A practical guide to analysing online support forums

Richard M. Smedley PhD [a], Neil S. Coulson PhD [a]

[a] Division of Rehabilitation and Ageing, Queen's Medical Centre, Nottingham, NG7 2UH

Address for correspondence:

Professor Neil S. Coulson
Division of Rehabilitation & Ageing, School of Medicine
Queen's Medical Centre
Nottingham
NG7 2UH
UK

+44 115 846 6642

Abstract

As the number of online support forums continues to grow and the amount of user-generated content increases, analysing forum messages can provide researchers and others working in health-related fields with important new insights into the needs, opinions and experiences of individuals who use them. There has been a lack of guidance regarding how to make the best use of user-generated content within forums as a data source in research studies, so this paper explains the practical, ethical and methodological issues associated with this type of research. It describes the benefits of forum research, the organisational structure of forums, how forums are selected for research, approaches to sampling, preparing data for analysis, and methods of analysis that can be used. The decisions that researchers need to make during each stage of the research process are explained, describing the options available and the ethical dilemmas that need to be considered to successfully develop, carry out and complete a research project.

Keywords: Online support communities; Forum; Sampling typology; Research methods; Ethics.

About the authors:

Richard M. Smedley has recently completed a PhD in Health Psychology in the School of Medicine at the University of Nottingham. His research examines how engaging with discussion forums can help individuals with the self-care and self-management of long-term health conditions.

Neil S. Coulson is a Professor of Health Psychology in the School of Medicine at the University of Nottingham. As a health psychologist, his programme of research examines how social media impacts on the experience of long-term illness. His current interests focus on the role of discussion forums and the links between active engagement and various health-related and psychosocial outcomes.

1 Introduction

Online support forums, also known as 'online health communities' or 'online support groups' (hereafter referred to as 'forums' for brevity), have been defined as "*online services with features that enable members to communicate with each other*" (Malinen, 2015). Forums can play an important role in helping and supporting individuals with the self-care and self-management of various health-related challenges (Coulson, 2013; Coursaris & Liu, 2009; Smedley, Coulson, Gavin, Rodham, & Watts, 2015). Individuals may join because they either have or are at risk of developing a specific illness such as diabetes (Loader, Muncer, Burrows, Pleace, & Nettleton, 2002), because they are a caregiver (Carter, 2009), for help with problem behaviours such as weight management (Bane, Haymaker, & Zinchuk, 2005) or because of other difficulties such as bereavement (van der Houwen, Stroebe, Schut, Stroebe, & van den Bout, 2010).

Forums provide a dynamic environment where individuals can interact with others to discuss health-related topics, learn how other people have dealt with problems or receive social support, comfort, empathy and companionship (Attard & Coulson, 2012; Chen, 2014; Coulson, 2013; Malik & Coulson, 2008). One of the unique aspects of forums is that they rely predominantly on members to both generate and consume content (Malinen, 2015), that is to say, these forums exist so that members can use them to interact with others by posting messages and reading replies. Indeed, members who participate in these forums acknowledge the potential scientific importance of studies analysing user-generated content, and recognise that this can provide researchers with important insights into the needs, opinions and lived experiences of individuals who are living with a health condition (Bond, Ahmed, Hind, Thomas, & Hewitt-Taylor, 2013).

For example, researchers have used forums to examine the self-care of health conditions (Malik & Coulson, 2011; Smedley et al., 2015), why people use online support (Coulson, 2013; Coulson et al., 2016), how individuals engage with forums (Smedley, Coulson, Gavin, Rodham, & Watts, 2016), the types of support exchanged (Coulson, Buchanan, & Aubeeluck, 2007; Coursaris & Liu, 2009), the

empowering effects of taking part (Mo & Coulson, 2012; van Uden-Kraan, Drossaert, Taal, Seydel, & van de Laar, 2009), the experiences of moderators (Coulson & Shaw, 2013; Smedley & Coulson, 2017), and a range of other important topics.

In the literature, researchers have discussed how to analyse forum data (e.g. Holtz, Kronberger, & Wagner, 2012; Im & Chee, 2006) and the ethical issues associated with this (e.g. Eysenbach & Till, 2001; Roberts, 2015), however few papers have drawn this together to provide guidance on every aspect of the research process.

The aim of this paper is to introduce researchers to health-related forums by providing a summary of the key potential benefits of working with user-generated content, an overview of their basic organisational structure, how to select a forum for research, sampling issues, preparing data for analysis as well as popular types of data analysis used by researchers working with this type of data.

1.1 Potential benefits and disadvantages of working with user-generated forum content

For researchers who are considering working with forums in order to analyse user-generated content, there are a range of potential benefits and disadvantages from this source of data (see Table 1).

<< Insert Table 1 about here >>

1.2 Organisation of online forums: boards, threads and messages

Most forums are asynchronous and have a hierarchical, tree-like structure, as shown in Figure 1. A forum can contain several distinct boards or sections that are arranged thematically, where each board will contain many different threads, and each thread will contain one or more individual messages written by members (Holtz et al., 2012).

<< Insert Figure 1 about here >>

At the top level, the forum usually has an index page listing the various boards or sections that are available, where each board provides an area where members can have a particular type of discussion. For example, a diabetes forum could have separate boards for introductions, newly diagnosed people, type 1 and type 2 diabetes, and various other issues related to this condition. Each of these boards will contain many different conversations, which are organised into groups of related messages called 'threads'. A thread consists of an initial message where a member starts a new discussion by asking a question, describing an experience or requesting advice, and other members contribute by posting replies (Cummings, Sproull, & Kiesler, 2002). Some forums allow members to reply to an individual message, giving threads a complex structure where a reply might receive several further replies of its own. Other forums are more limited and only allow members to post replies to the thread rather than responding to a specific message, making it more difficult to keep track of who members are replying to (Petrovčič, Vehovar, & Žiberna, 2012). The use of threading makes it possible for forum members to take part in several discussions simultaneously, by posting replies in one thread and then moving to another thread or board to contribute to discussions on other topics and with other members.

Each forum posting contains a message together with header information about the member who wrote it, such as their username and the date and time when it was posted to the group (Petrovčič et al., 2012). The username may consist of their real name (e.g. johnsmith), a nickname (e.g. mickeymouse), or a word or phrase that makes them anonymous (e.g. letsconnect). It is possible for more than one person to share an account, for example a husband and wife might share their password so they can both use the forum with the same username. Members will often 'sign' their posting by typing their name at the bottom of the message, in a similar way to signing an e-mail, which helps to clarify who wrote the message. When an individual joins multiple forums, they can either use the same username across all forums or choose a different username for each forum to help maintain their anonymity.

Some forums have user profiles that display additional information about each member, such as their age and gender, their geographical location, the date when they joined the forum, how many messages

they have posted, how many threads they have started, and possibly an avatar. An avatar is a small graphical image that is typically displayed alongside that member's username, showing their photograph or another picture that they have selected.

The message text may include quoted extracts from other messages that have previously been posted to the forum, making it easier to clarify which parts of a message that person is responding to. It is often possible to embed pictures into the message, attach files, or include links to other webpages. Another key feature of many forums is the use of emoticons (Im & Chee, 2006; van Uden-Kraan, Drossaert, Taal, Lebrun, et al., 2008). These are 'smiley faces' included in the text to indicate the emotional state of that member, as a substitute for non-verbal cues that are missing from written communication such as facial expressions and tone of voice. Two distinct styles of emoticon can be found within forums: Faces that are typed using normal keyboard characters such as ':-)' and ':-(' (a colon, dash and parenthesis that resemble a happy or sad face when the page is rotated sideways), and graphical images depicting various facial expressions that are embedded into the text.

Members may also include a signature block that is displayed below all of their messages. A signature block is an additional line or two of text that is appended to the end of the posting, which may contain their username, a favourite quote, a link to their personal website or blog, an image, or some other information that the user wishes to share with other members.

1.3 Overview of the research process

The flowchart in Figure 2 provides a simplified overview of the research process and the key decisions that need to be made when planning and carrying out a study, which some readers may find helpful. These decisions, the choices available, and the practical, methodological and ethical issues associated with them are discussed in the sections below. Data sources are discussed in Section 2, sampling forum content in Section 3, data preparation in Section 4, data analysis in Section 5, and ethics in Section 6.

<< Insert Figure 2 about here >>

1.4 The status of forum data

There is uncertainty over whether forum messages should be treated as a collection of individual voices (equivalent to the concept of participants in other forms of data collection) or as written documents (comparable to other textual documents such as books or newspaper columns). This distinction has important ethical implications.

Wilkinson and Thelwall (2011) draw a distinction between people and documents as research objects, noting that participants are protected by ethical procedures while documents can often be used for research purposes without ethical difficulties. They argue that personal information on publically-accessible websites should be treated as if they are electronic documents that have been published on the Internet, implying that research analysing forum postings involves studying electronic documents rather than human participants. This is different to the conventional approach of assuming that the authors of forum postings should be treated as research participants, and hence it could potentially free online observational research from some ethical dilemmas such as whether to obtain informed consent (ethical issues are discussed further in Section 6).

However, Lomborg (2013) points out that it is unclear whether forum postings and other textual data should be treated as documents or human participants. Answering this may depend upon the distance between the information to be analysed and the individual who created that information. With some data sources such as forum postings, the distance between an identifiable user and the text may be small, whereas a large study involving millions of messages might be less likely to involve human participants because of the longer distance between specific people and their online behaviour.

Throughout this paper, the word ‘participant’ has been used purely for consistency with conventional research terminology, and does not assume any particular epistemological position regarding the status of forum data.

2 Data sources for forum research

Researchers can conduct studies by either analysing messages from one or more existing forums, or by creating a bespoke forum as part of a research project.

2.1 Using existing forums

As a first step, researchers typically find forums by using one or more Internet search engines such as Google, Bing and/or DuckDuckGo and searching for relevant keywords, such as 'online/internet support group', 'forum', 'message board', 'discussion group' and the name of a health condition like 'diabetes', 'cancer' or 'Parkinson's disease' (Attard & Coulson, 2012; Coulson & Greenwood, 2012; Hwang et al., 2007; van Uden-Kraan, Drossaert, Taal, Lebrun, et al., 2008). Additionally, it may be possible to identify forums by consulting with colleagues and experts in the appropriate field (van Uden-Kraan, Drossaert, Taal, Lebrun, et al., 2008). When a list of potential forums has been produced, each forum should be examined to assess whether it is suitable for analysis. The researcher needs to consider whether the topics that are discussed in each forum will inform the research question, if members are likely to be representative of the desired target population, and whether the forum contains enough postings by a sufficiently large number of members to provide an appropriate volume and depth of data for analysis (Holtz et al., 2012).

Other potentially important considerations include: is the forum public or private (see Section 6.1)? Are there are any restrictions on who can join (Eysenbach & Till, 2001)?How popular is the website in terms of visitors and page views (Hwang et al., 2010)? Does the forum has a long history or has it won any awards (Coursaris & Liu, 2009)? The willingness of moderators and website administrators to support research activities might also be a consideration (e.g. Mo & Coulson, 2010; van Uden-Kraan, Drossaert, Taal, Shaw, et al., 2008). If other researchers have previously posted messages in the forum to recruit participants and they have received a negative or hostile response, then this may indicate that researchers are not welcome (Eysenbach & Till, 2001).

When selecting a forum as a data source it should be remembered that members will not normally be aware that their messages are being downloaded and analysed as part of a scientific study, and they may react badly if they realise that the forum is being monitored and used for research purposes without their knowledge (Eysenbach & Till, 2001; Hudson & Bruckman, 2004). Members join and participate in forums for a variety of personal reasons including looking for information, social support or friendship (Ridings & Gefen, 2004), and not because they want to act as 'guinea pigs' who are being observed and studied by researchers (Eysenbach & Till, 2001).

How the researcher proceeds may require careful thought, and will be guided by both the nature of the forum and whether the members are reasonably likely to perceive it as being a public or private place (Roberts, 2015). 'Closed' forums (requiring registration) are nearly always private, while 'open' forums (viewable without registration) could be public or private depending upon the nature of the forum (Eysenbach & Till, 2001). Forums with a smaller number of members are more likely to be private, as are forums that discuss sensitive topics, and forums intended for a specific target audience such as children (Eysenbach & Till, 2001). If a forum is public and also receptive to research activities then the researcher can proceed with sampling forum content, as described in Section 3. If the forum is private then the forum's moderators or administrators should be contacted for advice on how to proceed. They will either deny or permit access, possibly with their own conditions attached.

Whether a forum is public or private, the researcher's primary aim is to avoid any potential harm to both the forum and its members. For example, if one or more forum members were to leave as a result of researcher activity, then this could harm group dynamics as well as preventing those individuals from receiving the help and support that they needed. The ethical difficulties associated with this type of research, including important aspects of the distinction between public and private places, are explored further in Section 6.

2.2 Creating a bespoke forum

Alternatively, researchers can create and launch their own dedicated forum as part of a research project (Armstrong, Koteyko, & Powell, 2012; Im & Chee, 2006; Sanders, Rogers, Gardner, & Kennedy, 2011; Zrebiec & Jacobson, 2001). This may have advantages over downloading data from an existing forum because participants can be told when joining that the forum is part of a scientific study, and so they can consent to their data being used for research purposes. However, this may change group dynamics within the forum because members will be aware that they are being observed. Their motivation for joining the forum may also be different, i.e. members may join to take part in a scientific study rather than for mutual self-help (Smedley et al., 2016). Creating a new forum may considerably increase the expenses involved in conducting a study. Server space will be needed to host the forum on the Internet, specialist computer skills are required to set-up and maintain a forum, and it may be difficult and time-consuming to recruit enough forum members for the community to become self-sustaining and generate the required amount of data for analysis.

3 Sampling forum content

When sampling user-generated content within forums, there are a number of considerations that may be directly relevant including: forum characteristics; user characteristics; message characteristics and temporal focus. Each of these considerations are important potential influences on the sampling strategy adopted but need not all be relevant to every study.

3.1 Forum characteristics

Section 2 described some of the strategies that are commonly used by researchers to identify forums that might potentially be suitable for use as data sources in research studies. These techniques often lead to researchers selecting specific forums for use in research based on predetermined criteria to ensure they are suitable for the needs of the study. One important decision to be made is whether to collect data from a single forum (Coulson et al., 2007; Rodham, Gavin, Lewis, St. Denis, & Bandalli, 2013) or multiple forums (Brotsky & Giles, 2007; Eichhorn, 2008).

If data are collected from multiple forums then homogeneity may also be important, that is to say, how similar the forums are and if they are dedicated to discussions about the same health-related condition or multiple health conditions. Some studies use data from two or more similar forums to collect a larger and more representative selection of messages centred around a specific topic or health concern, such as Parkinson's disease (Attard & Coulson, 2012) or eating disorders (Eichhorn, 2008). Other studies collect data from diverse forums to examine broader issues that may apply across all forums rather than being specific to a particular health condition, such as looking at the activities performed by moderators (Smedley & Coulson, 2017) or investigating the potential disadvantages of using forums (van Uden-Kraan, Drossaert, Taal, Lebrun, et al., 2008). Other considerations may include how many members participate in each forum, how many messages have been posted, and whether the forum is active or dormant.

Some studies have been conducted where forums have been selected entirely at random. This is typically achieved by obtaining a list of potential forums and then using a random selection process to choose groups for inclusion in the study. Sometimes these groups are chosen entirely randomly regardless of what illness they represent or any other considerations (Hudson & Bruckman, 2004; Nonnecke & Preece, 2000; Yechiam & Barron, 2003), and sometimes they are 'filtered' by using inclusion criteria to exclude dormant groups or those that do not have a sufficient number of active members (McKenna, Green, & Gleason, 2002; Preece, Nonnecke, & Andrews, 2004; Voelpel, Eckhoff, & Förster, 2008).

3.2 User characteristics

Forum members can be selected based upon specific characteristics, such as their gender or age (Gustafson et al., 2001; Gustafson et al., 1998; Holbrey & Coulson, 2013; Malik & Coulson, 2008). Forum members sometimes adapt an anonymous online persona that may make it difficult to obtain relevant demographic information about participants (Brady, Segar, & Sanders, 2016), but it is often possible to obtain limited information from the self-disclosed content of postings or by examining

their online profiles (Eichhorn, 2008; Finfgeld-Connett, 2015; Holtz et al., 2012; Stieger & Göritz, 2006; Wilkinson & Thelwall, 2011).

Participants can also be randomly selected for inclusion in the study. The researcher will typically compile a complete list of all usernames, profiles or e-mail addresses within a forum, and then use random numbers to select which members to include as participants (Matzat & Rooks, 2014; Nosek, Banaji, & Greenwald, 2002; Stieger & Göritz, 2006; Thackeray, Crookston, & West, 2013), making a random selection process possible. Note that this type of randomization can only be used for certain types of analysis, such as content analysis.

One advantage of this technique is that, if participants are randomly selected from the overall group membership, then it is likely they will form a representative cross-section of those members (Levin, 2006). This should mean that less-active members of the forum are targeted in addition to focusing on the most-active members who might be responsible for producing the majority of group content (Matzat & Rooks, 2014). However, as with all message analysis studies, the resulting dataset will only reflect the experiences of active participants who have posted messages, and not 'lurkers' who read messages but do not take part in discussions (Nosek et al., 2002).

3.3 Message characteristics

Messages are typically selected for inclusion in a study by sampling at the level of either threads or messages. Sampling based on threads involves looking at complete discussion threads, where each thread is a self-contained conversation consisting of messages written by one or more forum users. Sampling on the level of messages involves selecting individual messages irrespective of the discussion threads that they were originally posted in. Within these options, a range of sampling criteria can be used.

3.3.1 Selecting messages/threads based on specific characteristics

Messages can be selected to fulfil specific criteria as required for the needs of the study, with several approaches being used in the literature. The *time period* when messages were posted can be a particularly important consideration, with researchers often selecting messages that were posted between a specific start and end date to capture all forum content over a particular number of days, weeks, months or years. There can be both practical and theoretical reasons for studying a particular time period. The main practical reason is to make the data more manageable, as illustrated in some of the studies described below. The theoretical reasons relate to particular research interests, such as studying how support processes develop during the first six months after a new forum is launched (Armstrong et al., 2012).

Some studies examine all messages posted during the chosen time period (Drentea & Moren-Cross, 2005; Esquivel, Meric-Bernstam, & Bernstam, 2006; Evans, Donelle, & Hume-Loveland, 2012; Finn, 1999; Galegher, Sproull, & Kiesler, 1998; Loader et al., 2002; Malik & Coulson, 2008; Tichon & Shapiro, 2003), while others only analyse a subset of those messages. For example, Coursaris and Liu (2009) selected all messages posted during a one-year period, divided these into batches of 100 messages, and then randomly selected 50 batches for analysis; Gavin, Rodham, and Poyer (2008) initially selected all messages posted during a one-week period, and then analysed a subset covering three-days to reduce both the volume of postings and repetition of key themes; Malik and Coulson (2010) selected all messages posted on seven boards during a four-month period, and then used a random number generator to select a subset of 500 messages from each board for analysis. In these studies, randomization differs from that described in Section 3.3.2 because it is only applied to a selected range of messages.

Alternatively, researchers can select a predetermined number of messages or threads to ensure they have the required amount of data for analysis, by selecting the most recent messages posted to the forum (Malik & Coulson, 2011; Rodham, Gavin, Lewis, et al., 2013). Some studies have looked at the first message posted by participants to examine various issues relating to how individuals introduce

themselves when they join a forum (Galegher et al., 1998; Joyce & Kraut, 2006; Rodham, Gavin, Coulson, Watts, & Heath, 2013; Smedley et al., 2016). Researchers have also looked at the number of replies that are received (Galegher et al., 1998; Smedley et al., 2016; Winzelberg, 1997), for example there is a possibility that some individuals might receive more replies than others, or that some health-related challenges may make it difficult to provide support over the Internet.

3.3.2 Randomly selected messages/threads

A random sampling strategy can be used to ensure that every thread or message posted in the forum has an equal probability of being selected for inclusion in the study. This technique may enhance the study's credibility by ensuring that content has not been subjectively chosen. Random sampling can be achieved by allocating a number to each thread or message and then using a random number generator to select which ones to include, or by randomly choosing a forum page and then randomly selecting a thread or message within that page (Attard & Coulson, 2012; Coulson & Greenwood, 2012; Elwell, Grogan, & Coulson, 2011; Ridings & Gefen, 2004). It should be noted that randomly selecting individual messages will remove those postings from the context in which they were originally written, which could make it more difficult for researchers to understand the content of those messages, while randomly selecting complete threads allows those messages to be examined in their original context (Coursaris & Liu, 2009).

3.3.3 Selecting all available content

Some studies have analysed the complete set of all messages posted to a forum, without filtering for participant or message characteristics. This means that all messages are selected for analysis regardless of how many messages were posted, who wrote them, when they were written, or any other considerations (Armstrong et al., 2012; Smedley et al., 2015; Zrebiec & Jacobson, 2001). By including all participants and messages in the analysis, it ensures that the results will be representative of all forum members. Depending on the forum used, this could potentially yield a very large number of messages for analysis and in some instances may render the dataset unmanageable.

3.4 Temporal focus

The final consideration is timeline, that is to say, whether the study is carried out using a retrospective or prospective approach. Retrospective studies involve looking back through message archives to examine previous forum messages that were written by member(s) at a particular period in time. Prospective studies are comparatively rare and typically use a longitudinal design to look forward from a point of entry to examine any changes that arise, develop and evolve over time (e.g. Wen, McTavish, Kreps, Wise, & Gustafson, 2011). However, prospective approaches may become important in the future as new techniques are developed for analysing forum messages longitudinally.

4 Preparing your data for analysis

The selected messages should be exported and stored in a text document such as a word processor file on the researcher's computer, in case the forum closes or becomes unavailable (Holtz et al., 2012). It is advisable to use a text file, database or spreadsheet to keep a record of the full URL for each individual message or thread that is downloaded, so that the original postings can be found again if it becomes necessary to refer back to them while analysing the data or carrying out any revisions that might be needed prior to publication.

A key advantage of using online data from forums is that it eliminates the need to transcribe data, which can yield major time savings and also removes the possibility of transcription errors. Forums use written communication that produces an automatic and error-free verbatim transcript of the interactions between members, rather than relying on a recording that needs to be subsequently transcribed in an attempt to capture as much detail as possible from the original interaction. However, the analysis of forum messages can be problematic. The process of importing messages into a word processor or other program may cause key features of the text to be lost, including emoticons and the position of line breaks (Im & Chee, 2006). This may be a particular problem with forum messages that use embedded graphical images for emoticons rather than emoticons that have been typed using standard keyboard characters, as described in Section 1.2.

The loss of some textual features might potentially threaten the credibility of the data because forums do not convey non-verbal cues such as facial expressions and tone of voice, and textual features such as spacing and emoticons could provide some clues to the emotional state of participants and the context of what is being said (Im & Chee, 2006). The researcher needs to decide which aspects of the forum are important for analysis so that important features can be retained while irrelevant features are removed from the downloaded version of the text. For example, the unedited forum data is likely to include the original formatting and HTML tables used to display the messages on the webpage, together with usernames, emoticons, avatars, embedded pictures, quotes from previous messages, and any advertisements that might have originally been shown alongside the postings (Holtz et al., 2012).

Some researchers find it useful to make printouts of all the messages, to retain a verbatim hardcopy of the original formatting, layout, emoticons, and other textual features. Alternatively it might be possible to take screenshots of messages, which would also retain the original formatting. It should be remembered that screenshots will convert the messages into graphical format, so it would be advisable to save a text copy of each message alongside the screenshots to avoid any potential difficulties converting the graphical screenshots back into plain text format.

When exporting, converting and saving messages, Holtz et al. (2012) recommend removing tables, converting the messages into plain text, and replacing all hard returns with soft returns. Hard returns are line breaks that usually denote the end of a paragraph, while soft returns mark the end of a line and are automatically inserted by the word processor's word wrapping facility. When copying text from a web browser into a word processor, hard returns are sometimes inserted at the end of each line. These may need to be converted into soft returns to restore the original text layout. Sometimes it can be useful to retain quoted extracts from previous messages because this can help to clarify precisely which message a member is replying to (Petrovčič et al., 2012). It may also be important to retain any website addresses or embedded images/videos that are present within the messages, so that these can be examined in order to fully understand the content and context of the message during analysis.

It may be possible to anonymise the data by using a word processor's 'find and replace' function to change personal details such as names or usernames, but care should be taken when doing this. Personal information can be located across the entirety of an individual's messages and it may be difficult to identify search terms that will replace all personal data without some details being overlooked such as the names of doctors and family members (Kaiser, 2009).

The participants in forum studies are the members who wrote the messages selected for analysis. Some forums have user profiles that members may use to provide demographic information about themselves such as their age and gender (Holtz et al., 2012), and it is sometimes possible to obtain demographic information about participants by examining the content of messages posted to the forum (Coulson et al., 2007; Malik & Coulson, 2010). It is difficult to confirm the accuracy of demographic information about participants (Holtz et al., 2012) but one study comparing profile data with self-reported information from online interviews found that the demographic information in online profiles was broadly accurate, with gender discrepancies only arising in 5.9% of participants and age discrepancies in 11.2% of participants (Stieger & Göritz, 2006).

5 Data analysis

This section provides a brief introduction to some of the qualitative analytical methods that can be used; interested readers are encouraged to pursue the supplied references for further information about these techniques.

Qualitative analysis is commonly used to study the content of messages and investigate what people talk about. This can involve using text-reducing or text-enhancing methods (Holtz et al., 2012). Text-reducing methods such as content analysis and thematic analysis reduce the complexity of qualitative data by identifying key themes or categories, while text-enhancing methods such as discourse analysis look at the meaning of different styles of communication. If desired, research teams have the option for more than one researcher to be involved in analysing qualitative data (e.g. Coulson et al., 2016).

This could help with identifying potential themes, reviewing them for coherency, and checking that the data has been coded and/or categorised correctly.

5. 1 Content analysis

Content analysis (Elo & Kyngäs, 2008; Pfeil & Zaphiris, 2010) is used to code rich textual data into categories, which are then used to produce simple statistics summarising the data by counting how many occurrences there are in each category. This can be done using an inductive or deductive approach. Inductive coding (bottom-up) involves generating coding categories from the dataset, and is useful when developing new theories or researching topics where little is known. For example, Hwang et al. (2010) hypothesised that forum social support might differ from face-to-face social support, so they used forum messages to create their own social support categories and examined how many messages fell within each category. Generating categories from the dataset should produce results that closely reflect the dataset, but it may mean that the results lack standardisation. Deductive coding (top-down) categorises the data into an existing framework, and is useful for testing existing theories or comparing different datasets. For example, Smedley et al. (2015) looked at social support in forums by coding messages into an existing set of support categories that have been used in several other studies. Using existing categories makes it easier for researchers to build a 'bigger picture' by comparing studies that have used the same coding scheme.

If desired, it is possible to use test-retest and inter-rater reliability testing with deductive content analysis. With test-retest reliability testing, a single researcher codes the data, waits a suitable time period, and then recodes either part or all of the data. With inter-rater reliability testing, the first researcher codes all the data while a second researcher independently codes either part or all of the data. The proportion of agreement between the coding sessions can be calculated using either a simple formula (e.g. percentage agreement) or Cohen's Kappa, as described by Ballinger, Yardley, and Payne (2004).

5. 2 Thematic analysis

Thematic analysis (Braun & Clarke, 2006) is used to identify patterns (or "themes") that are present across a dataset, to describe, analyse and interpret a phenomenon of interest. Thematic analysis is not linked to any particular theoretical framework and can be done using an inductive or deductive approach. The reasons for choosing an inductive or deductive approach, and the advantages and limitations associated with this, are the same as those for content analysis (Section 5.1). Inductive thematic analysis involves generating themes directly from the forum messages. For example, Ravert, Hancock, and Ingersoll (2004) examined what topics were discussed by adolescent members in diabetes forums. Deductive thematic analysis is similar to deductive content analysis, where the data is coded into an existing framework but without counting the number of occurrences. For example, Coulson and Greenwood (2012) looked at how social support was expressed in a forum for the families of children with cancer.

5. 3 Discourse analysis

Discourse analysis (Willig, 2003; Yardley & Murray, 2004) is a text-enhancing method that researchers use when they want to understand the 'meta' meaning of text rather than its semantic meaning (Hodges, Kuper, & Reeves, 2008). With discourse analysis language is used to construct social meanings rather than reflecting what people think, for example the statement 'I am in pain' might be used as a reason to withdraw from a social setting, to avoid completing a task at work, or to get sympathy from family and friends (see Yardley & Murray, 2004). For example, discourse analysis has been used to see how community identity shaped subsequent discussions in a forum for type-1 diabetes (Armstrong et al., 2012).

5.4 Other analytical approaches

Forum messages can also be analysed using interpretative phenomenological analysis and grounded theory. Interpretative phenomenological analysis (Smith & Osborn, 2003) is used when researchers want to examine how individuals make sense of their world, by exploring their perception of events rather than conducting an objective analysis. For example, Rodham, McCabe, and Blake (2009) used

this to investigate how individuals seek and provide support in a forum for complex regional pain syndrome. Grounded theory (Charmaz, 2003) takes individual cases or incidents and uses them to progressively build conceptual categories that can be used to identify and interpret patterns. For example, Wen et al. (2011) used grounded theory to conduct a longitudinal case study of a woman with breast cancer, analysing how her messages changes over the course of her illness from being diagnosed through to her death.

Other analytical approaches that have been used in the literature include social network analysis (Freeman, 2004; Luke & Harris, 2007) and semi-automated data mining techniques (Wiedemann, 2013). Studies that use these approaches may face similar concerns regarding data collection and ethical issues, but the methods of data analysis used are profoundly different and fall outside the scope of this paper. Interested researchers are encouraged to pursue the above references as a starting point for further information.

6 Ethical issues

Some national and international institutions provide guidelines and ethical codes of conduct to assist researchers when conducting online studies (e.g. American Psychological Association, 2017; British Psychological Society, 2017; NESH, 2014). If guidelines of this type are available in the researcher's country then they should always be examined and complied with as a matter of good practice. Where such guidelines exist, the ethical information provided here should be treated as a supplement to those guidelines.

Message analysis studies are associated with a range of complex ethical difficulties, some of which were briefly discussed in Section 2.1. In this section, these ethical difficulties will be explored further to examine the distinction between public and private places, informed consent, confidentiality, anonymity, debriefing and deception. Researchers should also consider the status of forum data, which is discussed in Section 1.4. There are no clear-cut answers to untangle the dilemmas surrounding these issues, so researchers have a responsibility to make the best decision they can on a

case-by-case basis. Researchers should also be mindful of the preferences and guidelines of any local ethical review panels which may consider this type of work.

It should be noticed that the ethical principles described here might not apply to all research paradigms. In particular, critical social psychology researchers may find it paradoxical that publically-available messages addressed to online strangers are treated by researchers as if they were private communication. Adopting such an extreme ethical position may preclude using forum data for discourse analytic research, making it seemingly impossible to conduct critical research into important social topics because such studies could be deemed ethically dubious (Jowett, 2015). In such studies, a fundamentally different ethical position may need to be taken on issues of privacy and anonymity.

6.1 Public and private places

The distinction between public (openly accessible) and private (walled off) sources of data on the Internet has important implications for researchers (Eysenbach & Till, 2001; Robinson, 2001). Some forums ('open' forums) are freely accessible so that anybody can read messages, while others ('closed' forums) are password-protected and can only be accessed by registered members.

Determining whether a forum perceives itself as being public or private can be complicated (Roberts, 2015). When messages are posted in open forums, some members may believe their postings are private without realising that they can be freely read by anybody who chooses to access them (see Lomborg, 2013). Password-protected forums are usually private, forums with a smaller number of members are more likely to be private than those with a larger membership base, and forums intended for a specific target audience such as children may also be private (Eysenbach & Till, 2001). To determine if a forum is public or private, researchers should consider whether it is openly accessible, how it is perceived by members, the sensitivity of discussions, and the intended audience (Roberts, 2015). As explained in Section 2.1, if a forum is private then the forum's moderators or administrators should be contacted for advice on how to proceed.

If a forum is public then researchers may still face ethical difficulties. Some researchers may be tempted to argue that reading or downloading messages from a public forum is comparable to observing people in a public setting, but Zimmer (2010) explains why this analogy is misleading. When a researcher observes people in a public place, random chance determines who happens to be present at the same time as the researcher. It is unlikely that the researcher will be able to observe everybody who is physically present so they will probably have to focus their attention on a small subset of those people. Guesswork may be needed to estimate the age, gender and other characteristics of the individuals under observation, so the data collected will necessarily be imperfect and incomplete. When a researcher is reading or downloading messages from a public forum, the situation is completely different. The researcher can simultaneously collect data about all forum members, they can study archives of messages that were written before the researcher discovered that forum, and they can potentially collect a wide range of precise data about those individuals without the need for guesswork.

Researchers should carefully consider whether a forum is likely to be public or private, and the difficult ethics of collecting data from public forums, before commencing with a study.

6.2 Informed consent

Informed consent involves ensuring that participants are aware of the purpose of the study, what kind of data will be collected, how the data will be used, what potential benefits might arise, and what risks are involved. Forum members have mixed opinions regarding whether informed consent is needed for message analysis studies: some members are happy for their messages to be used in research without their knowledge or permission, while others want researchers to obtain informed consent before using their data (Bond et al., 2013).

There are numerous challenges with obtaining informed consent for this type of research. Some individuals might be impossible to contact if they are no longer active within a forum (Roberts, 2015), and individuals may react badly if permission is requested to use their messages in research

(Eysenbach & Till, 2001; Hudson & Bruckman, 2004). Consent requests may produce feelings of resentment or distrust, they can disrupt the conversations that the researcher hoped to study, and if an individual responds by leaving the forum then it may prevent that person from obtaining the help they needed.

Researchers have a responsibility to avoid harming both individual members and the forum. When collecting data from public forums researchers may need to consider waiving the need for informed consent if the risks involved with obtaining consent are greater than the protection it offers, but consent should always be obtained when collecting data from private forums (Robinson, 2001).

If informed consent has been obtained then the researcher could explore the question of anonymity with participants. Researchers may assume that anonymity is the desired default position (see Section 6.4), but there is a possibility that some individuals may prefer to take ownership of their data by waiving the right to anonymity. There are risks and dilemmas associated with waiving anonymity, which are explored further by Giordano, O'Reilly, Taylor, and Dogra (2007). The researcher could also consider offering participants the opportunity to provide feedback on how their data has been used and analysed. This also has associated risks, for example the participants might be disinterested in reading a complicated research report outlining the findings, and it might be uncertain how to integrate their feedback into the finished work particularly if they disagree with any aspect of the results or conclusions (Kaiser, 2009).

6.3 Confidentiality

Researchers have an ethical duty to keep data about participants or collected from participants confidential. This means that data should not be disclosed to anyone who is not directly involved with the study, and if the results are published then participants can expect that any data relating to them should be anonymous so they cannot be identified. Confidentiality protocols typically mean that all data relating to a study should be securely stored, and personally identifiable data should be stored separately to anonymous research data. Ensuring confidentiality poses a number of difficulties when

conducting online research because of how the Internet works, which may make absolute confidentiality impossible.

In particular, using e-mail to discuss research or share data with collaborators could breach confidentiality. First, e-mails pass unencrypted through multiple servers before they are delivered to the intended recipient, and can potentially be viewed by others in a similar way to sending information on the back of a postcard (Kreindler, 2008). Second, e-mails containing sensitive information could be accidentally sent or forwarded to unintended recipients (Granberry, 2007). Third, some employers monitor the computer use of employees, ranging from monitoring e-mails through to tracking individual keystrokes as they are typed on a computer keyboard (Friedman & Reed, 2007). Any of these could lead to information being accidentally or unknowingly disclosed to third parties and hence violate the confidentiality of participants.

The use of cloud storage by researchers also has the potential to violate participant confidentiality. Cloud storage involves storing documents on a remote server so that they can be accessed from multiple devices or shared between collaborators. This leads to concerns regarding who could potentially access those files, for example hackers might maliciously gain access to the account, links to directly access files could be accidentally shared with unintended recipients, or privacy settings might inadvertently make documents publically searchable and viewable (Messier, 2014; Moura & Hutchison, 2016).

6.4 Anonymity

Whereas confidentiality involves managing private information, anonymity refers to obscuring or removing potentially identifiable information about participants and research sites, and can be a useful tool for maintaining confidentiality (Tilley & Woodthorpe, 2011).

In message analysis studies, the biggest threat to anonymity comes from disseminating the results (van den Hoonaard, 2003). Researchers typically include quoted extracts from messages as evidence

for their findings (e.g. Attard & Coulson, 2012; Malik & Coulson, 2008), and various approaches have been used to attribute these quotes. Quotes can be reported anonymously without identifying either the participant or forum they were taken from, participants can be given pseudonyms, or quotes can be attributed to the participants' real identity (Roberts, 2015). If data is taken from an 'open' forum then the use of quotes could threaten the anonymity of participants. It is often possible to use a search engine such as Google to trace quotes back to the person who wrote it, compromising their anonymity regardless of whether or not a pseudonym has been used (Eysenbach & Till, 2001; Kraut et al., 2004).

Consequently, researchers should check all quotes using a search engine to make sure they cannot be traced, or take alternative steps to protect participants' anonymity by paraphrasing quotes or using composite accounts that convey the intended meaning of multiple messages without directly quoting them (Roberts, 2015). Researchers should also avoid reporting any distinctive or personally-identifiable information that could potentially lead to participants being recognised from their descriptions using deductive disclosure (Kaiser, 2009; van den Hoonaard, 2003).

Regardless of the steps taken to protect the identity of participants, the use of unsolicited data from the Internet is associated with another ethical dilemma. Taking an individual's personal information from its intended domain of the forum where it was posted, and then storing it in a database or analysing it as part of a research study without that person's knowledge or consent, could be an affront to participants' human dignity and their ability to control how their information is used (Zimmer, 2010).

6.5 Debriefing

Debriefing participants in message analysis studies involves many of the same challenges and difficulties as obtaining informed consent (Section 6.2). If a study has been conducted without the knowledge or consent of participants, then those individuals will not be aware that their data has been used for research purposes and hence they cannot be debriefed afterwards.

Kraut et al. (2004) recommend that participants should be debriefed if deception is used during data collection, but there is a risk of harming participants by doing this. If a researcher joins a forum under false pretences to collect data that would otherwise be impossible to access (e.g. Brotsky & Giles, 2007), then debriefing participants to explain the deception is likely to produce feelings of anger and betrayal. Notifying forum members that a researcher has been collecting data under false pretences could lead to increased suspicions and an erosion of trust among members. The potential harm to individuals and the forum may be greater than the protection that debriefing is intended to provide, and hence researchers may need to consider waiving debriefing when conducting this type of research.

6.6 Deception

The concept of deception can be difficult to define. Withholding information from participants is not normally regarded as a form of deception, instead deception refers to deliberately and explicitly providing misleading or erroneous information to participants – in effect, lying to them (Hertwig & Ortmann, 2008). Deception violates a participant's ability to give informed consent to taking part in research because an individual cannot consent if they have been misled about the study and any risks that might be involved in taking part (Sell, 2008). Consequently, deception should only be used under exceptional circumstances where it is necessary for the integrity of the study.

Normally, observational studies require the consent of the people being observed unless the research is conducted in a public place where individuals would expect to be observed by strangers. However, as discussed in Section 6.1, it is uncertain whether the Internet should be regarded as a public or private place and there is considerable doubt over whether forum members can be observed in a way that is analogous to public observation.

If a researcher knowingly misrepresents him or herself by posing as an individual with a health condition in order to join a forum under false pretences (e.g. Brotsky & Giles, 2007) then this is likely

to be regarded as a form of deception. Whether collecting data from public forums without the consent of participants constitutes deception is less clear, and may require careful thought on the part of researchers before proceeding with a study.

7 Summary

Forums can help individuals with the self-care and self-management of many health challenges. By analysing the user-generated content of these forums, researchers can examine many different issues including the topics that members discuss, what kind of social support they exchange, how they overcome problems, and obtain new insights into the needs, opinions and experiences of individuals facing these health challenges. Forums have a hierarchical tree-like structure comprising boards, discussion threads and individual messages, with emoticons and other features to help convey information more effectively. Researchers can download data from existing forums or create their own bespoke forums, with both approaches having their own advantages and disadvantages depending on the nature of the research question. Sampling takes place on four levels comprising forum characteristics, user characteristics, message characteristics and temporal focus. The data are prepared for analysis by importing the selected messages into a text document, and can be analysed using a range of quantitative and qualitative techniques including content analysis, thematic analysis and discourse analysis.

Forum studies have become increasingly important and provide researchers with a different type of data to what is available through other research approaches and datasets. Many important decisions need to be made when designing and carrying out a study, and researchers need to be able to provide a clearly justified rationale for the decisions taken during each step of the research process. Forum research is associated with a range of ethical dilemmas, particularly in relation to informed consent, the participant's right to withdraw from a study, protecting their anonymity, and the distinction between public and private places on the Internet. The researcher's primary responsibility is to the forum, and it is vital for researchers to ensure that they do not harm either the forum or its members.

Particular care is required when preparing the results of a study for dissemination, which can harm the forum and/or individual members by violating their anonymity (van den Hoonaard, 2003).

Researchers typically include quoted extracts from forum messages as evidence for their findings (e.g. Attard & Coulson, 2012; Malik & Coulson, 2008). It is often possible to type these quotes into an Internet search engine such as Google, which can be used to identify the forum where it was posted and the individual member who wrote it, compromising the anonymity of both the forum and the participant (Eysenbach & Till, 2001; Kraut et al., 2004). Consequently, researchers should check all quotes using a search engine to make sure they cannot be traced, or take alternative steps to protect participants' anonymity by paraphrasing quotes or using composite accounts that convey the intended meaning of multiple messages without directly quoting them (Roberts, 2015). Researchers should also avoid reporting any distinctive or personally-identifiable information that could potentially lead to participants being recognised from their descriptions using deductive disclosure (Kaiser, 2009; van den Hoonaard, 2003).

Forums represent a vast and largely untapped source of naturalistic data on the Internet. By following the principles and guidelines described in this paper, researchers and others working in health-related fields will be able to successfully plan, develop and complete research projects that use forums as a data source while navigating the pitfalls and ethical difficulties associated with this type of research. Above all else, researchers must prioritise the needs of the forum and its members over their own needs and the needs of the study being conducted.

References

American Psychological Association 2017, 'Ethical principles of psychologists and code of conduct', viewed 13 July 2017, http://www.apa.org/ethics/code/ethics-code-2017.pdf.

Armstrong, N, Koteyko, N & Powell, J 2012, 'Oh dear, should I really be saying that on here?: issues of identity and authority in an online diabetes community', *Health*, vol. 16, pp. 347-365.

Attard, A & Coulson, NS 2012, 'A thematic analysis of patient communication in Parkinson's disease online support group discussion forums', *Computers in Human Behavior*, vol. 28, pp. 500-506.

Ballinger, C, Yardley, L & Payne, S 2004, 'Observation and action research', in DF Marks & L Yardley (eds.), *Research Methods for Clinical and Health Psychology*, SAGE Publications, London, pp. 102-121.

Bane, CMH, Haymaker, CMB & Zinchuk, J 2005, 'Social support as a moderator of the big-fish-in-a-little-pond effect in online self-help support groups', *Journal of Applied Biobehavioral Research*, vol. 10, pp. 239-261.

Bond, SC, Ahmed, HO, Hind, M, Thomas, B & Hewitt-Taylor, J 2013, 'The conceptual and practical ethical dilemmas of using health discussion board posts as research data', *Journal of Medical Internet Research*, vol. 15, pp. e112.

Brady, E, Segar, J & Sanders, C 2016, '"I always vet things": navigating privacy and the presentation of self on health discussion boards among individuals with long-term conditions', *Journal of Medical Internet Research*, vol. 18, pp. e274.

Braun, V & Clarke, V 2006, 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, vol. 3, pp. 77-101.

British Psychological Society 2017, 'Ethics guidelines for internet-mediated research', viewed 13 July 2017, https://beta.bps.org.uk/sites/beta.bps.org.uk/files/Policy%20-%20Files/Ethics%20Guidelines%20for%20Internet-mediated%20Research%20%282017%29.pdf.

Brotsky, SR & Giles, D 2007, 'Inside the "pro-ana" community: a covert online participant observation', *Eating Disorders*, vol. 15, pp. 93-109.

Carter, I 2009, 'Positive and negative experiences of parents involved in online self-help groups for autism', *Journal on Developmental Disabilities*, vol. 15, pp. 44-52.

Charmaz, K 2003, 'Grounded theory', in JA Smith (ed.), *Qualitative Psychology: A Practical Guide to Research Methods.* Sage Publications, London, pp. 81-110.

Chen, AT 2014, 'What's in a virtual hug? a transdisciplinary review of methods in online health discussion forum research', *Library & Information Science Research*, vol. 36, pp. 120-130.

Coulson, NS 2013, 'How do online patient support communities affect the experience of inflammatory bowel disease? an online survey', *JRSM Short Reports*, vol. 4, pp. 1-8.

Coulson, NS, Buchanan, H & Aubeeluck, A 2007, 'Social support in cyberspace: a content analysis of communication within a Huntington's disease online support group', *Patient Education and Counseling*, vol. 68, pp. 173-178.

Coulson, NS & Greenwood, N 2012, 'Families affected by childhood cancer: an analysis of the provision of social support within online support groups', *Child: Care, Health and Development*, vol. 38, pp. 870-877.

Coulson, NS & Shaw, RL 2013, 'Nurturing health-related online support groups: exploring the experiences of patient moderators', *Computers in Human Behavior*, vol. 29, pp. 1695-1701.

Coulson, NS, Smedley, R, Bostock, S, Kyle, SD, Gollancz, R, Luik, AI, Hames, P & Espie, CA 2016, 'The pros and cons of getting engaged in an online social community embedded within digital cognitive behavioral therapy for insomnia: survey among users', *Journal of Medical Internet Research*, vol. 18, pp. e88.

Coursaris, CK & Liu, M 2009, 'An analysis of social support exchanges in online HIV/AIDS self-help groups', *Computers in Human Behavior*, vol. 25, pp. 911-918.

Cummings, JN, Sproull, L & Kiesler, SB 2002, 'Beyond hearing: where the real-world and online support meet', *Group Dynamics: Theory, Research, and Practice*, vol. 6, pp. 78-88.

Drentea, P & Moren-Cross, JL 2005, 'Social capital and social support on the web: the case of an internet mother site', *Sociology of Health & Illness*, vol. 27, pp. 920-943.

Eichhorn, KC 2008, 'Soliciting and providing social support over the internet: an investigation of online eating disorder support groups', *Journal of Computer-Mediated Communication*, vol. 14, pp. 67-78.

Elo, S & Kyngäs, H 2008, 'The qualitative content analysis process', *Journal of Advanced Nursing*, vol. 62, pp. 107-115.

Elwell, L, Grogan, S & Coulson, NS 2011, 'Adolescents living with cancer: the role of computer-mediated support groups', *Journal of Health Psychology*, vol. 16, pp. 236-248.

Esquivel, A, Meric-Bernstam, F & Bernstam, EV 2006, 'Accuracy and self correction of information received from an internet breast cancer list: content analysis', *British Medical Journal*, vol. 332, pp. 939-942.

Evans, M, Donelle, L & Hume-Loveland, L 2012, 'Social support and online postpartum depression discussion groups: a content analysis', *Patient Education and Counseling*, vol. 87, pp. 405-410.

Eysenbach, G & Till, JE 2001, 'Ethical issues in qualitative research on internet communities', *British Medical Journal*, vol. 323, pp. 1103-1105.

Finfgeld-Connett, D 2015, 'Twitter and health science research', *Western Journal of Nursing Research*, vol. 37, pp. 1269-1283.

Finn, J 1999, 'An exploration of helping processes in an online self-help group focusing on issues of disability', *Health & Social Work*, vol. 24, pp. 220-231.

Freeman, LC 2004, *The Development of Social Network Analysis: A Study in the Sociology of Science*, Empirical Press, Vancouver, Canada.

Friedman, BA & Reed, LJ 2007, 'Workplace privacy: employee relations and legal implications of monitoring employee e-mail use', *Employee Responsibilities and Rights Journal*, vol. 19, pp. 75-83.

Galegher, J, Sproull, L & Kiesler, S 1998, 'Legitimacy, authority, and community in electronic support groups', *Written Communication*, vol. 15, pp. 493-530.

Gavin, J, Rodham, K & Poyer, H 2008, 'The presentation of "pro-anorexia" in online group interactions', *Qualitative Health Research*, vol. 18, pp. 325-333.

Giordano, J, O'Reilly, M, Taylor, H & Dogra, N 2007, 'Confidentiality and autonomy: the challenge(s) of offering research participants a choice of disclosing their identity', *Qualitative Health Research*, vol. 17, pp. 264-275.

Granberry, N 2007, 'Email-from "to" to "send"'. *AAOHN journal : official journal of the American Association of Occupational Health Nurses*, vol. 55, pp. 127-30.

Gustafson, DH, Hawkins, RP, Pingree, S, McTavish, FM, Arora, NK, Mendenhall, J, Cella, DF, Serlin, RC, Apantaku, FM, Stewart, J & Salner, A 2001, 'Effect of computer support on younger women with breast cancer', *Journal of General Internal Medicine*, vol. 16, pp. 435-445.

Gustafson, DH, McTavish, FM, Hawkins, RP, Pingree, S, Arora, NK, Mendenhall, J & Simmons, GE 1998, 'Computer support for elderly women with breast cancer', *JAMA: The Journal of the American Medical Association*, vol. 280, pp. 1305.

Hertwig, R & Ortmann, A 2008, 'Deception in social psychological experiments: two misconceptions and a research agenda', *Social Psychology Quarterly*, vol. 71, pp. 222-227.

Hodges, BD, Kuper, A & Reeves, S 2008, 'Discourse analysis', *British Medical Journal*, vol. 337, pp. a879.

Holbrey, S & Coulson, NS 2013, 'A qualitative investigation of the impact of peer to peer online support for women living with polycystic ovary syndrome', *BMC Women's Health,* vol. 13, pp. 51.

Holtz, P, Kronberger, N & Wagner, W 2012, 'Analyzing internet forums: a practical guide', *Journal of Media Psychology: Theories, Methods, and Applications*, vol. 24, pp. 55-66.

Hudson, JM & Bruckman, A 2004, '"Go away": participant objections to being studied and the ethics of chatroom research', *The Information Society*, vol. 20, pp. 127-139.

Hwang, KO, Farheen, K, Johnson, CW, Thomas, EJ, Barnes, AS & Bernstam, EV 2007, 'Quality of weight loss advice on internet forums', *The American Journal of Medicine*, vol. 120, pp. 604-609.e2.

Hwang, KO, Ottenbacher, AJ, Green, AP, Cannon-Diehl, MR, Richardson, O, Bernstam, EV & Thomas, EJ 2010, 'Social support in an internet weight loss community', *International Journal of Medical Informatics*, vol. 79, pp. 5-13.

Im, E-O & Chee, W 2006, 'An online forum as a qualitative research method', *Nursing Research*, vol. 55, pp. 267-273.

Jowett, A 2015, 'A case for using online discussion forums in critical psychological research', *Qualitative Research in Psychology*, vol. 12, pp. 287-297.

Joyce, E & Kraut, RE 2006, 'Predicting continued participation in newsgroups', *Journal of Computer-Mediated Communication,* vol. 11, pp. 723-747.

Kaiser, K 2009, 'Protecting respondent confidentiality in qualitative research', *Qualitative Health Research,* vol. 19, pp. 1632-1641.

Kraut, R, Olson, J, Banaji, MR, Bruckman, A, Cohen, J & Couper, M 2004, 'Psychological research online - report of board of scientific affairs' advisory group on the conduct of research on the Internet', *American Psychologist,* vol. 59, pp. 105-117.

Kreindler, DM 2008, 'Email security in clinical practice: ensuring patient confidentiality', *Open Medicine,* vol. 2, pp. e54-e59.

Levin, KA 2006, 'Study design III: cross-sectional studies', *Evidence-Based Dentistry,* vol. 7, pp. 24-25.

Loader, BD, Muncer, S, Burrows, R, Pleace, N & Nettleton, S 2002, 'Medicine on the line? computer-mediated social support and advice for people with diabetes', *International Journal of Social Welfare,* vol. 11, pp. 53-65.

Lomborg, S 2013, 'Personal internet archives and ethics', *Research Ethics,* vol. 9, pp. 20-31.

Luke, DA & Harris, JK 2007, 'Network analysis in public health: history, methods, and applications', *Annual Review of Public Health,* vol. 28, pp. 69-93.

Malik, SH & Coulson, NS 2008, 'The male experience of infertility: a thematic analysis of an online infertility support group bulletin board', *Journal of Reproductive and Infant Psychology,* vol. 26, pp. 18-30.

Malik, SH & Coulson, NS 2010, 'Coping with infertility online: an examination of self-help mechanisms in an online infertility support group', *Patient Education and Counseling,* vol. 81, pp. 315-318.

Malik, SH & Coulson, NS 2011, 'The therapeutic potential of the internet exploring self-help processes in an internet forum for young people with inflammatory bowel disease', *Gastroenterology Nursing,* vol. 34, pp. 439-448.

Malinen, S 2015, 'Understanding user participation in online communities: a systematic literature review of empirical studies', *Computers in Human Behavior,* vol. 46, pp. 228-238.

Matzat, U & Rooks, G 2014, 'Styles of moderation in online health and support communities: an experimental comparison of their acceptance and effectiveness', *Computers in Human Behavior*, vol. 36, pp. 65-75.

McKenna, KYA, Green, AS & Gleason, MEJ 2002, 'Relationship formation on the internet: what's the big attraction?', *Journal of Social Issues*, vol. 58, pp. 9-31.

Messier, R 2014, 'Chapter 4 - Storage in the cloud', *Collaboration with Cloud Computing*. Syngress, Boston, USA.

Mo, PKH & Coulson, NS 2010, 'Living with HIV/AIDS and use of online support groups', *Journal of Health Psychology*, vol. 15, pp. 339-350.

Mo, PKH & Coulson, NS 2012, 'Developing a model for online support group use, empowering processes and psychosocial outcomes for individuals living with HIV/AIDS', *Psychology & Health*, vol. 27, pp. 445-459.

Moura, J & Hutchison, D 2016, 'Review and analysis of networking challenges in cloud computing', *Journal of Network and Computer Applications*, vol. 60, pp. 113-129.

NESH 2014, 'Ethical guidelines for internet research', viewed 16 July 2017, https://www.etikkom.no/globalassets/documents/english-publications/ethical-guidelines-for-internet-research.pdf.

Nonnecke, B & Preece, J 2000, 'Lurker demographics: counting the silent', *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, ACM, The Hague, The Netherlands.

Nosek, BA, Banaji, MR & Greenwald, AG 2002, 'E-research: ethics, security, design, and control in psychological research on the internet', *Journal of Social Issues*, vol. 58, pp. 161-176.

Petrovčič, A, Vehovar, V & Žiberna, A 2012, 'Posting, quoting, and replying: a comparison of methodological approaches to measure communication ties in web forums', *Quality & Quantity*, vol. 46, pp. 829-854.

Pfeil, U & Zaphiris, P 2010, 'Applying qualitative content analysis to study online support communities', *Universal Access in the Information Society*, vol. 9, pp. 1-16.

Preece, J, Nonnecke, B & Andrews, D 2004, 'The top five reasons for lurking: improving community experiences for everyone', *Computers in Human Behavior*, vol. 20, pp. 201-223.

Ravert, RD, Hancock, MD & Ingersoll, GM 2004, 'Online forum messages posted by adolescents with type 1 diabetes', *Diabetes Educator*, vol. 30, pp. 827-834.

Ridings, CM & Gefen, D 2004, 'Virtual community attraction: why people hang out online', *Journal of Computer-Mediated Communication*, vol. 10, pp. 00-00.

Roberts, LD 2015, 'Ethical issues in conducting qualitative research in online communities', *Qualitative Research in Psychology*, vol. 12, pp. 314-325.

Robinson, KM 2001, 'Unsolicited narratives from the internet: a rich source of qualitative data', *Qualitative Health Research*, vol. 11, pp. 706-714.

Rodham, K, Gavin, J, Coulson, NS, Watts, LA & Heath, H 2013, 'Let me introduce myself: the role of first posts in establishing a common-identity community in an online support forum', *Conference on Psychology and Social Harmony*, Beijing, 24th May 2013, Scientific Research Publishing, USA.

Rodham, K, Gavin, J, Lewis, SP, St Denis, JM & Bandalli, P 2013, 'An investigation of the motivations driving the online representation of self-injury: a thematic analysis', *Archives of Suicide Research*, vol. 17, pp. 173-183.

Rodham, K, McCabe, C, & Blake, D 2009, 'Seeking support: an interpretative phenomenological analysis of an internet message board for people with complex regional pain syndrome', *Psychology & Health*, vol. 24, pp. 619-634.

Sanders, C, Rogers, A, Gardner, C & Kennedy, A 2011, 'Managing 'difficult emotions' and family life: exploring insights and social support within online self-management training', *Chronic Illness*, vol. 7, pp. 134-146.

Sell, J 2008, 'Introduction to deception debate', *Social Psychology Quarterly*, vol. 71, pp. 213-214.

Smedley, R, Coulson, N, Gavin, J, Rodham, K & Watts, L 2015, 'Online social support for complex regional pain syndrome: a content analysis of support exchanges within a newly launched discussion forum', *Computers in Human Behavior*, vol. 51, pp. 53-63.

Smedley, RM, Coulson, NS 2017, 'A thematic analysis of messages posted by moderators within health-related asynchronous online support forums', *Patient Education and Counseling*, vol. 100, pp. 1688-1693.

Smedley, RM, Coulson, NS, Gavin, J, Rodham, K & Watts, L 2016, 'Engagement in a newly launched online support community for complex regional pain syndrome: membership growth, header analysis and introductory messages', *International Journal of Web Based Communities*, vol. 12, pp. 147-164.

Smith, JA & Osborn, M 2003, 'Interpretative phenomenological analysis', in JA Smith (ed.), *Qualitative Psychology: A Practical Guide to Research Methods.* Sage Publications, London, pp. 51-80.

Stieger, S & Göritz, AS 2006, 'Using instant messaging for internet-based interviews', *CyberPsychology & Behavior*, vol. 9, pp. 552-559.

Thackeray, R, Crookston, BT & West, JH 2013, 'Correlates of health-related social media use among adults', *Journal of Medical Internet Research*, vol. 15, pp. e21.

Tichon, JG & Shapiro, M 2003, 'The process of sharing social support in cyberspace', *CyberPsychology & Behavior*, vol. 6, pp. 161-170.

Tilley, L & Woodthorpe, K 2011, 'Is it the end for anonymity as we know it? A critical examination of the ethical principle of anonymity in the context of 21st century demands on the qualitative researcher', *Qualitative Research*, vol. 11, pp. 197-212.

Van Den Hoonaard, WC 2003, 'Is anonymity an artifact in ethnographic research?', *Journal of Academic Ethics*, vol. 1, pp. 141-151.

Van Der Houwen, K, Stroebe, M, Schut, H, Stroebe, W & Van Den Bout, J 2010, 'Online mutual support in bereavement: an empirical examination', *Computers in Human Behavior*, vol. 26, pp. 1519-1525.

Van Uden-Kraan, CF, Drossaert, CHC, Taal, E, Lebrun, CEI, Drossaers-Bakker, KW, Smit, WM, Seydel, ER & Van De Laar, MAFJ 2008, 'Coping with somatic illnesses in online support groups: do the feared disadvantages actually occur?', *Computers in Human Behavior*, vol. 24, pp. 309-324.

Van Uden-Kraan, CF, Drossaert, CHC, Taal, E, Seydel, ER & Van De Laar, MAFJ 2009, 'Participation in online patient support groups endorses patients' empowerment', *Patient Education and Counseling*, vol. 74, pp. 61-69.

Van Uden-Kraan, CF, Drossaert, CHC, Taal, E, Shaw, BR, Seydel, ER & Van De Laar, MAFJ 2008, 'Empowering processes and outcomes of participation in online support groups for patients with breast cancer, arthritis, or fibromyalgia', *Qualitative Health Research*, vol. 18, pp. 405-417.

Voelpel, SC, Eckhoff, RA & Förster, J 2008, 'David against Goliath? group size and bystander effects in virtual knowledge sharing', *Human Relations*, vol. 61, pp. 271-295.

Wen, K-Y, McTavish, FM, Kreps, G, Wise, M & Gustafson, DH 2011, 'From Diagnosis to Death: A case study of coping with breast cancer as seen through online discussion group messages', *Journal of Computer-Mediated Communication*, vol. 16, pp. 331-361.

Wiedemann, G 2013, 'Opening up to big data: computer-assisted analysis of textual data in social sciences', *Forum: Qualitative Social Research*, vol. 14, art. 23.

Wilkinson, D & Thelwall, M 2011, 'Researching personal information on the public web: methods and ethics', *Social Science Computer Review*, vol. 29, pp. 387-401.

Willig, C 2003, 'Discourse analysis', in JA Smith (ed.), *Qualitative Psychology: A Practical Guide to Research Methods.* Sage Publications, London, pp. 159-183.

Winzelberg, A 1997, 'The analysis of an electronic support group for individuals with eating disorders', *Computers in Human Behavior*, vol. 13, pp. 393-407.

Yardley, L & Murray, M 2004, 'Qualitative analysis of talk and text: discourse and narrative analysis', in DF Marks & L Yardley (eds.), *Research Methods for Clinical and Health Psychology*, Sage Publications, London, pp. 90-101.

Yechiam, E & Barron, G 2003, 'Learning to ignore online help requests', *Computational & Mathematical Organization Theory*, vol. 9, pp. 327-339.

Zimmer, M 2010, '"But the data is already public": on the ethics of research in Facebook', *Ethics and Information Technology*, vol. 12, pp. 313-325.

Zrebiec, JF & Jacobson, AM 2001, 'What attracts patients with diabetes to an internet support group? a 21-month longitudinal website study', *Diabetic Medicine*, vol. 18, pp. 154-158.

Acknowledgements

The authors are grateful to Dr Heather Buchanan, Dr Laura Condon and Karen Shepherd for their helpful and constructive feedback on a previous draft of this paper.

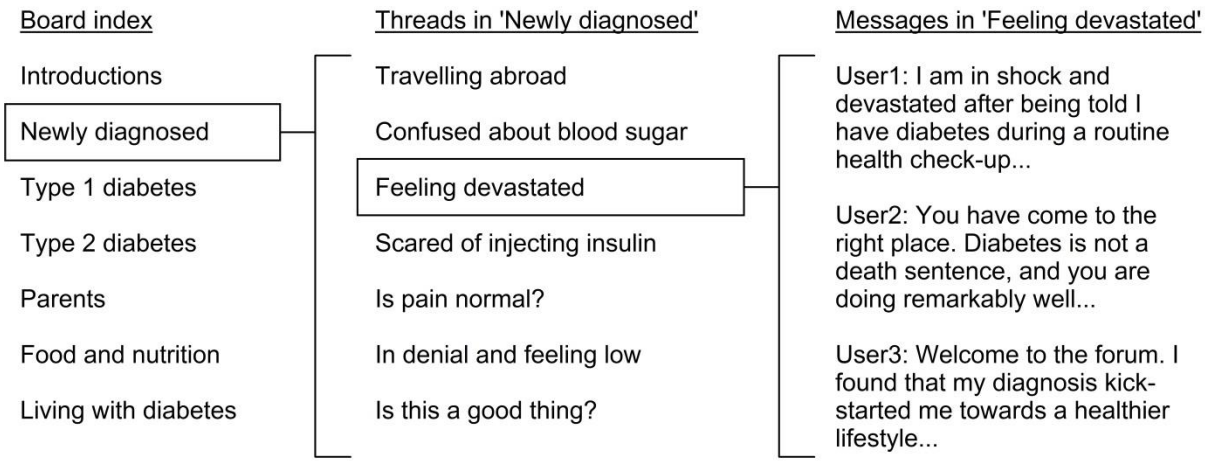
Table 1: Potential benefits and disadvantages of working with user-generated content.

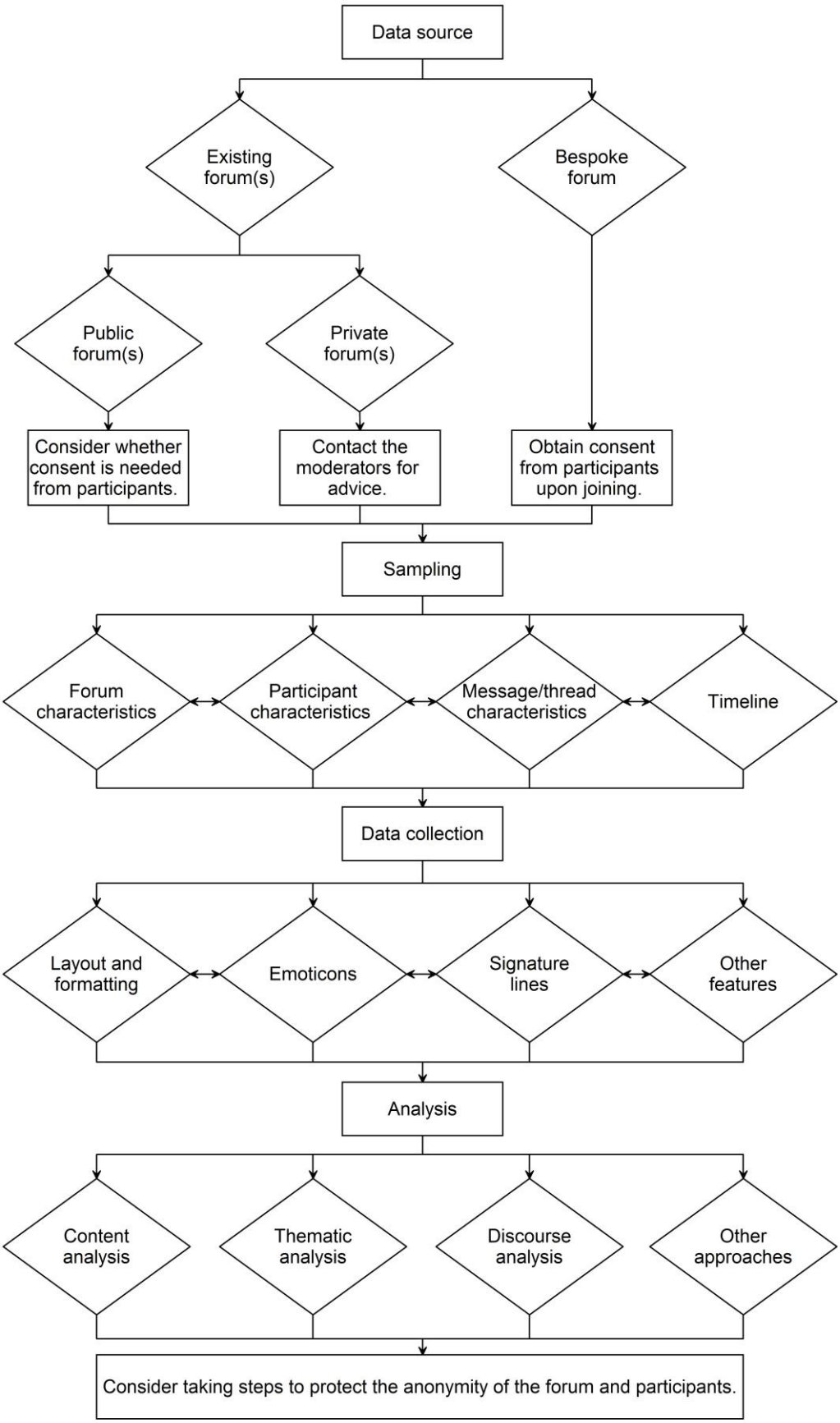
Benefits	Disadvantages
Volume of potential data The number of online forums continues to increase as does the number of people choosing to participate through posting messages. There is likely to be >1 million online forums devoted to health and illness topics, yielding several million conversation threads and messages.	*Data overload* Some forums have high levels of daily activity, and can generate such a large number of messages that it may not be possible for researchers to analyse all of them.
Naturalistic conversations Discussions between members within forums can be considered naturalistic since the researcher is typically an observer who does not influence the online interaction.	*Lack of social cues* The absence of social cues within messages may make it difficult for researchers to understand the intended meaning of messages, particularly if they are read out-of-context.
Accessing hard-to-reach populations Online forums can be used to research individuals who might otherwise be difficult to reach, such as those living with a rare illness or who have not sought professional help for a health issue.	*Membership uncertainty* Some members will be diagnosed with a health condition, some could be relatives/carers, and others might have joined dishonestly. They could be employees with a vested interest in promoting a particular drug or treatment.
Easy to access Messages posted in online forums can be read from anywhere in the world using an Internet-enabled computing device together with web browser software, without needing any special software or equipment.	*Unequal access* Internet access is not distributed equally across the population. Some people, such as those with disabilities, are less likely to be Internet users than others.
Ability to search through archives Forum messages are automatically stored in easily-accessible archives. Researchers can search through these archived collections of previously-posted discussions to identify messages that are relevant to the research topic.	*Overrepresentation of negative experiences* Some individuals may use forums when they are feeling particularly bad, and then stop posting messages when they feel better. This may lead to a bias, where some issues are overrepresented and others are underrepresented.
Public data Using public forums (see Section 6.1) gives researchers direct access to participants' messages, although in some cases the moderators may act as a 'gatekeeper' whose permission should be sought before carrying out any data collection or analysis.	*Lack of demographic information* The anonymous nature of forum profiles may make it difficult for researchers to obtain accurate sociodemographic data about forum members. When an individual reveals their age or gender in the content of messages, this information is not guaranteed to be accurate.
Reduced costs Studies may be cheaper and quicker to conduct by avoiding the time and costs associated with recruiting participants, conducting surveys or interviews, and transcribing data.	*Ethical concerns* There are major ethical concerns associated with analysing forum messages.

Figure captions

Figure 1: The hierarchical, tree-like structure of forums.

Figure 2: Key decisions in the research process.


Board index
Introductions
Newly diagnosed
Type 1 diabetes
Type 2 diabetes
Parents
Food and nutrition
Living with diabetes
Threads in 'Newly diagnosed'
Travelling abroad
Confused about blood sugar
Feeling devastated
Scared of injecting insulin
Is pain normal?
In denial and feeling low
Is this a good thing?
Messages in 'Feeling devastated'
User1: I am in shock and devastated after being told I have diabetes during a routine health check-up...
User2: You have come to the right place. Diabetes is not a death sentence, and you are doing remarkably well...
User3: Welcome to the forum. I found that my diagnosis kick-started me towards a healthier lifestyle...


Data source
Existing forum(s)
Bespoke forum
Public forum(s)
Private forum(s)
Consider whether consent is needed from participants.
Contact the moderators for advice.
Obtain consent from participants upon joining.
Sampling
Forum characteristics
Participant characteristics
Message/thread characteristics
Timeline
Data collection
Layout and formatting
Emoticons
Signature lines
Other features
Analysis
Content analysis
Thematic analysis
Discourse analysis
Other approaches
Consider taking steps to protect the anonymity of the forum and participants.