Quality of Life Research (2019) 28:1441–1455
https://doi.org/10.1007/s11136-019-02132-w

Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery

Yuelin Li[1,2,5] · Bruce Rapkin[3] · Thomas M. Atkinson[1] · Elizabeth Schofield[1] · Bernard H. Bochner[4]

Accepted: 6 February 2019 / Published online: 23 February 2019

Abstract

Purpose As we begin to leverage Big Data in health care settings and particularly in assessing patient-reported outcomes, there is a need for novel analytics to address unique challenges. One such challenge is in coding transcribed interview data, typically free-text entries of statements made during a face-to-face interview. Latent Dirichlet Allocation (LDA) offers statistical rigor and consistency in automating the interpretation of patients' expressed concerns and coping strategies.

Methods LDA was applied to interview data collected as part of a prospective, longitudinal study of QOL in $N=211$ patients undergoing radical cystectomy and urinary diversion for bladder cancer. LDA analyzed personal goal statements to extract the latent topics and themes, stratified by time, and on things patients wanted to accomplish and prevent. Model comparison metrics determined the number of topics to extract.

Results LDA extracted seven latent topics. Prior to surgery, patients' priorities were primarily in cancer surgery and recovery. Six months after the surgery, they were replaced by goals on regaining a sense of normalcy, to resume work, to enjoy life more fully, and to appreciate friends and family more. LDA model parameters showed changing priorities, e.g., immediate concerns on surgery and resuming employment decreased post-surgery and were replaced by concerns over cancer recurrence and a desire to remain healthy and strong.

Conclusions Novel Big Data analytics such as LDA offer the possibility of summarizing personal goals without the need for conventional fixed-length measures and resource-intensive qualitative data coding.

Keywords Latent Dirichlet Allocation · Big Data analysis · Text analysis · Qualitative data · Bladder cancer

Introduction

Bladder cancer is the sixth most commonly diagnosed cancer with an estimated 81,190 cases in 2018 in the USA alone [1]. It is the second most common malignancy of the genitourinary tract. Approximately 2.3 percent of men and women will be diagnosed with bladder cancer at some point during their lifetime. The majority of patients with high-grade aggressive non-muscle invasive bladder cancer or with muscle invasive tumors are treated with radical surgery and urinary diversion, the most common of which include two options: (1) the cutaneous ileal conduit urinary diversion and (2) the orthotopic neobladder. A patient who has the ileal conduit diversion will need to use an appliance (external stoma bag) to collect the urine at the level of the abdominal wall, which may adversely affect body image. With the orthotopic neobladder, a continent urinary reservoir is reconstructed from a section of small intestine or colon, which is attached to the native urethra, allowing for volitional voiding via the normal anatomic pathway. Neobladder patients must learn to void via a strain voiding technique. Either of these two reconstructive options potentially may affect the patient's daily activities and quality of life.

✉ Yuelin Li
liy12@mskcc.org

1 Department of Psychiatry & Behavioral Sciences, Memorial Sloan Kettering Cancer Center, New York, NY, USA

2 Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

3 Albert Einstein College of Medicine, Montefiore Health System, Bronx, NY, USA

4 Urology Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA

5 641 Lexington Avenue, 7th Floor, New York 10022, NY, USA

QOL appraisal in illness and recovery

The prospect of a urostomy can induce a great deal of stress in patients who earnestly want a near-normal lifestyle after surgery. Personal goals, priorities, and concerns that comprise individuals' "frame of reference," are an essential component of the theoretical model of quality-of-life appraisal [2–5], operationalized in the Quality-of-Life Appraisal Profile (QOLAP) assessment tool [3]. This theoretical framework posits that QOL ratings arise from a personal *appraisal* process engendering an idiosyncratic meaning of quality of life from four elements. In addition to frame of reference (7 items), appraisals are also affected by the way individuals "sample experiences" (14 items) within their frame of reference, the "standard of comparison" (8 items) by which individuals evaluate their experiences, and a "combinatorial algorithm" (16 items) to weigh up all elements. The frame of reference items are the only open-ended questions, which involve coding and scoring by trained qualitative data analysts to yield a scale, a time intensive and difficult coding exercise despite an effort to streamline it [6]. Its psychometric properties have been evaluated, including its ability to complement conventional QOL measures [7] and more recently on its discriminant validity [8] and construct validity [9].

QOL in bladder cancer

Evidence to date shows no consistent effect on QOL between the diversions. In one of the largest studies in bladder cancer, Hart et al. [10] reported no statistically reliable difference in QOL between the diversions after controlling for age and gender differences. Other studies have reported a significant advantage in ileal conduit diversion [11–13]. Several recent meta analyses found divergent results. Yang et al. [14] and Ali, Hayes, Birch, Dudderidge, and Somani [15] found no differential QOL effects, but Cerruto et al. [16] found slightly better QOL in orthotopic neobladder.

Why not just ask the patients?

Broad and open-ended questions such as "How well are you taking care of yourself?" may yield useful information on aspects of QOL most important to each individual patient. Such questions may complement standard QOL assessments by giving patients the opportunity to express what matters to them the most without being asked about specific health problems. However, coding of interview data is time consuming and not easily scaled up as free-text data accrue.

Need for analytic approaches to quantify goals in free-text data

Latent Dirichlet Analysis (LDA), a quantitative text analysis technique, distils vast amounts of free-text data into underlying topics [17–19]. This article aims to assess the use of LDA in coding and scoring free-text data to complement conventional QOL assessments:

Aim 1 Apply LDA to extract primary topics in goals, priorities, and concerns ("frame of references") expressed by bladder cancer patients undergoing surgery.

Aim 2 Characterize the changes in the primary topics in patients' goals and priorities before and after cancer surgery.

Aim 3 Investigate the strengths and limitations in LDA in coding free-text data to complement conventional QOL measures.

With respect to Aim 3, we share our experience using LDA in QOL research, rather than a fuller, more comprehensive investigation such as contrasting LDA against conventional qualitative data analysis [20]. We first give a brief overview on a practical level of understanding so that a reader may gain enough working knowledge of LDA. Then we demonstrate LDA by applying it to free-text responses from bladder cancer patients before and after bladder radical cystectomy and urinary diversion.

Methods

All procedures and assessments were approved by the MSKCC Institutional Review Board. Below we describe how the free-text data were collected as part of the QOLAP measure. Below we briefly summarize the procedure pertaining to only the open-ended portion of the QOLAP assessment.

Procedure of the QOLAP open-ended questions

The interviewer first prompted the respondent to think about the meaning of quality of life, that it could change over time, and that we were interested in knowing what was most important to him/her. The following open-ended questions, printed on a paper questionnaire, were presented:

- In a sentence, what does the phrase "quality of life" mean to you at this time?

- In order to have the most satisfying life possible….

1. What are the main things that you want to accomplish?
   _____
   _____
   _____
   _____
2. What are the main problems that you want to solve?
3. What situations do you want to prevent or avoid?
4. What things do you want to keep the same as they are now?
5. What things do you want to accept as they are?
6. What demands and responsibilities do you want to let go of or reduce?

The six questions were presented on paper in the order shown, each followed by four blank lines for written responses. The basic structure of the research interaction asks patients to name the primary goals they would like to accomplish, what problems they would like to solve, what they would want to prevent or avoid, what they would want to keep the same as they are now, and what commitments they would want to let go, what things that they want to be able to accept as they currently are, and what special events or milestones they are looking forward to reach in order to have the most satisfying life possible. For example, one patient wrote three sentences in response to things he/she wants to accomplish before surgery: (1) "survive the surgery"; (2) "have a manageable recovery process"; and (3) "return to a normal life after recovery."

Analytic plan

The parent study of the current LDA application was originally planned to recruit a sample of $N = 550$ bladder cancer patients (allowing for attrition, a complete sample of $N = 500$ would detect a Cohen's $d = 0.25$ with an 80% power in the QOL outcomes between the ileal conduit and neobladder diversions, at a two-sided type-I error rate of 0.05). Data from $N = 537$ were collected. However, the QOLAP assessment was introduced in an amendment to the original study, approximately half-way after recruitment had begun ($N = 211$ reported below for LDA).

Data analysis began by assembling free-text entries of patients' goal statements. For instance, the three separate sentences in the example above were combined into one single *document* for this patient, with punctuations removed and words converted into lower-case letters as per standard text analysis. This was repeated so that each patient had one document representing his or her goals at baseline and another document for goals at 6 months post-surgery. This step yielded 104 unique documents at baseline and 211 documents at 6 months post-surgery on things patients wanted to *accomplish*. There were fewer baseline entries because the ideographic assessments were introduced after study recruitment had begun. The same steps were applied to what patients wanted to *prevent*. Henceforth, they are called the *accomplish* and *prevent* documents.

LDA overview

LDA was developed by Blei et al. [17] to process large quantities of unlabeled data, and used primarily in extracting latent topics from enormous amounts of digital data such as online postings [17], scientific articles [18], educational materials [19], photographs on social media [21], and music notes and cords [22, 23]. LDA can quickly distil information from any discrete data, from sources that are otherwise too expensive to comprehend by conventional methods.

Figure 1 provides a visual explanation of LDA with 3 hypothetical *documents*, verbatim statements from patients on things they would like to accomplish. Document 1 draws words from the topic of 'cancer treatment' with probability 1.0. Document 2 draws a mixture of words from 'cancer treatment' with probability 0.5 and from 'family and life' with probability 0.5, and Document 3 draws words entirely from 'family and life' with probability 1.0.

If the number of topics is fixed at T topics, then we can write the probability of the ith word in a given document as generated from the following mixture process:

$$p(w_i) = \sum_{j=1}^{T} p(w_i | z_i = j) p(z_i = j),$$

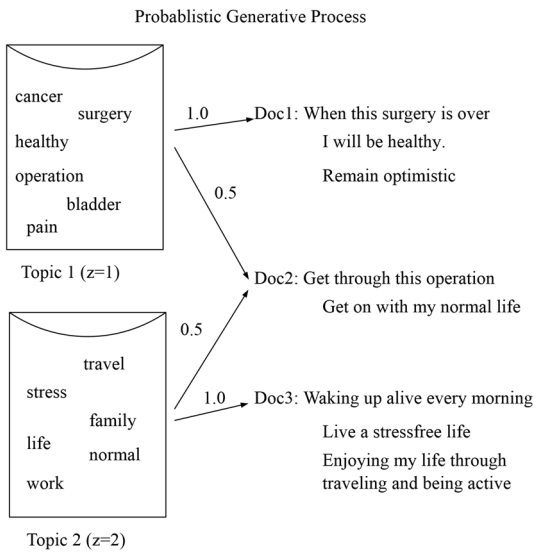

Fig. 1 Visual representation of the LDA model. Each document includes all statements made at a given time point for a single patient

where $p(w_i|z_i = j)$ is the probability of the ith word under the jth topic and $p(z_i = j)$ is the probability of choosing a word from topic j in the current document, which will vary across different documents. Words in the ith document emerge from the jth topic with probabilities $p(w_i|z_i = j)$, which are distributed according to a set of discrete multinomial distributions with parameters ϕ^{z_i}. The T by W *topic-vocabulary* matrix, $p(w_i|z_i = j)$, shows the association between the words and the jth underlying topic, analogous to the factor loading matrix in a conventional exploratory factor analysis. The second set of parameters $p(z_i)$, the *document-topic* matrix, allows documents to be sampled by mixing words from different topics. It is analogous to the factor score matrix in a conventional factor analysis (used to assign each person a score on a factor). It represents the probability that a patient's expressed goals arise from the underlying topics.

Blei et al. [17] designed LDA to distil topics from text documents so enormous that comprehension cannot be feasibly attempted by reading them. They applied LDA to 2500 news articles with a vocabulary size of 38,000 unique words and 1400 technical abstracts in a statistics literature with 8000 unique words. Griffiths and Steyvers [18] applied LDA to scientific abstracts published in 1991–2000 in the Proceedings of the National Academy of Sciences. LDA extracted topics that correspond well with keywords in this large corpus of documents. Similar findings showed that language used on social media is associated with personality characteristics [24], and Big Five personality traits [25].

Overview of LDA analytic plan

Preprocessing

The LDA computation was primarily done by a publicly accessible toolkit called scikit-learn, computation tools written in the Python computer programming language [26]. We followed the typical steps in text analysis: (1) preprocessing; (2) determining the number of topics; and (3) setting the control parameters. Text preprocessing was done by first correcting misspellings and typos in the raw data. Additional processing was done by the CountVectorizer() procedure, which encoded the corpus of documents into numeric vectors by frequency count to extract feature words. Additionally, it also removed stop words and added bigrams. Stop words (e.g., 'the', 'of', 'from', words that carry limited information) were based on the built-in stop words dictionary. Bigrams were added (e.g., two-word phrases that appeared consecutively in a sentence, such as 'continue traveling', 'cancer-free', and 'make money'). Bigrams were included to provide contextual information in the sequences of words. For example, "survive the surgery" would yield two distinct words 'survive', 'surgery' and one bigram of 'survive surgery.' The sentence "have a manageable recovery process" would yield bigrams 'manageable recovery' and 'recovery process' in addition to the unique words. Bigrams were analyzed as distinct text patterns. Namely, bigrams would be ranked high in the LDA results if many patients used the same bigrams.

Number of topics

Four different model comparison metrics in the R package ldatuning [27] were used to establish the desired number of topics to extract in LDA [28–30, 18], using all available data pooled over time. All subsequent analyses were fixed at this number to make a consistent and streamlined presentation. We opted for this pragmatic approach to use all available data, given the relatively small number of short documents.

LDA parameter settings

LDA was done using the LatentDirichletAllocation() procedure in scikit-learn to yield the *document-topic* and *topic-vocabulary* matrices. An essential input parameter is the number of topics T, which tunes other default parameters. The default priors for both matrices are set specifically to $1/T$. For example, the default prior on the document-topic matrix is set to a symmetric Dirichlet(1/7, 1/7, …, 1/7) for a Dirichlet prior with 7 categories. For values of parameters < 1, the distribution concentrates in the corners and along the boundaries of the simplex, thus expressing the preference for few topics in a document (an intuitive, animated explanation can be found in Ipeirotis [31]). The variational Bayes algorithm offers a fast LDA solution. Additionally, following Griffiths and Steyvers [18], we used the document-topic matrix to quantify patients' priorities over time. The document-topic matrix contains the probability that the topics are expressed in the 315 goal statements (104 at baseline and 211 at 6 months post-surgery). For example, we can calculate the average probability of topic 1 over all 104 baseline goals to represent the patients' priority on topic 1 at baseline. Similarly, the average probability of topic 1 over all 211 post-surgery goal statements represents the patients' priority on topic 1 at post-surgery. Changes in the priority scores offer a descriptive statistic on how patients' priorities changed over time.

Exploratory and descriptive analyses were done with the tm package in R [32, 33] for visualization. The computer code is available from the authors upon request.

Results

Participant characteristics at baseline

Table 1 summarizes the baseline characteristics of N = 211 patients who gave at least one interview. Most patients

Table 1 Baseline characteristics ($N=211$) of patients who gave us at least one free-text data entry

Characteristics[a]	Neobladder	Ileal conduit	Continent cutaneous
	$N=95$ (45%)	$N=107$ (51%)	$N=8$ (4%)
Sex			
Male	88 (93%)	76 (71%)	3 (38%)
Female	7 (7%)	31 (29%)	5 (62%)
Age			
Mean (range)	64.06 (33–82)	72.00 (43–91)	56.28 (36–71)
Race/ethnicity			
White	92 (97%)	99 (93%)	7 (88%)
Non-White	3 (3%)	8 (7%)	1 (12%)
Marital status			
Married/partnered	71 (75%)	82 (77%)	7 (88%)
Not-married	24 (25%)	25 (23%)	1 (12%)
Employment			
Employed	49 (52%)	41 (38%)	3 (38%)
Not-employed	46 (48%)	66 (62%)	5 (62%)
Stage			
Ta, Tis, T1	44 (46%)	52 (49%)	3 (38%)
T2–T4	51 (54%)	54 (50%)	5 (62%)
TX	0 (0%)	1 (1%)	0 (0%)

[a]Diversion type data not available in 1 patient

underwent a neobladder (45%) or an ileal conduit (51%) urinary diversion. Few patients underwent a continent cutaneous diversion (4%). Patients were mostly White, male, and married or partnered. The average age was 65.8 years. There were approximately equal representations of patients with lower-grade disease and muscle invasive disease ("T2–T4").

Examples of things patients wanted to *accomplish* at baseline

Table 2 provides illustrations of free-text entries on what patients wanted to *accomplish* prior to surgery. These are illustrative examples from the combined total of 315 unique statements. The first patient expressed goals on a successful cancer surgery, thereafter a sense of normalcy and the enjoyment of maximum time spent with family. The second patient's goals were similar, with an additional consideration on traveling. The third person expressed a concern over money and paying bills. There appears to be shared commonality in patients' goals, a successful surgery, uneventful recovery, and continuation with normal and independent life, work, with a renewed prioritization on living life to its fullest.

Frequency distributions of words at baseline and at post-surgery

Figure 2 shows the distributions of word counts on what patients wanted to accomplish at baseline and at post-surgery. Words used scarcely were excluded. Changes were visible in the frequency distribution of words from baseline to post-surgery. For example, before surgery, the most frequently used words included 'surgery', 'life', 'healthy', 'continue', and 'cancer.' After surgery, the most frequently used words were 'back', 'family', 'life', and 'work.'

How many topics?

Next, a series of LDA models were fitted to examine the desired number of topics in the LDA model using the combined documents on things patients want to accomplish. Figure 3 plots the changes in model comparison metrics when the number of extracted topics increases. The Griffiths and Steyvers [18] metric indicated a model with 7 topics as shown in the curve plateau. The Cao et al. [29] metric indicated approximately 6 to 7 topics. The other two metrics provided limited guidance because of the monotonic patterns. By considering all four metrics, on balance, a 7-topic LDA model was used in subsequent analyses.

Table 2 Examples of things patients would like to *accomplish* at baseline

Statements		
First	Second	Third
Get through the operation successfully	Be with my family as much as possible	Do what I normally do
Get this bladder thing behind me	Do some more traveling	Spend some more time with grandkids
Take time off work	Not worrying about paying bills	–[a]
Beat the cancer	Go back to work	Live like I used to
To continue to be fully able to take care of myself	Good emotional and spiritual quality of life	Continue to work

[a]Most patients gave 3 statements, some gave only 2

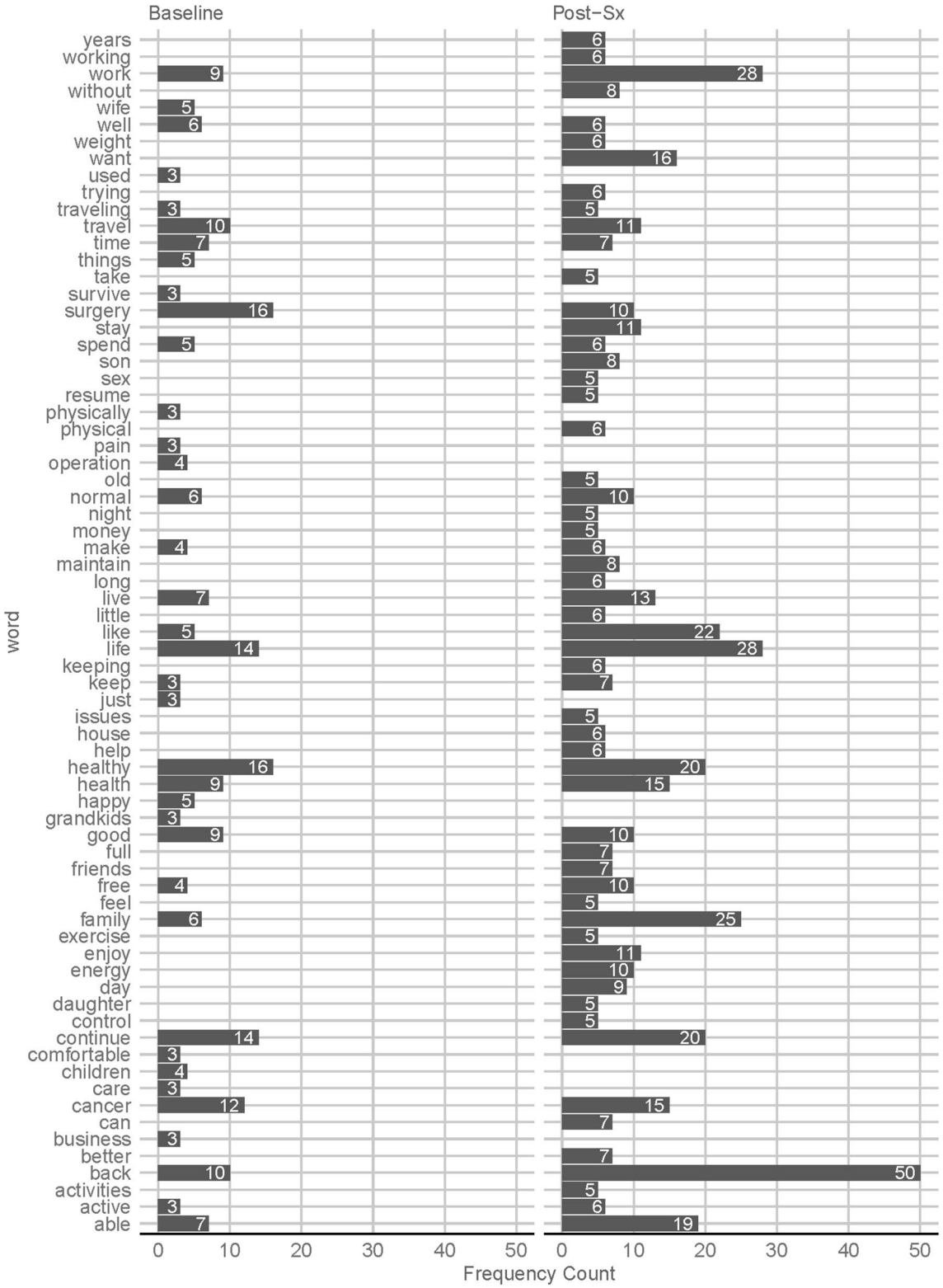

Fig. 2 Frequency distributions of words used at baseline and at post-surgery. Words used infrequently were not tallied to reduce clutter (fewer than 3 at baseline and fewer than 5 at post-surgery)

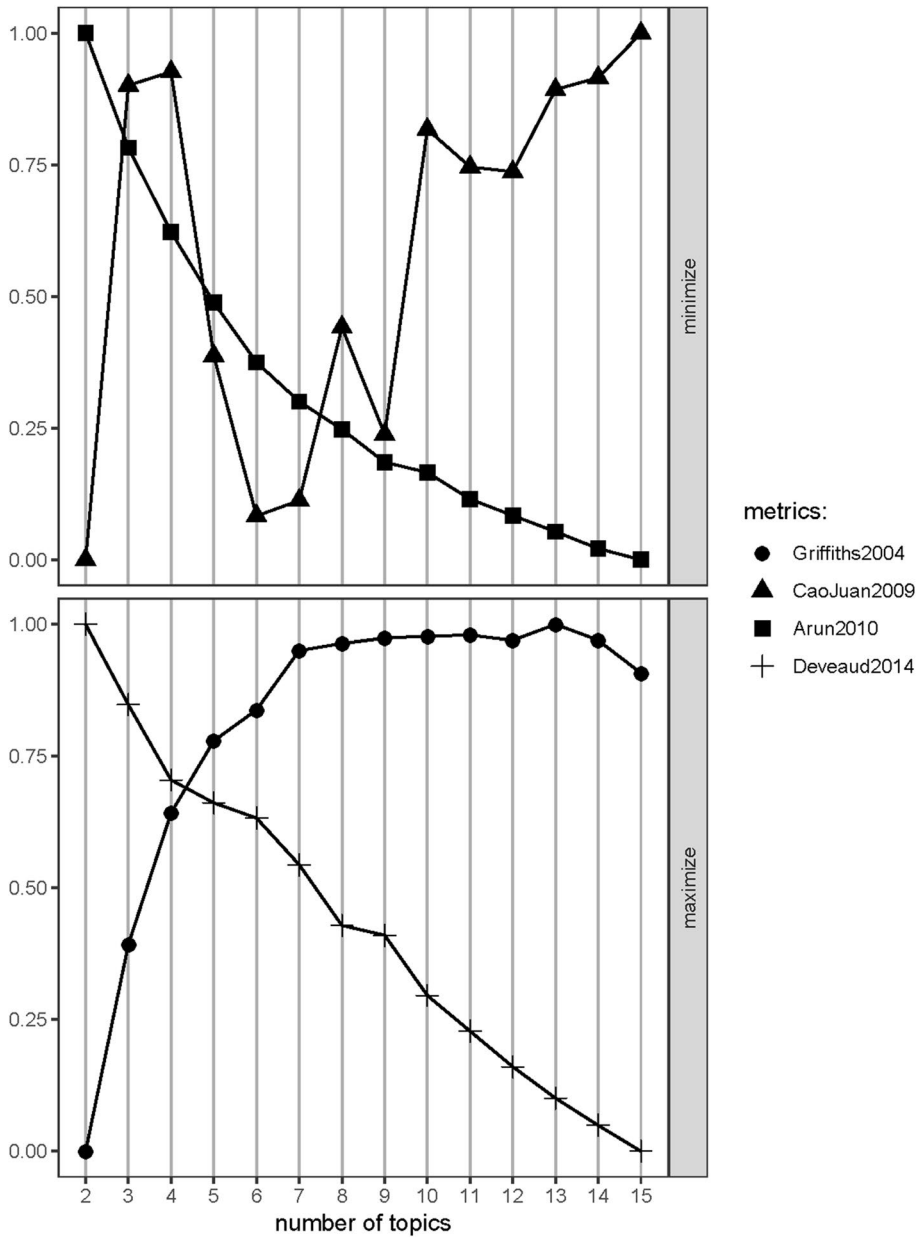

Fig. 3 Model comparison metrics to estimate the desired number of topics, using the things patients would like to accomplish at baseline and post-surgery combined. The Griffiths and Steyvers [18] criterion suggested 8 to 10 topics are suitable. Considerable variabilities indicate that not all criteria agree on the number of topics

Personal goals before and after bladder cancer surgery

Table 3 summarizes the topics in what patients would like to accomplish, using the goal statements separately collected at baseline and 6 months post-surgery, respectively. The featured words were the top 10 words that had the highest probability of belonging to a topic. At baseline, topic 1 appeared to be related to optimism in a successful surgery that restores health. Topic 2 was related to the hope that surgery will bring about a healthy life so that one can travel and be productive ('doing things'). Topic 3 was related to a desire to travel, recover, and restore physical health to normal. Topic 4 appeared to be related to living a full life after surgery, to travel, be active, and to have a normal health. Topic 5 mapped onto a desire to be cancer-free and to maintain a healthy life. Topic 6 mapped onto the wish to have a successful surgical operation and to resume the role as a father to the family. Finally, topic 7 expressed a wish to remain healthy, to be able to work and remain happy. These topics appeared to encompass several broad areas, such as the goal to have a successful surgery and be cancer-free (topics 5 and 6), to resume roles as a 'father' to the family (topic 6), to be able to 'work' (topic 7), 'travel' (topics 2 and 3), and remain 'optimistic' (topic 1). The desire to travel and to have a sense of normalcy and emotional wellbeing appeared in several topics. Overall, patients' expressed concerns over cancer and surgery were in a context of family, marriage,

Table 3 Topics extracted by LDA on things patients want to *accomplish* prior to surgery and 6 months after surgery

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6		Topic 7	
Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.
Featured words: Prior to surgery ($n=104$ documents, vocabulary size = 632)													
Health	0.0193	Surgery	0.0343	Travel	0.0194	Life	0.0217	Healthy	0.0169	Good	0.0146	Continue	0.0284
Surgery	0.0103	Make	0.0118	Time	0.0132	Live	0.0132	Cancer	0.0169	Operation	0.0146	Healthy	0.0228
Getting	0.0103	Life	0.0118	Spend	0.0132	Surgery	0.009	Free	0.0128	Father	0.0146	Good	0.0172
Healthy	0.0103	Health	0.0118	Spend time	0.0132	Travel	0.009	Things	0.0087	Work	0.0078	Work	0.0172
With	0.0103	Continue	0.0118	Recovery	0.0132	Normal	0.009	Cancer-free	0.0087	Family	0.0078	Getting	0.0172
Happiness	0.0103	Travel	0.008	Physical	0.0132	Used	0.009	Life	0.0087	Open	0.0078	Happy	0.0116
Optimistic	0.0103	Time	0.008	Surgery	0.007	Wife	0.009	Going	0.0047	Possible	0.0078	Cancer	0.0116
Surgery healthy	0.0103	Things	0.008	Traveling	0.007	Able	0.009	Stay	0.0047	Live	0.0078	Able	0.0088
Remain-optimistic	0.0103	Doing	0.008	Return normal	0.007	Active	0.009	stay-healthy	0.0047	Successfully	0.0078	Just	0.0088
Remain	0.0103	Design	0.008	Return	0.007	cancer	0.009	Business	0.0047	Lot	0.0078	Doing	0.0088
Featured words: after surgery ($n=211$ documents, vocabulary size = 1000)													
Cancer	0.025	Family	0.032	Continue	0.015	Life	0.026	Life	0.037	Travel	0.023	Like	0.043
Work	0.024	Health	0.028	Issues	0.015	Work	0.018	Live	0.017	Work	0.021	Able	0.017
Cancer-free	0.019	Maintain	0.018	Clean	0.010	Weight	0.010	Long	0.015	Walk	0.010	Getting	0.013
Free	0.019	Time	0.017	Active	0.008	Live	0.010	Healthy	0.013	Continue	0.008	Life	0.012
Energy	0.018	Spend	0.013	Having	0.008	Continue	0.009	Disease	0.011	Healthy	0.008	Normal	0.011
Want	0.018	Want	0.012	Improve	0.008	Help	0.008	Long healthy	0.010	Recovery	0.008	Day	0.010
Health	0.011	Able	0.012	Feeling	0.008	Lose	0.008	Healthy life	0.010	Family	0.009	Friends	0.008
Able	0.010	Spend time	0.011	Functioning	0.008	Lose weight	0.008	Continue	0.009	Lives	0.008	Continence	0.008
Good	0.010	Time-family	0.010	Sexual	0.008	Continue-work	0.008	Working	0.009	Health	0.007	Erection	0.007
Trying	0.009	Stay	0.010	Sexual functioning	0.008	Close	0.007	Live long	0.009	Finish	0.005	Family	0.007

and an optimism that health will be restored so that work and travel would be possible.

At 6 months post-surgery, topic 1 appeared to express a desire to be 'cancer-free' and to be able to work and have energy. Topic 2 mapped onto family and spend time with family. Topic 3 appeared to involve in addressing post-surgery issues such as 'clean', 'active', and 'sexual functioning.' Topic 4 appeared to convey a goal to be able to work, to live, and to lose weight. Topic 5 expressed a desire to live a long healthy life and to continue working. Topic 6 conveyed the wish to travel, work, to be healthy and be with family. And topic 7 appeared to communicate a desire to resume a normal life without problems related to urinary incontinence and erection. There appeared to be a general theme on recovery, a long and healthy life, family, a sense of normalcy, enjoyment of life, friends and family, resuming employment, and free of concerns over incontinence and erectile dysfunction.

Table 4 summarizes the topics on things patients would like to *prevent*. Important at baseline were concerns over the loss of bladder and cancer metastasis (topic 1), medical decisions in cancer operation, possibly in the choice of urinary diversion (topic 2). Generally, the remaining topics included concerns about surgical complications, general health issues, falling, operation, depression, immobility, becoming dependent on a caretaker, and death. At post-surgery, concerns over cancer recurrence, relapse, and illness remained (topics 2, 5, 6, and 7). Concerns over leakage, accidents, problems and incontinence at night emerged (topics 1, 3 and 4), and a desire to prevent death so that the patient may enjoy survivorship and reach an old age.

Changes in priorities after bladder cancer surgery

Figure 4 displays the changes in patients' priorities on things they wanted to accomplish. The desire for a long healthy life (topic 3) had an average priority score of 0.32 at baseline. An estimated 32% of the goal statements at baseline expressed a desire for a long and healthy life. At post-surgery, it reduced to 0.26, a slight reduction in the expression of topic 3 at post-surgery. Concerns over surgery and recovery (topic 5) reduced from an average priority of 0.18 at baseline to 0.15 at post-surgery. Increased in valence were topics 2 (a desire to work and be more physically active) and 7 (to have good health, feeling strong, and erection). The plot on the right shows the relatively unchanging goals on spending time with family, maintaining health and happiness, and being cancer-free. Note, however, that among all seven topics, topics 3 and 5 remained relatively high in importance at both time points.

Figure 5 shows the changes in priorities on things patients wanted to prevent. Goals that decreased in prominence at post-surgery were on surgery, death, and complications and avoiding future illness and additional surgery. Goals that became more prominent were cancer recurrence and becoming ill again. Relative stable were priorities on concerns over post-surgery leakage/accidents, having additional health problems including cancer recurrence, and minimization on future issues associated with illness in general and specific problems with the stoma bag.

Changes in priorities between urinary diversions

To further examine the extent to which patients' priority change depended on their urinary diversions, we plotted Fig. 6 to stratify the priority by time and urinary diversion type. Confidence intervals (95%) were added to represent the variabilities. Overall, the overlapping confidence intervals indicate no statistically reliable differences between the estimated probabilities. Although subtle differences were visible, e.g., in subplot (A), that patients with the neobladder diversion showed a slightly greater reduction after surgery in topic 3 ('life,' 'healthy,' 'continue,' etc.) than ileal conduit patients. In subplot (B), increase in concerns over topic 1 ('cancer,' 'recurrence,' 'cancer recurrence') appeared to be greater in ileal conduit patients than neobladder patients.

Scoring each person's goals

Figure 7 offers a visual explanation on how patients' statements are scored against the topics. Such a plot is often used in LDA to highlight the roles words play in documents [17, 18]. Words are tagged with topic labels as superscripts, representing the topic most strongly associated with each word. Words without superscripts are excluded from the vocabulary (e.g., stop words). Moreover, we can calculate the probability that a specific word belongs to the most prevalent topic in a document, which can be used to identify the latent topic and its constituent words most important to a person at a given time point. This probability is a graded measure, here used to set the contrast of the words so that words with a high contrast (darker font) are important to the underlying topic. It is important to note that, under a perfect LDA, all words would have the same tag and in dark font if a patient has one and only one goal.

Patient 526 scores the highest on topic 5 (successful surgery/recovery and then normalcy) at baseline and then topic 7 (cancer-free and good health) at post-surgery. Patient 418 expresses goals in topic 5 in both assessment time points. Patient 504 expresses goals in both topics 4 and 6 at baseline, then changes to topic 3 at post-surgery. The low contrast in many words indicate that LDA mapping is not perfect, due in part to patients having multiple goals. However, the crude information offers a proof of principle. LDA offers a pragmatic measure to quantify individual patient's

Table 4 Topics extracted by LDA on things patients want to *prevent* prior to surgery and 6 months after surgery

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6		Topic 7	
Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.
Featured words: prior to surgery ($n=104$ documents, vocabulary size = 434)													
Loss	0.0203	Want	0.0276	Surgery	0.0596	Death	0.0525	Effects	0.0171	Surgery	0.0229	Cancer	0.0806
Bladder	0.0203	Going	0.0188	Complications	0.0328	Falling	0.0231	Operation	0.0171	Dying	0.0229	Cancer recurrence	0.0366
Bag	0.0203	Operation	0.0100	Complications-surgery	0.0222	Life	0.0158	Effects-operation	0.0171	Want	0.0156	Recurrence	0.0366
Having	0.0138	Medical	0.0100	Want	0.0168	Want	0.0158	Depression	0.0171	Sure	0.0156	Coming	0.0166
Loss bladder	0.0074	People	0.0100	Long	0.0115	Prevent	0.0084	Immobility	0.0171	Make	0.0156	Cancer coming	0.0166
Cancer spreading	0.0074	Life	0.0100	Knee	0.0115	Issues	0.0084	Sickness	0.0021	Dependent	0.0156	Cancer spreading	0.0126
Spreading	0.0074	Operation-going	0.0100	Function	0.0115	Health	0.0084	Illness	0.0021	Make-sure	0.0156	Spreading	0.0126
Long	0.0074	Compared	0.0100	Infection	0.0115	Health issues	0.0084	Death	0.0021	Cancer	0.0156	Prevent	0.0086
Pain	0.0074	Change-choice	0.0100	Recurrence	0.0115	Stress	0.0084	Surgery	0.0021	Continue	0.0083	Getting	0.0086
Surgery loss	0.0074	Choice-compared	0.0100	Illness	0.0115	Surgery	0.0084	Cancer	0.0021	Bladder cancer	0.0083	Avoid	0.0086
Featured words: After surgery (n = 211 documents, vocabulary size = 825)													
Avoid	0.019	Getting	0.044	Want	0.016	Avoid	0.028	Cancer	0.114	Illness	0.019	Wan	0.026
Getting	0.019	Cancer	0.033	Death	0.012	Accidents	0.013	Recurrence	0.075	Having	0.016	Avoid	0.017
Sick	0.017	Prevent	0.022	Problem	0.009	Going	0.013	Cancer recurrence	0.055	Negative	0.015	Dying	0.014
Leakage	0.017	Cancer coming	0.017	Night	0.009	Like	0.010	Avoid	0.031	Prevent	0.014	Old	0.010
Going	0.014	Coming	0.017	Problems complaint	0.009	Prevent	0.010	Prevent	0.020	Prevent having	0.011	Relapse	0.010
Getting sick	0.014	Sick	0.015	Concerned-problems	0.009	Avoid going	0.010	Avoid cancer	0.015	Illness-negative	0.011	Years	0.010
Health	0.014	Getting sick	0.015	Concerned	0.009	People	0.007	Weight	0.012	Family	0.008	Years old	0.010
Things	0.014	Future	0.013	Complaint	0.009	Ill	0.007	Having cancer	0.011	Happening	0.008	Getting	0.010
Prevent	0.012	Sickness	0.010	Incontinence	0.009	Leakage	0.007	Recurrence cancer	0.010	Exhaustion	0.008	Dying-years	0.007
Prevent getting	0.010	Getting cancer	0.010	Incontinence-night	0.009	Know	0.007	Having	0.009	Ridden exhaustion	0.008	Avoid relapse	0.007

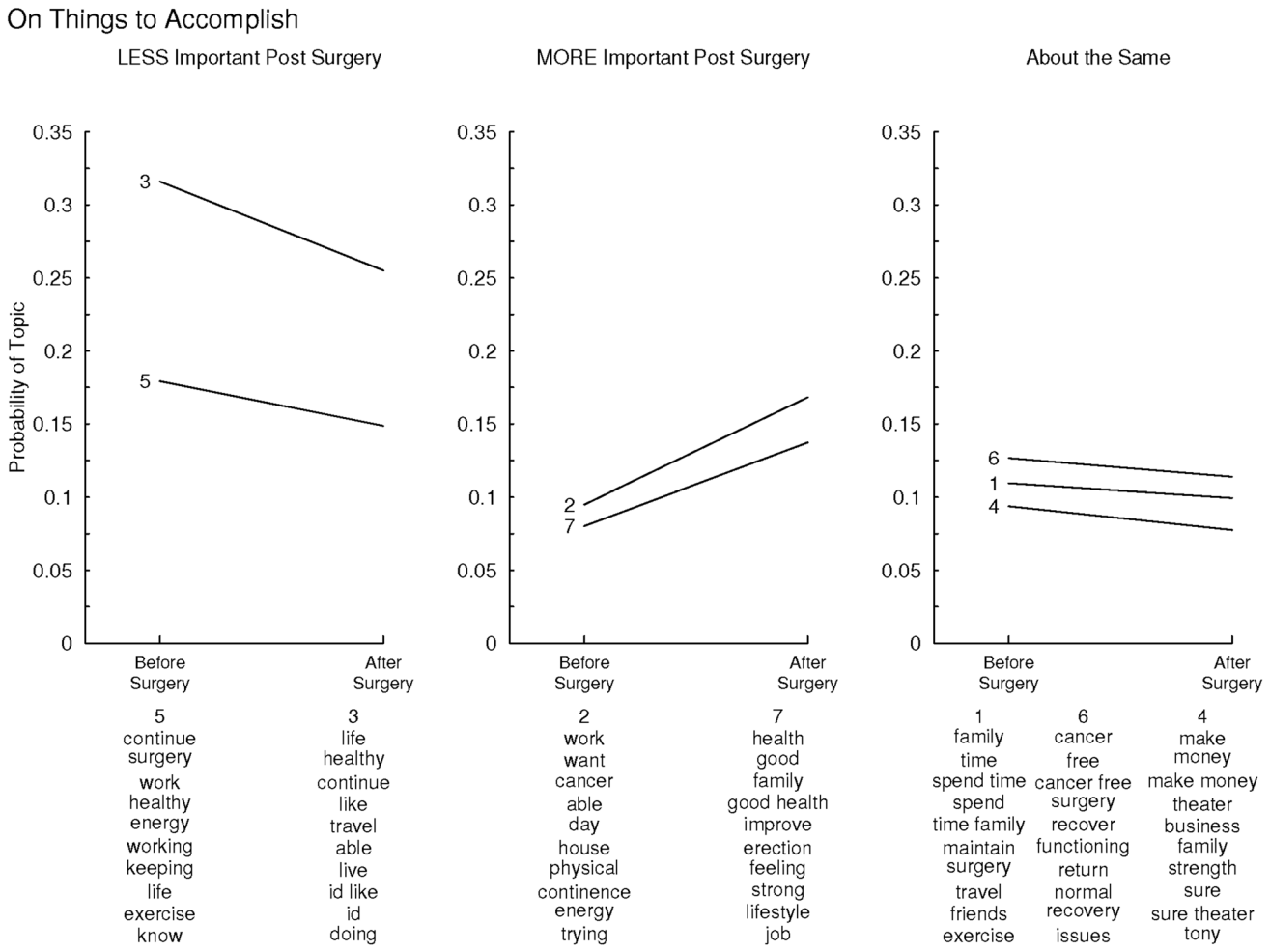

Fig. 4 Changes in patients' priorities on things they wanted to accomplish before and after bladder cancer surgery. The priority scores were calculated using the combined 315 documents as described in the "Methods" section to derive the probability of each person's goals mapping onto the 7 latent topics, and averaged across persons and time

underlying concerns over time purely from their verbatim goal statements.

Discussion

LDA makes qualitative data more accessible to QOL researchers. The automated LDA process reduces the need for a researcher to read and immerse in the qualitative data. Moreover, patients are given the opportunity to express what matters the most to them personally, in their own words (and not the words chosen for them). Researchers can capitalize on LDA to go beyond the typical QOL research paradigm, with its fixed-length QOL measurement tools and conventional psychometric scoring methods. LDA emulates quantitative information important in scale scoring and in examining changes in priorities. LDA is but one method among numerous novel methodological advances that offer QOL researchers additional tools, tools that cater to the need to comprehend vast amount of electronic data on social media and other platforms of technology advances.

The probabilistic priority scores offer a way to score people's changing priorities over time. Immediate concerns over surgery and recovery at baseline are replaced by goals on regaining a sense of normalcy, to resume work, to enjoy life more fully, and to appreciate friends and family more. A desire to be cancer-free and stay free of concerns over cancer recurrence are on the top of the patients' priorities. The extracted topics may complement conventional quality-of-life survey questionnaires. Patients' idiosyncratic concerns and coping strategies in a life-threatening illness, reported in their own words, are now available to researchers who want to go beyond conventional standardized questionnaires. For example, a desire to resume employment (see topic 2 in Fig. 4) has high importance throughout the surgery and recovery process, implying a concern over financial security

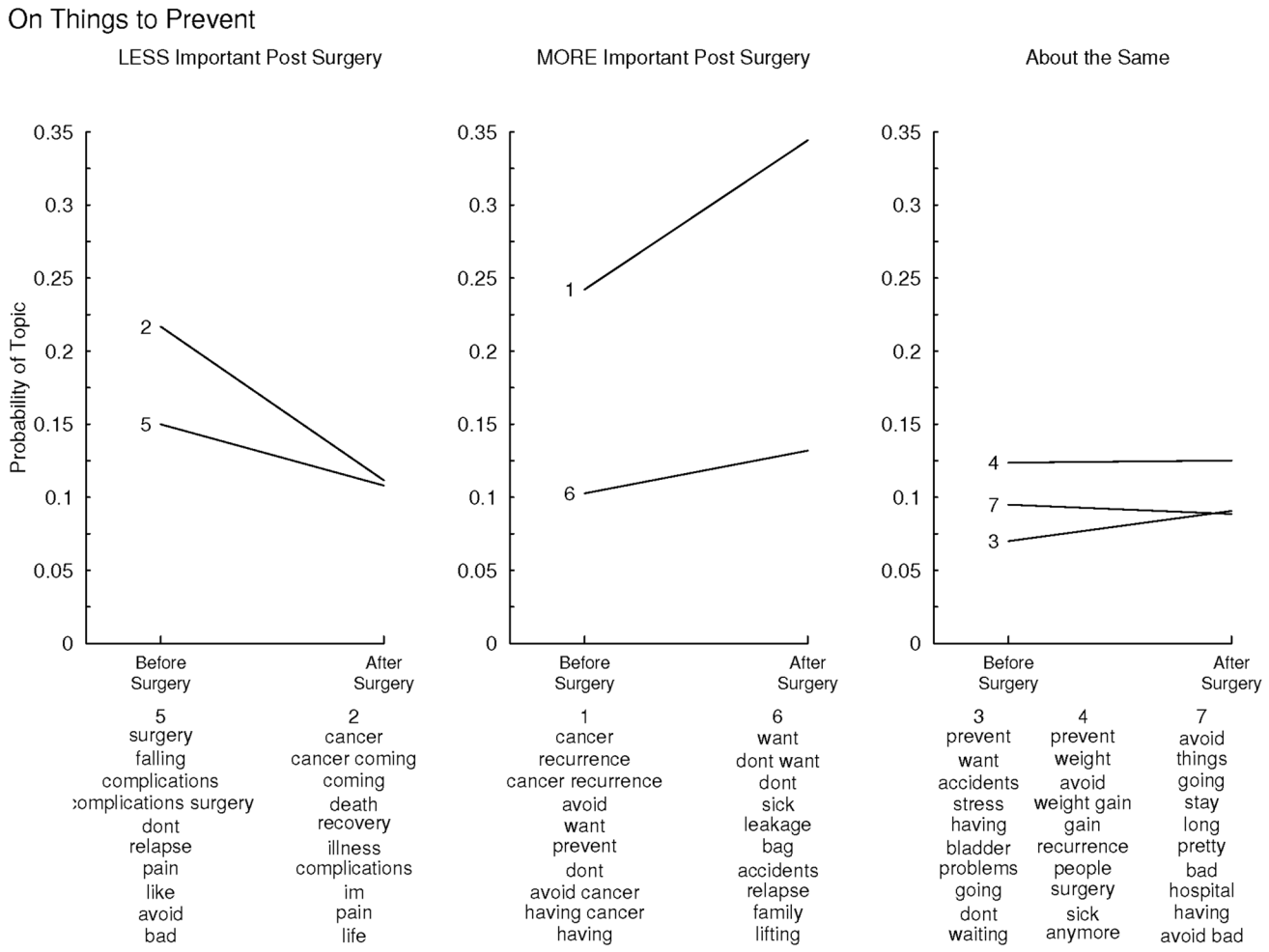

Fig. 5 Changes in patients' priorities on things they wanted to prevent before and after bladder cancer surgery

during retirement and perhaps for loved ones. Patients explicitly state that they want to "be healthy, be financially comfortable," "live the rest of my life in a comfortable way," and "make enough money so my family lives comfortably." The financial aspects of one's QOL are not typically a primary domain in existing measures for cancer patients. Different individuals, depending on their priorities and goals, may respond very differently to the same surgical complications and/or incontinence concerns. Asking patients to express their personal goals and priorities may shed light on puzzling results from conventional survey questionnaires alone. This is the beginning of a line of inquiry not typically attempted in behavioral research in health.

We believe we are the first in applying LDA to patient's verbatim utterances on their goals during cancer treatment and recovery. We have observed the strengths of LDA, primarily in the automation of a rudimentary but pragmatic interpretation of free-text data in an otherwise time-consuming conventional coding by a qualitative data analyst. The model parameters were leveraged to reflect changes in priorities, quantities not typically obtained in conventional quality data analysis. Griffiths and Steyvers [18] showed that, with abundant data and structured keywords in abstracts of scientific articles, LDA yielded high concordance with the keywords given by the authors themselves. LDA may be more consistent and less prone to interpretation biases because it follows a stochastic yet concrete algorithm.

All these advantages notwithstanding, our LDA application identifies several limitations that may guide future research. First, short documents tend to yield topics that are crude and vague. Our patients typically gave 2 to 3 short sentences per probe. LDA is known to not work well in short documents such as Tweets [34]. There is simply not enough information for LDA to leverage. Lack of information plagues any statistical procedure, which is not an inherent problem specific to LDA. For instance, in an exploratory factor analysis, sometimes the factors are not distinct, and sometimes the questionnaire items do not clearly load onto the latent factors. Documents with hundreds of words produce strong and coherent topics [17–19]. Therefore, moving

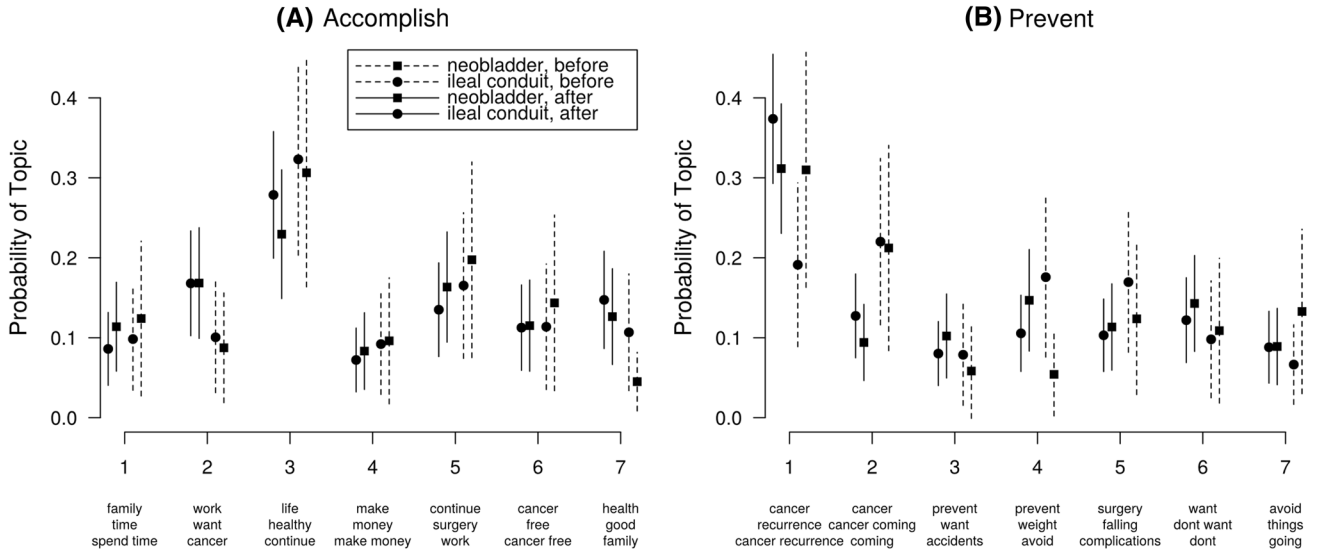

Fig. 6 Changes in patients' priorities stratified by urinary diversion and time. Subplot **a** shows subtle differences. For example, after surgery, slightly higher weights were given to topics 2 and 7. Patients lowered weights to topic 3, slightly more so in the neobladder patients than the ileal conduit patients. Subplot **b** shows that, on things patients wanted to prevent, "cancer recurrence" was highly salient for all patients. There appeared to be a slightly greater reduction in topic 1 in the ileal conduit patients

ID	Baseline	6 Months Post-Surgery
526	survive[1] the surgery[4] have a manageable[5] recovery[5] process[6] return[2] to a normal[5] life[5] after recovery[5]	i would like[7] to get more information about colostomy reversal[7]
418	get through with operation[5] return[2] to normal[5]	live[5] a long[5] and happy[4] life[5]
504	get healthy[6] being actively[6] employed[1] after surgery[4]	extending interval between urination[3] improve[3] sleep[3] patterns[3]

Fig. 7 Selected examples of free-text goal statements on things patients would like to accomplish at baseline and 6 months post-surgery. The superscripts indicate the topics to which individual words were assigned in a single free-text entry. The grayscale contrast level of the words reflects the probability of a word belonging to the most prevalent topic in the entry

forward, LDA should be best applied to diaries or electronic records longer than a handful of sentences. However, cancer patients undergoing intensive treatments may have difficulty keeping a diary. Electronic medical records may be an alternate source of information. A recent study showed that physicians may already use open-text entries in electronic medical notes to document cancer patients' symptoms [35]. Another limitation is that the inclusion of bigrams may cause concerns on the conditional independence assumption in LDA. For example, "cancer recurrence" may not be strictly independent from either "cancer" or "recurrence." However, we have found no pragmatic computation solutions or alternative model to incorporate correlations between words. Thus, we must be mindful of this assumption and interpret the results with care.

Further readings

While this may be an early attempt at using LDA techniques in quality-of-life research, there are many resources available for those interested in LDA. Here are a few that we found instructive. There is the publicly accessible electronic book by Tufts [36]. Steyvers and Griffiths [19] and Reed [37] provide the basics, and Ponweiser [38] offers a step-by-step guide on how to reproduce the analyses in Griffiths and Steyvers [18]. A head-to-head comparison between LDA and conventional qualitative data analysis by Baumer et al. [20] may be particularly interesting to researchers who have always relied on human interpretation of text data. These authors believe that LDA may help

a human coder to organize documents into groups that are likely to be thematically coherent. To go beyond a basic LDA, Rosen-Zvi et al. [39] offer the author-topic model to map authors with topics. Roberts et al. [40] offer a structural topic model to analyze covariate effects (e.g., gender of the author) on topic prevalence and applied the model on open-ended survey responses. Recent methodology work encourages researchers to always include sensitivity analyses as an integral part of LDA [41, 42]. For example, the Dirichlet prior for the document-topic matrix can be assigned different values to encourage fewer or more topics. Then the analyst can select the model with the best interpretable number of topics. Methodological enhancements have continuously been made since the original LDA work. We hope that the current study offers a useful introduction to LDA in Quality-Of-Life research.

Acknowledgements The authors thank Patient-Centered Outcomes Research Institute Grant ME-1306-00781 (PI: Rapkin); National Institute of Health Grant P30 CA008748 to Memorial Sloan Kettering Cancer Center; Sidney Kimmel Center for Prostate and Urological Cancers at Memorial Sloan Kettering Cancer Center, Pin Down Bladder Cancer; and the Michael A. and Zena Wiener Research and Therapeutics Program in Bladder Cancer.

Funding This study was funded by (1) Patient-Centered Outcomes Research Institute Grant ME-1306-00781 (PI: Rapkin); (2) National Institute of Health Grant P30 CA008748 to Memorial Sloan Kettering Cancer Center; and (3) Sidney Kimmel Center for Prostate and Urological Cancers at Memorial Sloan Kettering Cancer Center, Pin Down Bladder Cancer, and the Michael A. and Zena Wiener Research and Therapeutics Program in Bladder Cancer (PI: Bochner).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval for human subject research All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. National Cancer Institute, S. P. (2018). *Cancer stat facts: Bladder cancer*. Retrieved from https://seer.cancer.gov/statfacts/html/urinb.html.
2. Rapkin, B. (2000). Personal goals and response shifts: Understanding the impact of illness and events on the quality of life of people living with AIDS. In C. A. Schwartz & M. A. G. Sprangers (Eds.), *Adaptation to changing health: Response shift in quality-of-life research* (pp. 53–71). Washington, DC: American Psychological Association.
3. Rapkin, B., & Schwartz, C. E. (2004). Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. *Health Quality of Life Outcomes, 2*, 14.
4. Rapkin, B. D., Smith, M. Y., DuMont, K., Correa, A., Palmer, S., & Cohen, S. (1993). Development of the ideographic functional status assessment: A measure of the personal goals and goal attainment activities of people with AIDS. *Psychology and Health, 9*, 111–129.
5. Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality-of-life research: A theoretical model. *Social Science and Medicine, 48*, 1507–1515.
6. Schwartz, C. E., Finkelstein, J. A., & Rapkin, B. D. (2017). Appraisal assessment in patient-reported outcome research: methods for uncovering the personal context and meaning of quality of life. *Quality of Life Research, 26*(3), 545–554. https://doi.org/10.1007/s11136-016-1476-2.
7. Li, Y., & Rapkin, B. (2009). Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *Journal of Clinical Epidemiology, 62*(11), 1138–1147. https://doi.org/10.1016/j.jclinepi.2009.03.021.
8. Rapkin, B. D., & Schwartz, C. E. (2016). Distilling the essence of appraisal: a mixed methods study of people with multiple sclerosis. *Quality of Life Research, 25*(4), 793–805. https://doi.org/10.1007/s11136-015-1119-z.
9. Morganstern, B. A., Bochner, B., Dalbagni, G., Shabsigh, A., & Rapkin, B. (2011). The psychological context of quality of life: a psychometric analysis of a novel idiographic measure of bladder cancer patients' personal goals and concerns prior to surgery. *Health Quality of Life Outcomes, 9*, 10. https://doi.org/10.1186/1477-7525-9-10.
10. Hart, S., Skinner, E. C., Meyerowitz, B. E., Boyd, S., Lieskovsky, G., & Skinner, D. G. (1999). Quality of life after radical cystectomy for bladder cancer in patients with an ileal conduit, cutaneous or urethral Kock pouch. *The Journal of Urology, 162*, 77–81.
11. Dutta, S. C., Chang, S. C., Coffey, C. S., Smith, J. A. Jr., Jack, G., & Cookson, M. S. (2002). Health related quality of life assessment after radical cystectomy: Comparison of ileal conduit with continent orthotopic neobladder. *Journal of Urology, 168*, 164–167.
12. Gerharz, E. W., Weingartner, E., Dopatka, T., Kohl, U. N., Basler, H. D., & Riedmiller, H. N. (1997). Quality of life after cystectomy and urinary diversion: Results of a retrospective interdisciplinary study. *Journal of Urology, 158*, 778–785.
13. Hobisch, A., Tosun, K., Kinzl, J., Kemmler, G., Bartsch, G., & Holtl, L. (2001). Life after cystectomy and orthotopic neobladder versus ileal conduit urinary diversion. *Seminars in Urologic Oncology, 19*, 18–23.
14. Yang, L. S., Shan, B. L., Shan, L. L., Chin, P., Murray, S., Ahmadi, N., & Saxena, A. (2016). A systematic review and meta-analysis of quality of life outcomes after radical cystectomy for bladder cancer. *Surgical Oncology, 25*(3), 281–297. https://doi.org/10.1016/j.suronc.2016.05.027.
15. Ali, A. S., Hayes, M. C., Birch, B., Dudderidge, T., & Somani, B. K. (2015). Health related quality of life (HRQoL) after cystectomy: comparison between orthotopic neobladder and ileal conduit diversion. *Eur J Surg Oncol, 41*(3), 295–299. https://doi.org/10.1016/j.ejso.2014.05.006.
16. Cerruto, M. A., D'Elia, C., Siracusano, S., Gedeshi, X., Mariotto, A., Iafrate, M.,.. . Artibani, W. (2016). Systematic review and meta-analysis of non RCT's on health related quality of life after radical cystectomy using validated questionnaires: Better results with orthotopic neobladder versus ileal conduit. *European Journal of Surgical Oncology, 42*(3), 343–360. https://doi.org/10.1016/j.ejso.2015.10.001.

17. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022. http://jmlr.org/papers/v3/blei03a.html. doi.
18. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.
19. Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara & S. Dennis, & K. W. (Eds.), *Latent semantic analysis: A road to meaning*. Hillsdale: Laurence Erlbaum.
20. Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology, 68*(6), 1397–1410.
21. Mittal, V., Kaul, A., Sen Gupta, S., & Arora, A. (2017). Multivariate features based Instagram post analyiss to enrich user experience. *Procedia Computer Science, 122*, 138–145.
22. Glickman, M., Brown, J., & Song, R. (2018). *Assessing authorship of Beatles songs from musical content: Bayesian classification modeling from bags-of-words representations*. Paper presented at the 2018 Joint Statistical Meeting, Vancouver, Canada. https://ww2.amstat.org/meetings/jsm/2018/onlineprogram/AbstractDetails.cfm?abstractid=329336.
23. Simon, S. H. (2018). A songwriting mystery solved: Math Proves John Lennon wrote 'in my life': National Public Radio.
24. Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M.,… Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791.
25. Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*(1), 150–159. https://doi.org/10.1016/j.paid.2017.12.018.
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,.. . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
27. Nikita, M. (2016). ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters: R package version 0.2.0.
28. Arun, R., Suresh, V., Veni Madhavan, V. C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran & V. Pudi (Eds.), In *Advances in knowledge discovery and data mining* (pp. 391–402). Heidelberg: Springer Berlin.
29. Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive lDA model selection. *Neurocomputing—16th European Symposium on Artificial Neural Networks, 72*, 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011.
30. Deveaud, R., SanJuan, É, & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique, 17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84.
31. Ipeirotis, P. (2007). Visualizing the Dirichlet. Retrieved from https://www.behind-the-enemy-lines.com/2007/10/visualizing-dirichlet.html.
32. Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1–54.
33. Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13.
34. Hong, L., & Davison, B. D. (2010). *Empirical study of topic modeling in Twitter*. Paper presented at the Proceeding SOMA '10 Proceedings of the First Workshop on Social Media Analytics, Washington DC.
35. Forsyth, A. W., Barzilay, R., Hughes, K. S., Lui, D., Lorenz, K. A., Enzinger, A.,.. . Lindvall, C. (2018). Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. *J Pain Symptom Manage, 55*(6), 1492–1499. https://doi.org/10.1016/j.jpainsymman.2018.02.016.
36. Tufts, C. (2018). *The little book of LDA an overview of Latent Dirichlet Allocation & Gibbs Sampling*. Retrieved from https://ldabook.com.
37. Reed, C. (2012). *Latent Dirichlet allocation: Towards a deeper understanding*. Retrieved from http://obphio.us/pdfs/lda_tutorial.pdf.
38. Ponweiser, M. (2012). *Latent Dirichlet Allocation in R*. WU Vienna University of Economics and Business. Retrieved from http://epub.wu.ac.at/id/eprint/3558.
39. Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence*, 487–494. https://dl.acm.org/citation.cfm?id=1036902.
40. Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103.
41. Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2019). A review of best practice recommendations for text analysis in R (and a user-friendly App). *Journal of Business and Psychology, 33*(4), 445–459. https://doi.org/10.1007/s10869-017-9528-3.
42. Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A.,… Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754.