Making sense of patients' internet forums:

a systematic method using discourse analysis

INTRODUCTION

A survey of access to the internet in the UK conducted in 2008 revealed that the internet is used by 79% of men and 75% of women of all ages including 72% of people aged 55–64 years and 32% of people aged ≥65 years.[1] Internet data that is freely and publicly accessible are now being used for research purposes.[2,3]

Internet communities offer an increasingly important source of information expressed openly by individuals. In particular, the internet offers access to hard-to-reach groups who are often excluded (or exclude themselves) from traditional research studies.

DISCOURSE ANALYSIS

Discourse analysis (DA), an approach to analysing naturally occurring language, is a technique that is particularly suited to examining internet data.[4,5] DA is pertinent to health care for it has the potential to reveal the dimensions of health beliefs, the doctor–patient relationships and the dissemination of health information. The focus of DA is on communicative behaviour.[6] Within internet forums communicative behaviour is the manner in which individuals communicate through written text.

At a basic level, interrogation with linguistic analysis software reveals word frequency. Frequency is a simple way to identify problems and issues. We can look at how patterns of words colocate together and uncover associations between words (that is, concordances) that may provide insights into people, groups, and ideas. With the development of computers, linguistics has become involved using concordancing where keywords from a body of text, often termed a corpus, are highlighted in their surrounding context.

Search engines like Google and Yahoo are, at heart, simple concordancers in their browsing functions. They offer the casual user the opportunity to search a very large database for examples of a single word or phrase (one's own name, for instance). Linguistic concordancing offers the opportunity for more sophisticated language-based analyses. The techniques are very widely-used in language studies; for example, in forensic linguistics (assessing whether a document has been forged or to examine witness statements). The approach has been applied to healthcare studies since the 1990s, though not yet with great frequency.[7–11]

Box 1. Biases of traditional research approaches versus internet forum-based discourse analysis

Traditional research approaches	Internet forum-based discourse analysis
• Participants signing-up to studies come from a selected population (for example, high-adherers)	• Participants using internet (selection bias is internet users)
• Patients aware that their behaviour is being monitored (reactivity and self-representation biases)	• No reactivity or self-presentation bias
• Data collected are prompted by pre-structured research questions	• Self-initiated views with no time, length, or behavioural constraints
• Data mediated/focused through research facilitator	• Data stem from participants' agenda
• Time consuming: small set of data	• Semi-automatic: large set of data

BIASES OF TRADITIONAL RESEARCH APPROACHES VERSUS INTERNET FORUM-BASED DISCOURSE ANALYSIS

The assessment of patients' behaviour for research purposes is fraught with difficulties (Box 1).[12] One of the problems of measuring behaviours is that the act of measurement can itself influence behaviour. The measurement of behaviour is vulnerable to reactivity and self-presentational bias on the part of the patient. Reactivity is the tendency of attention from others to influence behaviour. If patients are aware that their behaviour is being monitored, this might stimulate a specific behaviour simply by drawing attention to it. This is because of self-presentational bias. Patients may perceive that a certain behaviour, for example, adherence to treatment, is one of the duties expected of the 'good patient' and may be reluctant to admit different behaviour because they fear that this will offend or disappoint their doctor or risk their disapproval. For example, if the topic of interest is adherence to treatment, patients may create a falsely-elevated adherence score by taking more medication immediately prior to testing or by under-reporting non-adherence. In addition, patients who are low adherers might be less likely to sign up as participants in adherence studies.

Studies conducted across a range of chronic illnesses and involving patients from different countries and cultural groups and using qualitative and quantitative methods have consistently found that health behaviour is related to patients' perceptions of their illness, as well as to social representations of medicines in general and need for treatment. Research is needed to assess whether healthcare interventions can be modelled to help patients make decisions that are informed by realistic assessments of their healthcare needs and behaviours and are not based on mistaken premises or misplaced beliefs about illnesses and treatment.

We advocate the use of DA on internet forums as a means of assessing patients' behaviour in a way that bypasses reactivity and self-representation bias. Internet forum dialogues stem from participants' own agenda. These posts represent self-initiated views of people communicating with each other without time, length, or behavioural constraints, unlike in a traditional research study.

Limitations of this approach will emerge as more research is performed using DA methodology on internet forums. For example, some participants may use online forums to express frustration or anger against a perceived unfair situation and exaggerate particular circumstances.

THE TALKSTROKE INTERNET FORUM

TalkStroke is an online discussion forum from The Stroke Association website, which includes 22 173 individual entries from around 1000 registered participants. It is a collection of self-initiated views of a population that includes stroke survivors, their friends and/or family members and/or caregivers and sheds light on issues relevant to them.

Some participants consult the forum on behalf of stroke survivors unable to

"We propose discourse analysis of internet forums as a relatively quick and inexpensive means to better understand patients' issues and behaviours, with the view of better informing healthcare interventions and policies."

communicate their needs due to severe cognitive, communication, and physical impairments. Such needs might otherwise remain unknown to the public and to health services.

PRELIMINARY RESULTS

The TalkStroke forum entries have been collected into a body of text, a corpus, and readied for linguistic and textual analysis through the use of specialist software. Preliminary investigations of word frequency are reported here. The corpus is large: there are 42 337 different words with many of them repeated on multiple occasions. Frequency gives us markers of how important certain concepts are to people. If we prepare a list of the most frequent words in any corpus, we first come across 'grammar' words like determiners (such as 'the', 'a', and 'an') and pronouns (such as 'he', 'she', and 'it'). These are closely followed by very common verbs, (variants of 'be', 'do', 'go', 'think', and 'know'). If we remove these very common terms, key vocabulary items that are specific and important to this forum begin to emerge (Table 1).

Table 1. Most frequent items in TalkStroke Forum, ('grammar' words removed)

Number	Frequency	Item
1	15 884	stroke
2	6035	time
3	5541	help
4	4457	hospital
5	3511	home
6	3424	mum
7	3395	care
8	3246	hope
9	3016	need
10	2927	brain
11	2754	dad
12	2588	work

'Help' is one of the most frequently used words, followed by 'hospital' and 'home', possibly suggesting that unmet needs of participants and care settings are important argument of discussion. It is noticeable that the first two relationships in terms of frequency within the collection are 'mum' and 'dad', underscoring the importance of family relationships and family care-giving to participants. In addition, it is also notable how often names of individual doctors are expressed within the corpus itself. A preliminary glance suggests that forum participants often base the strength of knowledge claims on or around the names of doctors rather than on the content of the knowledge itself. This is further supported when we employ the software to examine reporting verbs. The word 'told' is frequently used to initiate reported speech, especially in narrative-rich text, such as this forum.

The word immediately preceding 'told' (for example, in a phrase such as 'The doctor told me ...') is very likely to tell us who is doing the telling. By far the most common agents are representatives of the health professions: doctor(s), consultant(s), physio(therapist)(s), hospital, GP, nurse, speech therapist(s), neurologist, in order of frequency; doctor(s) being the most nominated professional.

CONCLUSION

We propose DA of internet forums as a relatively quick and inexpensive means to better understand patients' issues and behaviours that might not be captured by traditional research studies, with the view of better informing healthcare interventions and policies.

This approach could lead to novel healthcare related studies, where discourse analysis is combined with more traditional qualitative approaches.

For example, findings reported here from the TalkStroke forum analysis could inform the design of topic guides for more qualitative work with stroke patients and their families, studying with an established methodology like content analysis arguments made around discussions on 'hospital' and 'home' and the type of 'help' participants are looking for within the forum.

Internet forums are available to other groups of patients suffering from chronic diseases, such as diabetes and heart disease. Such resources could allow comparison of relevant issues and behaviours from patients affected by different diseases and for which different healthcare interventions have been implemented.

The potential of internet forums for understanding what people say to each other in an unguarded context (for better or worse) is only just beginning to be understood. The software resources for an analysis of electronically stored language exist, and offer an opportunity to explore communication on needs, beliefs and treatments in a manner which holds substantial promise, and with a methodology which is at once well-validated and relatively novel.

ADDRESS FOR CORRESPONDENCE

Andrew Shanks
University of Birmingham, School of Health and Population Sciences, 90 Vincent Drive, Birmingham, B15 2SP.

E-mail: a.j.shanks@bham.ac.uk

Anna De Simoni,
Clinical Research Associate,The Primary Care Unit, University of Cambridge, Cambridge.

Andrew Shanks,
Lecturer in Behavioural Medicine, Primary Care Clinical Sciences, School of Health and Population Sciences, University of Birmingham, Birmingham.

Jonathan Mant,
Professor of Primary Care Research,The Primary Care Unit, University of Cambridge, Cambridge.

John R Skelton,
Professor of Clinical Communication, Primary Care Clinical Sciences, School of Health and Population Sciences, University of Birmingham, Birmingham.

Provenance
Freely submitted; externally peer reviewed.

REFERENCES

1. Goldfarb A, Prince J. Internet adoption and usage patterns are different: Implications for the digital divide. *Information Economics and Policy* 2008; **20(1):** 2–15.
2. Harvey KJ, Brown B, Crawford P, *et al.* 'Am I normal?' Teenagers, sexual health and the internet. *Soc Sci Med* 2007; **65(4):** 771–781.
3. Seale C, Charteris-Black J, MacFarlane A, McPherson A. Interviews and internet forums: a comparison of two sources of qualitative data. *Qual Health Res* 2010; **20(5):** 595–606.
4. Slembrouck S. Discourse and discourse analysis. In: Cummings L, ed. *The pragmatics encyclopedia*. London: Routledge, 2010: pp. 116–120 and 120–122.
5. Stubbs M. *Discourse analysis. The sociolinguistic analysis of natural language.* Oxford, Blackwell, 1983.
6. Hymes D. *Foundations in sociolinguistics: an ethnographic approach*. Philadelphia, PA: UPenn Press, 1975.
7. Skelton J, Hobbs FDR. Concordancing: the use of language-based research in medical communication. *Lancet* 1999; **353:** 108–111.
8. Skelton JR, Wearn AM, Hobbs FDR. 'I' and 'we': a concordancing analysis of doctors and patients use first person pronouns in primary care consultations. *Fam Pract* 2002; **19(5):** 484–448.
9. Adolphs S, Brown B, Carter R, *et al.* Applying corpus linguistics in a health care context. *J Appl Linguistics* 2004; **1(1):** 9–28.
10. Metcalfe A, Plumridge G, Coad J, *et al.* Parents and children's communication about genetic risk: qualitative study learning from families' experiences. *Eur J Hum Genet* 2011: **19(6)**: 640–646.
11. Greenhalgh T, Procter R, Wherton J, *et al.* The organising vision for telehealth and telecare: discourse analysis. *BMJ Open* 2012; **2**: e001574. DOI:10.1136/bmjopen-2012-001574.
12. Gordis L. Conceptual and methodological problems in measuring patient compliance. In: Sackett DL, Taylor DW and Haynes RB (eds). *Compliance in health care*. London: John Hopkins University Press, 1979: 23–45.