others in their cultural community (Rabinow, 1986).

Patterns across the narratives elucidate the distribution and evolution of young people's social representations (Moscovici, 1981) of HIV prevention and treatment options, of people living with HIV, and of gender. Social representations communicate norms and values in symbolic form. They act as both barriers to and facilitators of social and behavioral change – in the form of image-rich and emotionally-laden stereotypes or progressive cultural narratives – and are amenable to influence by communication efforts (Markova & Wilkie, 1987). Social representations are properties of social groups rather than individuals (Howarth, 2006); often preconscious, they tend to be more emotionally-charged and less susceptible to social desirability bias than cognitively-based attitudes.

Narratives have been identified as a particularly valuable and underused data source for the study of social representations (Jovchelovitch, 2002; Laszlo, 1997; Murray, 2002) and sense-making. One of the reasons we tell stories is "to 'make sense' of what we are encountering in the course of living…" (Bruner & Lucariello, 1989, p. 79). Narratives allow people to formulate and articulate the causes and consequences of human actions and, as such, underlie social knowledge (Bruner, 1990). They offer researchers the opportunity to explore normative issues in a way that approximates to the complexities with which they are surrounded in reality and thus provide access to ethnographically rich social representations.

With notable exceptions (Hirsch et al., 2009), few ethnographic studies have used similar data collection and analytical techniques across multiple cultural settings. The analysis of a large sample of narratives authored by young people from a diverse range of countries over two decades can add significantly to our understanding of the cross-cultural and longitudinal variation in cultural meanings and contextual factors that inform sexual behavior and social practice.

Just as there are multiple theories of narrative (Bruner, 1991; Labov & Waletzky, 1967), hailing from a range of disciplinary and theoretical perspectives, so there is a spectrum of approaches to narrative analysis, leading Mishler (1995) to comment on the state of near anarchy in the field. Quantitative/qualitative and inductive/deductive to varying degrees, these approaches range from the highly quantitative forms of content analysis through socio-linguistics to literary analysis (Czarniawska, 2004; Elliott, 2005; Franzosi, 1998; Oliver,

1998). Our study draws from – and triangulates between – a range of methods to develop a methodology tailored to the specific research objectives.

Study Settings and Sampling Strategies

The Global Dialogues archive includes stories from 80 countries, 49 of which are in sub-Saharan Africa. The sheer size and scope of this dataset and our resulting longitudinal sample precludes the possibility of analysis of every story. Moreover, as the sample is not representative, there would be no merit in doing so (Savage, 2011). We therefore developed a stratified random sampling strategy to generate a sample of approximately 2,000 narratives spanning the eight time points (1997, 2000, 2002, 2005, 2008, 2011, 2013, 2014).

Our longitudinal study builds on an NIH-funded cross-sectional study which analyzed narratives written in 2005. For that study, we selected six non-contiguous countries in which at least 500 stories had been contributed. The estimated adult HIV prevalence rates of those countries in 2005 were distributed along a near-exponential curve from 1 to 33%: Senegal (0.9%), Burkina Faso (2%), Nigeria (3.9%), Kenya (6.1%), Namibia (19.6%), and Swaziland (33.4%) (UNAIDS, 2006). We use these same countries for our longitudinal study. However, there was insufficient longitudinal data from Namibia to warrant its inclusion. Our study therefore focuses on the five remaining countries that have diverse epidemiological and sociocultural profiles. The countries also differ significantly in other important respects, including demographics, socioeconomics, urbanization, education and health.

We applied methodologies piloted and validated in our preliminary study of 2005 data to our longitudinal sample extending across eight time points between 1997 and 2014. Contest participants provide demographic information on the contest entry form, which is included in the contest leaflet. Every contest entry since 1997 is documented by Global Dialogues in a year- and country-specific Microsoft Excel file. A spreadsheet row is devoted to each scenario received and includes the demographic information provided by the young author on the contest entry form, including year, country, age, sex, location (capital, suburbs of the capital, another big city, medium-sized town, village), and whether the individual took part in the contest alone or as part of a team.

We sampled from all eight time points for the quantitative and originally from three time points (1997, 2005, and 2014) for the qualitative dimensions of the study. Preliminary analysis of quantitative attributes suggested that the 2008 narratives may have some distinctive characteristics in terms of message framing. We therefore decided to add this year to the time points we would analyze qualitatively.

Inclusion/Exclusion Criteria and Sampling

Only narratives about HIV/AIDS written in English or French by authors between the ages of 10 and 24 from the five study countries were eligible for inclusion in the sample. In order to ensure comparability across countries based on demographics of individual authors, narratives were ineligible if they were team-authored. As indicated, they were also ineligible if written in response to one of the thematic story-starters provided on the contest leaflet

between 1997 and 2011, but not if there were written in response to a broad suggested topic area 2014. Approximately one third of submissions were either non-text-based (e.g. pictures, video cassettes) and/or non-narrative (e.g. essays) and therefore eliminated from the study if sampled. A text was eligible for inclusion as long as it incorporated a story component. In some cases, this was preceded or followed by commentary from the narrator, however, the entire text was included in our analysis; for convenience, we use the term *narrative* to refer to it. In light of the size and cultural diversity of the Nigerian population, only those narratives from the Igbo-speaking Southeast were eligible. If more than one narrative by the same author was sampled, only the first text selected was eligible in order to maximize the social diversity of the sample. Any entry for which the age, gender or place of residence of the author could not be ascertained was not deemed eligible. Participants self-reporting current residence in “a small or medium-sized town” or “a village” were categorized as rural and the remainder as urban.

In order to maximize representation of participants across demographic strata and allow us to explore the distribution of social representations within subgroups, thereby reducing the effects of selection biases in the country samples, we stratified our data by sex, urban/rural location and age (10-14, 15-19, 20-24).

In the existing year-specific spreadsheets, we eliminated scenarios that did not comply with the inclusion criteria on grounds of age, team-authorship, or story starter inspiration. For each country and each year, we randomly selected up to 10 narratives for each of the twelve strata, having determined that a higher number of stories per stratum did not substantially increase the richness of the data. In line with the iterative nature of qualitative research, if we find that we are not reaching saturation across certain particularly relevant themes, for example, same-sex attraction, we will purposively resample for that particular topic. Using a random number generator, we systematically and sequentially selected narratives and assessed in each case whether the text fulfilled the remaining eligibility criteria (text-based, in English or French, narrative-based), clearly documenting reasons for ineligibility.

Not all strata were complete for each country (Figure 3). Because of the challenges of achieving consistency in the demarcation between urban and rural in countries at such divergent levels of socio-economic development, stratification by place of residence was designed to maximize geographical representation but was not intended to allow systematic comparison between urban and rural stories. Stories submitted to the competition in general did not match the urban-rural distribution of the country population. It was therefore decided to oversample some locales to increase the likelihood that 20 stories were selected for each age/sex stratum. There was no pattern to this over-sampling: in some countries urban areas were over-sampled and in some countries rural areas were over-sampled.

As we do not have data from Nigeria, Kenya and Swaziland until 2005 and as the overwhelming majority of country samples for specific time-points contain fewer than 120 scenarios, we obtained a final sample of 1,937 narratives. Each sampled narrative was given a unique identifier indicating country and year of origin, sex, age, and urban/rural location of author.

Data & Management Methods

The narratives range in length from a few lines to many pages. They are overwhelmingly handwritten, with a small proportion submitted as word-processed text. Narratives were labelled with a unique identifier that includes the contest country and year as well as author sex, age and residence; these identifiers facilitate both the sorting of narratives based on demographic information and also pattern identification across the narratives. Narratives were transcribed verbatim (complete with spelling and grammatical errors) in English or French before being entered into MAXQDA qualitative data analysis software (VERBI Software, 1989-2010).

Our qualitative analytical approach is situated at the intersection of grounded theory (Corbin & Strauss, 2008) and thematic narrative analysis (Riessman, 2008). In our preliminary study, we developed triangulating methodologies that are tailored to the unique data source, its scope, and our research questions. We combine three primary approaches:

1. analysis of quantifiable characteristics of the narratives;
2. a narrative-based approach, focusing on plot summary and thematic keywords; and
3. thematic qualitative data analysis, focusing on thematically-related text segments and memoing for emergent themes.

The approach was developed to enable cross-national and longitudinal comparison and has three main advantages: it grounds the analysis in three distinct, though intersecting, dimensions of the data; allows for triangulation; and facilitates the generation and validation of interpretive hypotheses. It is, however, important to stress that the narratives themselves, in their entirety, are our constant point of reference, providing a holistic perspective to counteract any fragmentation and decontextualization of the data resulting from other analytical approaches.

1. Quantifiable Characteristics of the Narratives

Given the number of narratives included in our sample, we found it helpful to quantify discrete components of each narrative in order to allow for identification of patterns. A list of 40 characteristics were identified iteratively via dialogue within the research team based on prior analysis of the 2005 sample and updated in line with biomedical and social developments across the time points. Quantifiable characteristics include author demographics, narrative subtopic (prevention, infection, post-infection, combination of these or other) and mode of transmission (via blood route, mother-to-child transmission, or sexual transmission), among others. Using a detailed protocol, complemented by periodic trainings, the quantifiable characteristics were double entered by research assistants into Qualtrics data analysis software (Qualtrics, Provo, UT). Data were downloaded in Excel files and, where inconsistencies were noted, consensus was reached either via dialogue or via adjudication by a third team member. The finalized Excel files were combined across countries and years and imported into SAS software, Version 9.4, for analysis. Using appropriate tests for significance (dependent on cell size), we will examine differences in quantitative indicators by country, year, age, gender, urban/rural location, and over time. For example, are HIV-

related deaths (including suicides) less likely to occur over time? Are narratives becoming less focused on infection? Are representations of testing becoming more frequent? Is there any change in the proportion of negative tests?

In addition, we imported the quantitative data into MAXQDA for all countries and years, with particular focus on the 1997, 2005, 2008 and 2014 files, which are also thematically coded. This allows us to easily isolate and compare qualitatively narratives which share certain quantitative attributes, for example, those with a male vs. female protagonist or in which an HIV death does or does not occur. For the thematically-coded time points, we also exported quantifiable data from MAXQDA (specifically dichotomous variables on the presence or absence of specific codes) into Excel to allow us to conduct the statistical analyses described above on a thematic subset of data. For example, does male-to-female transmission occur more often than female-to-male transmission in narratives that address the theme of infidelity? It is important to notes that none of these analyses stand alone. They do, however, help identify patterns in the data and lead to the generation of hypotheses that can be examined using qualitative components of our triangulating methodologies.

2. Narrative Summary and Keywords

Traditional qualitative techniques are often inadequate for the analysis of narrative texts, because they “fracture… texts in the service of interpretation and generalization” (Riessman, 1983, p. 3), by decontextualizing segments for the purpose of comparison and analysis. In order to address this limitation and respect the narrative specificity of the data, we decided to complement traditional thematic qualitative analysis techniques with a narrative oriented approach. As used in relation to life histories (Viney & Bousfield, 1991), narrative analysis characteristically involves breaking down a narrative into defining structural characteristics. The most central and complex component of narrative analysis is the core narrative (Mishler, 1986), which “represents the essential meaning of the story in terms of its informational content, its interpersonal impact and the language in which it has been told” (Viney & Bousfield, 1991, p. 759).

We prepared a one-paragraph summary or core narrative for each narrative (n=1,937), comprising the key elements of plot and message. These paragraphs were limited to 10-15 sentences, and summarized the plotline and key themes; those creating the summaries were instructed to withhold any interpretation when summarizing. Written instructions were provided, as well as examples of “good” and “bad” summaries. We originally attempted to break down the narrative into structural categories. However, it was challenging to apply these categories consistently to our sometimes rudimentary narratives, with their often undefined tone, and contradictory master plots. In addition, using these categories to structure the summary was unwieldy, counter-intuitive, and unnecessarily disruptive to the flow of meaning-making within the narrative. We therefore simplified our instructions to research assistants for the preparation of summaries. In repeated tests, we found a high level of consistency in the narrative summaries and consensus on what merited inclusion. It is important to note that these narrative summaries in no way replace the narratives themselves for analytic purpose, but are intended instead to provide an *aide-memoire*, which is valuable given the size of the sample. They are particularly useful in allowing us to easily situate a

segment of text in the context of its plotline and in facilitating identification of common narrative arcs across a thematically-linked body of texts, for example, those focusing on sexual abstinence.

The summaries were imported into MAXQDA for all narratives, where each was labelled with keywords to allow us to easily locate texts focusing on specific themes. For those time points that were not transcribed and thematically coded (those from 2000, 2002, 2011, and 2013), each narrative was scanned and stored so that it could be easily accessed by the research team if needed.

While it was impossible to eliminate all interpretative bias from the allocation of keywords, our methods sought to minimize it. For the longitudinal study, the list of keywords used in the preliminary study was updated and revised using the methodologies for codebook revision described above. Up to 6 keywords were double-entered and any discrepancies were resolved by means of dialogue and reviewed by a third research team member to ensure consistency.

The keywords function in combination with thematic coding, allowing us to isolate both individual text segments related to a specific theme, for example ART, and those narratives in which ART is a central theme. They are particularly important for themes that recur with frequency across the texts but that may vary dramatically in intensity. For example, HIV testing may be thematically central to a narrative which describes the experience of getting tested, or it may be thematically peripheral in the many narratives in which a positive HIV diagnosis provides the turning point for the plot; the keywords allow us to make this differentiation. In our analysis we always review both keyworded and thematically-coded text on a specific theme to compensate for any biases in the allocation of keywords. Once again, these analyses do not stand alone, but interact with other components of our triangulating methodologies to generate hypotheses for further examination.

3. Thematic Codes

The codebook of descriptive codes (Miles & Huberman, 1994) used in our preliminary study of 2005 data was developed via an iterative team-based process and drew on the recommendations of African colleagues who read the narratives when selecting the winning contest entries (Winskell & Enger, 2009). Codes were both deductive, in line with the specific research questions, and inductive, in response to themes emerging from the data.

We provisionally revised and updated the 2005 codebook by reviewing comments and recommendations from the 2014 contest juries in the five countries in order to identify thematic priorities not addressed in the 2005 codebook. This allowed us to incorporate into codebook revisions the perspectives of a broad range of cultural insiders who were well-versed in HIV prevention and treatment at the community and national levels. We then randomly selected an initial 20 narratives from 2014, five from each country, to read and extensively memo with the purpose of identifying emerging themes and potential codebook revisions. These emerging, inductive themes were identified based on recurrence and on similarities and differences noted across the texts (Ryan & Bernard, 2003).

Our research team comprised a US-based team and African consultants from each of the five study countries. Using Skype group calls, we discussed our findings and recommendations with the African team members until we arrived at a consensus for recommended changes to the codebook. Following revisions resulting from the steps described above, the codebook was applied to a second sample of 20 narratives submitted in 2014. This coding was compared and discrepancies discussed with the purpose of identifying revisions to code definitions, inclusion/exclusion criteria, etc. This final step formed part of our training of research team members. Codebook refinement was thus the product of an extensive process of discussion and consensus-building in which provisional codes were applied to the data and iteratively refined (MacQueen, McLellan-Lemal, Bartholow, & Milstein, 2008).

The codebook comprises 54 codes and includes a detailed description of each code, inclusion and exclusion criteria, examples of the code in use, and specific guidance for coders (e.g. “things to look for specifically” and “vital questions to answer”) (MacQueen et al., 2008). It is organized architecturally, with each code falling under one of five primary headings: values/moral/message; inequalities & vulnerabilities; infection or post-infection; prevention/risk; and interpersonal interactions. In addition to the codes, research members memoed on both deductively-identified themes not adequately captured by codes, such as “social support” and “gender,” and emerging themes identified inductively (Corbin & Strauss, 2008). Two research members reviewed each narrative, and any discrepancies in coding or memos were resolved via dialogue. Questions about cultural context were referred to our African partners for insight, with a view to ensuring the validity and the ongoing applied relevance of the data.

While this initial coding allowed us to easily isolate text segments on specific general themes, such as “condoms” or “unmarried romantic relationships and values”, further coding was needed for in-depth analysis. We therefore develop fine or second-level codes for the themes we are analyzing in depth. For example, representations of sexual violence in the narratives encompass a range of experiences from coercion to rape. The descriptive codebook included codes such as “sexual violence,” “power differential” – referring to the coercion facilitated via economic and educational hierarchies – and “partner pressure.” Within the representations of sexual coercion and violence are many fine-grained themes such as the disclosure of assaults, parental awareness of risk and the social support (or lack thereof) in response to the sexual coercion or violence. We identify these fine codes inductively through the data and deductively via the literature, and apply these to the theme-specific data, memoing extensively throughout.

The current article addresses data management to facilitate analysis using thematic and narrative approaches. We will present our methodologies for contextual interpretation and analysis in a future publication. Suffice it to say, in our analytical approaches, we triangulate between the methodological approaches described above, for example, using the narrative and thematic analyses to illuminate the quantitative data, and vice versa. Our qualitative analyses also draw on code frequencies and intersections. For example, descriptive quantitative analysis of the distribution of thematic codes across the sample also allows us to broadly delineate thematic priorities across countries, time points and demographic characteristics. This is important in light of the size and scope of our data and our focus on

comparing across countries. We have found that generating histograms comparing code frequencies and intersections facilitates the identification of patterns and generation of hypotheses.

Discussion

Our 2005 study provided several cross-national cross-sectional insights into youth sense-making (Winskell, Brown, Patterson, Burkot, & Mbakwem, 2013) round subjects including condoms (Winskell, Obyerodhyambo, & Stephenson, 2011), stigma (Winskell, Hill, & Obyerodhyambo, 2011), abstinence (Winskell, Beres, Hill, Mbakwem, & Obyerodhyambo, 2011), and testing (Beres, Winskell, Neri, Mbakwem, & Obyerodhyambo, 2013). Cross-national comparison with Demographic and Health Survey (DHS) data (Winskell, Hill, et al., 2011; Winskell, Obyerodhyambo, et al., 2011), ethnographic studies (Winskell et al., 2013), and existing literature on a range of themes provided validation of our findings. The qualitative/quantitative balance of our analytical approaches has shifted in line with our research objectives. For example, we drew on quantitative and qualitative analyses to present a cross-national overview of symbolic stigma across our sample of almost 600 narratives (Winskell, Hill, et al., 2011); another article, in contrast, highlighted the value of the creative narratives to our understanding of stigma by presenting a more literary, fine-grained narrative analysis of just three texts (Winskell et al., 2015). We are currently employing the methods described above in longitudinal analyses across a range of HIV-related themes. The use of quantification and visualization techniques to support the recognition of patterns and generation of hypotheses will become increasingly important.

Research on Social Representations has employed methods ranging from the statistical to the ethnographic (Flick, Foster, & Caillaud, 2015). However, few studies have either sought to integrate qualitative and quantitative approaches or attempted to track changes in social representations over time. If we are to increase our potential to gain insights into processes of sociocultural change, we need to transcend methodological binaries without sacrificing either systematic rigor or ethnographic insight.

Scholars from a range of fields have explored innovative ways to do this. The field of oral history has long struggled with the challenges of managing complex, “messy” data, primarily relying on keyword indexes attached to summaries, not unlike our narrative summaries and keywords approach. Certain projects, such as the Millennium Memory Bank project, coupled data content management with data collection – a strategy becoming increasingly easy for oral history as digital, including mobile, technologies evolve.

Approaches such as those facilitated by Sensemaker software blend quantitative approaches that allow for statistical and visual analyses, while still maintaining access to the original narrative data. Sensemaker is an approach in which micronarratives – or anecdotes – are communicated and then interpreted by the person telling the narrative. This interpretation occurs through the use of *signifiers* or a *semiconstrained index set* that allows researchers to delineate the boundaries within which participants may interpret their micronarratives, while still allowing flexibility for emergent signification if the Sensemaker users desire (Snowden, 2011). Sensemaker is able to quantify the signifiers and then visualize the data using

visualization techniques such as fitness landscapes, an approach originally developed in biology (Wright, 1932). David Snowden, the architect of Sensemaker, argues that "visualization and statistical instruments allow patterns to be detected in the metadata, and once a pattern is detected, the ability to go to the supporting narrative, immediately and without interpretation, enables more effective decision making with respect to that pattern" (Snowden, 2011 232).

In its combination of narrative data, quantification, visualization and qualitative methodologies, Global Dialogues has some parallels with the Sensemaker approach. However, as the Global Dialogues narratives are contributed by young people as young as 5 years old and are contributed as part of a public health program rather than for research purposes, self-indexing of stories is not feasible. Retroactive indexing of the narratives as secondary data by the research team nonetheless has the advantage of ensuring that it is closely aligned with current research priorities and with our study's applied public health objectives.

Strengths and weaknesses

While our data management strategies to support our analytical methodologies are well suited to our research aims, the study and strategies underlying it have both strengths and weaknesses. Some scholars have turned to strictly quantitative approaches to analyze narrative data, using relational database management systems and semantic coding schemas that combine traditional content and linguistic analyses (Franzosi, 2010); ours, however, is a resolutely qualitative study of narrative data. This lack of automatization means that it is labor-intensive.

As contest participants self-select, the sample is not constructed to be representative of the youth population of the five countries in respect to ethnicity, religion, literacy, access to television, etc. Depending on the context, contest participants may be better educated, and more knowledgeable and motivated about HIV than the general youth population. However, as the product of the contest mechanism, these biases are likely to be consistent across the five countries and over time, hence the samples from different countries at different time points, though not representative, are comparable for our purposes. While our stratified random sampling approaches cannot retroactively influence the demographics of those who choose to participate in the competition, it – and the sheer size of our sample – can minimize bias. For example, the inclusion of urban/rural residence in our random stratified sampling process goes some way to mitigating socio-economic bias in the sample, ensuring that the sample is not solely composed of more educated urban residents. We account for these limitations in interpretation of our findings.

It is important to contextualize these limitations within those of other study designs. Most qualitative studies are conducted in a small number of often highly circumscribed settings and without a clear point of comparison outside the study site; their findings are nonetheless extremely valuable and are often used to inform nation-wide programs. The majority of qualitative studies address adult rather than youth populations and few studies address children aged 10 and up as a future risk group. Operating at national and cross-national level with a much larger sample size than the vast majority of qualitative studies, our research

project has the potential both to inform national-level HIV communication strategies and to benefit from cross-national comparability. In addition, in light of the impact of cross-national forces such as the President's Emergency Plan for AIDS Relief, evangelical Christian movements, and ideologies of companionate marriage, there is need to situate young people's social representations within a broad sub-Saharan context and systematically study their evolution over time. Our large qualitative data source with its rich ethnographic detail is distinctively positioned to do this.

We follow Farmer and Good (1991) in acknowledging the potential role that performative and rhetorical considerations may be playing in the narrative-based representations in our study: the young authors' motivation to tell what they consider to be a good story – and thereby win the scriptwriting contest – may be influencing the ways in which they represent HIV and related phenomena. Our preliminary studies demonstrate, however, that this does not represent an impediment to our study aims. On the contrary, it reinforces the proposed study's applied utility. The scriptwriting process offers young people an opportunity to create their own narratives about HIV for widespread dissemination, thereby situating HIV within their own cultural and moral logic (Watkins & Swidler, 2009). This has the advantage of allowing us to understand how HIV is constructed in the collective lay imagination, while at the same time permitting identification of communication needs – ranging from the cognitive (misconceptions and information gaps) to the ideological (stigmatizing cultural narratives that blame specific populations for spreading the disease) – and providing youth perspectives to inform education and communication efforts.

The insights provided by consultants in sub-Saharan Africa with extensive programmatic experience and stakeholder networks through ongoing communication and annual meetings help ensure the contextual sensitivity of our analysis. In addition to increasing validity, they also help ensure that our findings have direct relevance to programmatic practice. Despite the level of interpretation required in the analysis of our narrative data, we have sought to ensure a high level of consistency through our dialogue-based instruments development and ongoing training. In addition, the level of triangulation between the different methodological approaches helps to compensate for biases resulting from any one methodological approach.

Conclusion

With increasing digitization, large qualitative data sources are becoming increasingly common and accessible. As interest in various forms of storytelling increases, there is a need for theoretically-grounded and methodologically rigorous approaches to analyzing large qualitative datasets across space and time. Our study seeks to integrate seemingly divergent approaches with a view to shedding light on young Africans' changing representations of HIV over time and place. The data management strategies we have developed to support our study will allow us to easily identify patterns across countries and time-points, while also allowing us to conduct in-depth analysis of thematically-linked groups of texts. We will therefore be able to systematically compare over time contextualized social representations of HIV in countries with different epidemiological, socio-cultural, programmatic, and policy histories without sacrificing ethnographic richness. We have shared our own research priorities, our theoretical framework, our sampling approaches using systematic inclusion

and exclusion criteria and stratified randomization, and our triangulating methodologies tailored to the specificities of our narrative data. While recognizing that our dataset is highly distinctive, we hope that the approaches we have developed to facilitate this integrative analysis may be helpful to others seeking to manage and analyze large qualitative data sources.

Acknowledgments

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R01HD085877. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported by the Emory Center for AIDS Research (P30 AI050409) and the Mellon Foundation.

References

Beres LK, Winskell K, Neri EM, Mbakwem B, & Obyerodhyambo O (2013). Making sense of HIV testing: Social representations in young Africans' HIV-related narratives from six countries. Global Public Health, 8(8), 890–903. doi: 10.1080/17441692.2013.827734 [PubMed: 24004339]

Bruner J (1990). Acts of Meaning. Cambridge: Harvard University Press.

Bruner J (1991). The Narrative Construction of Reality. Critical Inquiry, 18(1), 1–21. doi:10.1086/448619

Bruner J, & Lucariello J (1989). Monologue as narrative recreation of the world. Narratives from the crib, 73–97.

Corbin J, & Strauss A (2008). Basics of qualitative research: techniques and procedures for developing grounded theory. Thousand Oaks, California: Sage Publications.

Czarniawska B (2004). Narratives in Social Science Research. London: Sage.

Elliott J (2005). Using Narrative in Social Research: Qualitative and Quantitative Approaches. London: Sage.

Farmer P, & Good BJ (1991). Illness Representations in Medical Anthropology: A Critical Review and a Case Study of the Representation of AIDS in Haiti In Skelton JA & Croyle RT (Eds.), Mental Representations in Health and Illness (pp. 132–162). New York: Springer-Verlag.

Flick U, Foster J, & Caillaud S (2015). Researching social representations In Sammut G, Andreouli E, Gaskell G, & Valsiner J (Eds.), Cambridge Handbook of Social Representations (pp. 64–80). Cambridge: Cambridge University Press.

Franzosi R (1998). Narrative analysis—or why (and how) sociologists should be interested in narrative. Annual Review of Sociology, 24(1), 517–554. doi:10.1146/annurev.soc.24.1.517

Franzosi R (2010). Quantitative Narrative Analysis. Thousand Oaks: Sage Publications, Inc.

Geertz C (1973). The Interpretation of Cultures. New York: Basic Books.

Germain J, Harris J, Mackay S, & Maxwell C (2017). Why should we use online research methods? Four doctoral health student perspectives. Qualitative Health Research, 1049732317721698.

Hirsch J, Wardlow H, Smith D, Phinney H, Parikh S, & Nathanson C (2009). The Secret: Love, Marriage and HIV. Nashville, TN: Vanderbilt University Press.

Howarth C (2006). How Social Representations of Attitudes have informed Attitude Theories: The Consensual and the Reified. Theory & Psychology, 16(5), 691–714. doi:10.1177/0959354306067443

Jovchelovitch S (2002). Social representations and narrative: Stories of public life in Brazil In Laszlo J & Rogers W. Stainton (Eds.), Narrative Approaches in Social Psychology. Budapest, Hungary: New Mandate.

Labov W, & Waletzky J (1967). Narrative analysis: Oral versions of personal experience In Helm J (Ed.), Essays on the verbal and visual arts. Seattle: University of Washington Press.

Laszlo J (1997). Narrative organisation of social representations. Papers on Social Representations, 6, 155–172.

MacQueen KM, McLellan-Lemal E, Bartholow K, & Milstein B (2008). Team-based Codebook Development: Structure, Process, and Agreement In Guest G & MacQueen KM (Eds.), Handbook for Team-Based Qualitative Research. Lanham, New York, Toronto, Plymouth: Altamira Press.

Markova I, & Wilkie P (1987). Representations, concepts and social change: The phenomenon of AIDS. Journal for the Theory of Social Behaviour, 17(4), 389–409. doi:10.1111/j.1468-5914.1987.tb00105.x

Miles MB, & Huberman MA (1994). Qualitative Data Analysis: An Expanded Sourcebook (Second Edition ed.). Thousand Oaks, London, New Delhi: Sage Publications.

Mishler EG (1986). The analysis of interview-narratives In Sarbin T (Ed.), Narrative psychology: The storied nature of human conduct (pp. 233–255). Westport, CT: Praeger/Greenwood.

Mishler EG (1995). Models of narrative analysis: A typology. Journal of Narrative and Life History, 5(2), 87–123. doi:10.1075/jnlh.5.2.01mod

Moscovici S (1981). On social representations In Forgas JP (Ed.), Social cognition: Perspectives on everyday understanding (pp. 181–209). London: Academic Press.

Murray M (2002). Connecting narrative and social representation theory in health research. Social Science Information, 41(4), 653–673. doi:10.1177/0539018402041004008

Oliver KL (1998). A journey into narrative analysis: A methodology for discovering meanings. Journal of Teaching in Physical Education, 17(2), 244–259. doi:10.1123/jtpe.17.2.244

Rabinow P (1986). Representations are Social Facts: Modernity and Post-Modernity in Anthropology In Clifford J & Marcus GE (Eds.), Writing Culture: The Poetics and Politics of Ethnography (pp. 234–261). Berkeley and Los Angeles, CA: University of California Press.

Riessman CK (1983). Women and medicalization: a new perspective. Social Policy, 14(1), 3. [PubMed: 10264493]

Riessman CK (2008). Narrative methods for the human sciences. Thousand Oaks: Sage.

Ryan GW, & Bernard HR (2003). Techniques to identify themes. Field Methods, 15(1), 85–109. doi:10.1177/1525822x02239569

Savage M (2011). Using archived qualitative data: Researching socio-cultural change In Mason J & Dale A (Eds.), Understanding Social Research: Thinking Creatively about Method (pp. 169–180). London: Sage Publications.

Snowden D (2011). Naturalizing sensemaking In Mosier KL & Fischer UM (Eds.), Informed by knowledge: Expert performance in complex situations (pp. 223–234). New York: Taylor & Francis Group.

UNAIDS. (2006). Report on the Global AIDS Epidemic. Retrieved from http://data.unaids.org/pub/report/2006/2006_gr_en.pdf:

VERBI Software. (1989-2010). MAXQDA, software for qualitative data analysis (Version 2007). Berlin-Marburg-Amöneburg, Germany.

Viney LL, & Bousfield L (1991). Narrative analysis: A method of psychosocial research for AIDS-affected people. Social Science & Medicine, 32(7), 757–765. doi:10.1016/0277-9536(91)90301-R [PubMed: 2028270]

Wang T (2013). Big Data Needs Thick Data. Retrieved from http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/

Watkins SC, & Swidler A (2009). Hearsay ethnography: Conversational journals as a method for studying culture in action. Poetics, 37(2), 162–184. doi:10.1016/j.poetic.2009.03.002 [PubMed: 20161457]

Winskell K, Beres LK, Hill E, Mbakwem BC, & Obyerodhyambo O (2011). Making sense of abstinence: social representations in young Africans' HIV-related narratives from six countries. Culture, Health & Sexuality, 13(8), 945–959. doi:10.1080/13691058.2011.591431.

Winskell K, Brown PJ, Patterson AE, Burkot C, & Mbakwem BC (2013). Making Sense of HIV in Southeastern Nigeria. Medical Anthropology Quarterly, 27(2), 193–214. doi: 10.1111/maq.12023. [PubMed: 23804317]

Winskell K, & Enger D (2005). Young voices travel far: a case study of Scenarios from Africa Media and glocal change: Rethinking communication for development, 403–416. Buenos Aires/Goteborg: Clasco/Nordicom.

Winskell K, & Enger D (2009). A new way of perceiving the pandemic: the findings from a participatory research process on young Africans' stories about HIV/AIDS. Culture, Health & Sexuality, 11(4), 453–467. doi: 10.1080/13691050902736984.

Winskell K, Hill E, & Obyerodhyambo O (2011). Comparing HIV-related symbolic stigma in six African countries: Social representations in young people's narratives. Social Science & Medicine, 73(8), 1257–1265. doi:10.1016/j.socscimed.2011.07.007. [PubMed: 21864965]

Winskell K, Holmes K, Neri E, Berkowitz R, Mbakwem B, & Obyerodhyambo O (2015). Making sense of HIV stigma: Representations in young Africans' HIV-related narratives. Global Public Health, 10(8), 917–929. doi: 10.1080/17441692.2015.1045917. [PubMed: 26132087]

Winskell K, Obyerodhyambo O, & Stephenson R (2011). Making sense of condoms: Social representations in young people's HIV-related narratives from six African countries. Social Science & Medicine, 72(6), 953–961. doi: 10.1016/j.socscimed.2011.01.014. [PubMed: 21388731]

Wright S (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Paper presented at the Proceedings of the Sixth International Congress on Genetics, Austin, TX.

Table 1

Sampling Pool from 5 Countries, 1997-2014

	1997	2000	2002	2005	2008	2011	2013	2014	TOTALS
Senegal	1,202	1,166	1,245	2,824	1,643	20	228	360	**8,688**
Burkina Faso	450	390	1,099	4,821	3,769	283	1,043	1,190	**13,045**
Nigeria	--	--	--	1,869	1,156	409	574	1,316	**5,324**
Kenya	--	--	--	673	117	39	137	3,007	**3,973**
Swaziland	--	--	--	510	716	56	211	236	**1,729**
TOTALS	**1,652**	**1,556**	**2,344**	**10,697**	**7,401**	**807**	**2,193**	**6,109**	**32,759**

--: no data available

Table 2

Final Sample

	1997	2000	2002	2005	2008	2011	2013	2014	TOTALS
Senegal	86	109	104	107	79	--	14	67	**566**
Burkina Faso	44	45	103	112	100	27	--	56	**487**
Nigeria	--	--	--	120	93	34	65	88	**400**
Kenya	--	--	--	88	25	--	45	116	**274**
Swaziland	--	--	--	72	50	--	48	40	**210**
TOTALS	**130**	**154**	**207**	**499**	**347**	**61**	**172**	**367**	**1,937**

--: no data available