Bridging Qualitative and Quantitative Methods for User Modeling: Tracing Cancer Patient Behavior in an Online Health Community

Zachary Levonian, Drew Richard Erikson, Wenqi Luo, Saumik Narayanan,
Sabirat Rubya, Prateek Vachher, Loren Terveen, Svetlana Yarosh
GroupLens Research, University of Minnesota, Minneapolis, MN 55455, USA
{levon003,eriks074,luoxx498,naray114,rubya001,vachh007,terveen,lana}@umn.edu

Abstract

Researchers construct models of social media users to understand human behavior and deliver improved digital services. Such models use conceptual categories arranged in a taxonomy to classify unstructured user text data. In many contexts, useful taxonomies can be defined via the incorporation of qualitative findings, a mixed-methods approach that offers the ability to create qualitatively-informed user models. But operationalizing taxonomies from the themes described in qualitative work is non-trivial and has received little explicit focus. We propose a process and explore challenges bridging qualitative themes to user models, for both operationalization of themes to taxonomies and the use of these taxonomies in constructing classification models. For classification of new data, we compare common keyword-based approaches to machine learning models. We demonstrate our process through an example in the health domain, constructing two user models tracing cancer patient experience over time in an online health community. We identify patterns in the model outputs for describing the longitudinal experience of cancer patients and reflect on the use of this process in future research.

1 Introduction

Social media data offers the promise of human behavioral insight that is temporally linked and captured contemporaneously with that behavior (Olteanu, Varol, and Kiciman 2017). While much of this data is unstructured, methods for identifying patterns—such as supervised machine learning—are increasingly being used to extract structure for further analysis (Kuksenok et al. 2012). Developing computational models to do this data extraction requires defining a taxonomy: the explicit structure to extract from the underlying social media data. For example, to identify targets of online hate, Salminen et al. defined a complex taxonomy capturing the nuances of hate speech (2018). Researchers use taxonomies that are created by experts, derived unsupervised from the data, or adapted from prior work. In this paper, we create taxonomies directly from themes identified in qualitative research.

The incorporation of qualitative research into user modeling is beneficial because mixed methods enable researchers to triangulate their understandings, refine theory, and make use of the strengths of both qualitative and quantitative methods (Shah and Corley 2006). However, the themes and implications described in qualitative work cannot be taken "as is" as a taxonomy. Even when qualitative themes have an appropriate level of granularity for the research question at hand (Zhang, Culbertson, and Paritosh 2017), an explicit mapping of themes to divisions in the data must be constructed to derive quantitative models. In this paper, we focus on the problem of *bridging* existing qualitative work to computational user models built from social media text data. We consider bridging as a two-stage process involving (1) *operationalization* of qualitative themes into a taxonomy, and (2) *classification* of the data based on that taxonomy. This process has received little explicit focus in prior research; we argue that developing an operationalization process for qualitative themes can better enable the incorporation of qualitative insights into user model taxonomies.

We implement an operationalization method for identifying taxonomic boundaries for two qualitative frameworks in the cancer domain, identifying critical challenges in this process. These taxonomies seek to support modeling based on user-generated text. Two common approaches to this are identifying keywords that signify inclusion in a particular taxonomic category versus supervised machine learning based on human annotation of text into the taxonomic categories. We conduct empirical comparisons of these two approaches for classification of categories in the derived taxonomies.

Our present study is motivated by research questions related to cancer patients' labor and their use of online health communities (OHCs). Substantial sources of social media data capture the experiences of cancer patients, but no existing operationalizations bridge these data to the extensive qualitative work describing the experiences and needs of cancer patients. By bridging existing qualitative health theories into computational models of patients' OHC use, these models can inform the delivery of digital services (Jacobs, Clawson, and Mynatt 2014). From qualitative frameworks developed by Jacobs, Clawson, and Mynatt (2016) and Hayes et al. (2008), we iteratively develop taxonomies for classifying cancer patient responsibilities and temporal cancer phases. We use supervised machine learning to construct computational models that trace cancer patients' experiences through their OHC posts.

Preprint. To appear in the proceedings of ICWSM 2020.
Last edited July 31, 2019.

The contributions of this work are (1) an articulation of a bridging process between qualitative themes and quantitative models, (2) a comparison of two classification methods for taxonomies—supervised machine learning and keyword-based classifiers, and (3) the extension of two existing qualitative frameworks to a novel social media context. Our proposed bridging process builds towards researcher triangulation of findings across methodological approaches to build more robust user models. We describe our application of the two stages of the bridging process in the Operationalization and Classification sections, then reflect on the two models' validity and predictions in the Model Analysis section. In the Discussion, we identify implications for future researchers using this method.

2 Related Work

Social media data contains traces of human activity that, if structured, can reveal human behavior (Olteanu, Varol, and Kiciman 2017; Kulkarni et al. 2018). The unstructured text of social media data constitutes a trace of human behavior, and those texts can inform us about humans' behaviors and beliefs (Pennebaker, Mehl, and Niederhoffer 2003). Social media text has been used to infer ideology (Zhang and Counts 2015), personality (Kulkarni et al. 2018), nutrition (De Choudhury, Sharma, and Kiciman 2016), and other aspects of human experience. Behavioral analysis via social media is often used to explore human behavior during periods of change like the birth of a child or a health crisis, as we do here (De Choudhury et al. 2013; Paul, White, and Horvitz 2015). In the next two sections, we discuss background on taxonomies and classification.

2.1 Operationalization of Taxonomies

To create user models, researchers define taxonomic categories of behavior from three non-exclusive sources: unsupervised machine learning, experts, and qualitative inquiry. Unsupervised machine learning defines categories and the boundaries between them directly from patterns in the data, but it can be hard to validate automatically-inferred patterns or to determine their relevance to the research question at hand (Sachdeva, Kumaraguru, and De Choudhury 2016). But, questions can be asked and answered using the resulting taxonomies without strictly adhering to prior expectations (Concannon et al. 2018).

Expert-derived taxonomies are built from close collaboration with domain experts (Liu, Weitzman, and Chunara 2017; Kiciman, Counts, and Gasser 2018), a manual reading of existing literature in the target domain (Paul, White, and Horvitz 2015; Zhang et al. 2017), or from codebooks of keywords uncovered from "expert" Internet sources (Huang et al. 2017). While these taxonomies gain validity from their basis in expert knowledge, this top-down approach may limit the ability to detect novel categories in the data and in many cases the relevant domain expert may not exist.

An alternative is to operationalize a taxonomy from qualitative work, which is the approach we explore. Zhang, Culbertson, and Paritosh aimed to develop a taxonomy from prior work, but found that existing work was too narrow, instead iteratively developing their own taxonomy with experts (2017). Singer et al. used hand-coded survey responses to construct a taxonomy and validated it with an additional survey (2017). While it is ideal for quantitative researchers to collaborate closely with qualitative ones on the same research questions, requiring that qualitative and quantitative experts work together synchronously limits the community's ability to learn from the existing body of qualitative work (Morgan 1998). By articulating a process of taxonomy operationalization from qualitative themes, user models benefit from existing bottom-up work.

2.2 Classification of Social Media Data

Once taxonomies are defined, two primary approaches are used to classify available text data: the use of specific word patterns by lexical analysis of texts through the discovery of words closely related to a desired category (Fast, Chen, and Bernstein 2016) (i.e. keyword-based approaches) and supervised machine learning (ML).

Keyword-based approaches are appealing because they are interpretable and require no human annotation of data. These approaches often involve soliciting keywords from an expert (Kiciman, Counts, and Gasser 2018). The line between building a taxonomy from "constructs of interest" (Fast, Chen, and Bernstein 2016) and selecting keywords to use in that taxonomy is often blurred e.g. in (Geiger and Halfaker 2017). Such approaches run the risk of missing important variants of the phenomena under study (Salminen et al. 2018) and may need additional human validation (Birnbaum et al. 2017).

In contrast, supervised ML can result in higher precision than keyword lists on social media data (Birnbaum et al. 2017) and find patterns that are more generalizable and robust (Zhang, Culbertson, and Paritosh 2017). We compare supervised ML to keyword-based approaches to further articulate the trade-offs of interpretability versus robustness.

3 Study Design

We investigate the proposed bridging process in the context of cancer patients' OHC posts. In this section, we provide the relevant qualitative background (3.1), describe the OHC (3.2), and discuss the selected data (3.3). Subsequent sections describe the operationalization, classification, and finally analysis of the model outputs, with each section addressing the methods used and our results.

3.1 Cancer patients and OHCs

Online health communities (OHCs) are used by patients and caregivers to seek social support (Gui et al. 2017). We focus on patient use of CaringBridge, an online health community. Responding to the call for catalyzing social support by understanding and enhancing OHCs (Skeels et al. 2010), we use unstructured text of patient posts to model their use of CaringBridge. In contrast, most prior user modeling health research has relied primarily on structured health information like self-reported condition (Tamersoy, De Choudhury, and Chau 2015). In the next sections, we discuss the theoretical foundations from which we operationalize taxonomies.

Phases and transitions The concept of cancer *phases* are used by patients to self-characterize their needs (Eschler, Dehlawi, and Pratt 2015), in medical research to organize programs of care (O'Brien et al. 2014), and as the basis for prior HCI research (Jacobs, Clawson, and Mynatt 2014). In this work, we adopt the phase model of cancer articulated by Hayes et al. (2008) and adapted by Jacobs, Clawson, and Mynatt (2016) to describe commonalities in patients' experiences of their cancer journeys.

While we are the first to use Hayes et al.'s phases in quantitative modeling, Wen and Rose used an earlier iteration of this phase model to identify cancer disease trajectories, although their emphasis is on phase boundary identification via automatic event extraction (2012). Liu, Weitzman, and Chunara utilized supervised ML of social media posts to identify drinking behavior through a series of discrete stages (2017). Although conceptually similar to phases, their stage taxonomy was developed with the input of domain experts. We utilize a similar modeling approach and follow their lead in the use of active learning. Other established stage/phase models, like the widely used transtheoretical model (TTM) of health behavior changes, are used as the basis for taxonomies that are tweaked by experts (MacLean et al. 2015). The TTM has been refined through both theory-building and empirical validation over many years (Prochaska and Velicer 1997); in contrast, the Hayes et al. phase model is based directly on qualitative work and has not yet been explored in diverse contexts. Our operationalization contributes to a broader effort of theoretical refinement (Adcock and Collier 2001). On the quantitative side, concepts similar to phases have been operationalized via discrete observable keyword-patterns e.g. for the identification of recovery events (Chancellor, Mitra, and De Choudhury 2016).

Cancer journey framework Jacobs, Clawson, and Mynatt articulated a cancer journey framework (CJF) from qualitative interviews with cancer patients (2016). The CJF is organized into three dimensions: responsibilities, challenges, and how the cancer journey influenced patients' daily life. We focus only on the responsibilities, defined by Jacobs, Clawson, and Mynatt as "the multiple tasks that are placed on patients during each of the cancer journey phases", referring to the phases described by Hayes et al. (2008). Qualitative exploration of the dataset indicated that the other two dimensions were seldom visible in the details of patients' posts. The responsibilities and their corresponding phase assignments are listed in Table 1, along with abbreviated responsibility codes used where space is limited. Responsibilities are purposeful and goal-oriented tasks that are required of the patient because of a cancer diagnosis; for example, one task associated with the Preparation responsibility would be getting a wig fitting in advance of anticipated hair-loss due to treatment.

CaringBridge is designed to support patients' communication with their extended support networks (Massimi, Dimond, and Le Dantec 2012). Therefore, we expected that patients would discuss their responsibilities with their CaringBridge support network. While there is a tension between managing self-presentation and "sharing information related to specific needs and desires" (Newman et al. 2011), we treat patient's discussions of their responsibilities on CaringBridge as veridical representations (Star and Strauss 1999) of their real-world responsibilities. In particular, we assume patients may *omit* responsibilities from discussion on CaringBridge but will not *fabricate* them, such that our computational models can be taken as a high-precision view of cancer patients' responsibilities. By classifying these responsibilities on CaringBridge, we aim to conceptualize patients' communication of their labor.

We selected the CJF and the Hayes et al. phase model for use in this bridging process based on our broader research question, which was related to understanding patient labor needs over time so that we can design more effective, personalized online interventions to meet or reduce those needs. We offer no guidance on the identification and selection of qualitative frameworks for this bridging process other than alignment with the research question of interest; this is an important theoretical problem that deserves additional attention in future work. We acknowledge a broader tension in qualitative research regarding the generalizability of qualitative work; while not all qualitative work is intended to generalize, we select frameworks that comprise "in-depth analysis of specific, local phenomena, with the intention of generalizing to other sites and other people" (Muller et al. 2016). Our bridging process builds on that intention.

End-of-life The CJF was developed through retrospective patient interviews. Thus, one limitation is that it necessarily omits cancer journeys that conclude with the death of the patient. OHCs have a role to play in end-of-life situations, as the use of technology to aide in communication and support coordination is important to patients' quality of life during hospice (Heyland et al. 2006). Online hospice communities have been studied for their role facilitating social support during hospice care (Buis 2008), but OHCs like CaringBridge have not been specifically investigated in this context. While most studies of technology use at end-of-life have relied on retrospective interviews (Ferguson et al. 2014), CaringBridge provides an opportunity to explore the use of technology at end-of-life contemporaneous with the dying experience. As communication and decision-making labor passes from the patient to their caregivers near death (Prendergast and Puntillo 2002), we expect that many aspects cannot be captured via the patient's own writing; however, these data remain a unique opportunity to analyze responsibilities articulated during the end-of-life phase.

3.2 CaringBridge research collaborative

This work was conducted during a research collaboration between CaringBridge (CB) and the University of Minnesota[1]. CB is a global, nonprofit social network dedicated to helping family and friends communicate with and support loved ones during a health journey.

[1] CaringBridge has engaged in collaborations with several research groups and has requested that we represent their organization consistently across research efforts, thus this section may be nearly identical to text describing the platform in other papers.

Code	Responsibility	Phase
CO	Communicating the disease to others	PT
IF	Information filtering and organization	PT
CD	Clinical decisions	PT
PR	Preparation	PT
ST	Symptom tracking	T
CS	Coordinating support	T
SM	Sharing medical information	T
CP	Compliance	T
MT	Managing clinical transition	T
FM	Financial management	T
CM	Continued monitoring	NED
GB	Giving back to the community	NED
BC	Health behavior changes	NED

Table 1: Patient responsibilities in the CJF and the phase within which that responsibility was organized. Phase is either *pretreatment* (PT), *treatment* (T), or *no evidence of disease* (NED).

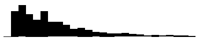
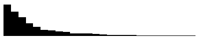
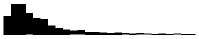
Platform description CaringBridge.org offers individual *sites* for users—free, personal, protected websites for patients and caregivers to share health updates and gather their community's support. Each site prominently features a *journal*, which is a collection of multiple health *updates* by or about a patient. Updates are comprised of text and are timestamped with a creation date and time. This terminology reflects that used by Ma et al. (2017).

Data description and ethical considerations The complete dataset used for this analysis includes de-identified information from 588,210 CaringBridge sites created between June 1, 2005 and June 3, 2016. The site data were acquired through collaboration with CB leadership in accordance with CB's Privacy Policy & Terms of Use Agreement. This study was reviewed and deemed exempt from further IRB review by the University of Minnesota Institutional Review Board. We acknowledge the tension in HCI between open data dissemination (Hornbk et al. 2014) and the ethical necessity to protect participants' rights and privacy (Bruckman et al. 2017). As CB data are highly sensitive, we opt not to publicly release the dataset used for analysis in this paper or to use crowdsourcing for annotation. In compromise between replicable science and the ethical protection of participants' privacy, we welcome inquiries about the dataset by contacting the authors. We do release our taxonomy definitions and analysis code.[2]

3.3 Study data selection

Most sites in the CB dataset are not relevant to this study, as the CJF and phase model articulate themes only for cancer patients. We include only sites that self-reported cancer as the health condition category at the time of site creation. For ethical reasons, we further omit sites deleted by the site authors. To account for shifts in the design and demographics of CaringBridge over time, we include only sites created in 2009 or later. We focus on completed sites, ones with their final journal updates made before April 1st, 2016 (two months before the end of the dataset's span). We analyze only sites active enough to capture part of the patient's cancer journey, which we define as sites with at least five journal updates spanning at least one month.

Finally, as sites may have multiple authors and we are only interested in sites written by patients themselves, we exclude sites in which fewer than 95% of the updates were authored by the patient. We identify updates as patient-authored or not using a binary Vowpal Wabbit logistic regression classifier with L2 regularization (Langford, Li, and Strehl 2007). Hashed unigram and bigram bag-of-words features were used. During data exploration, two researchers annotated updates as evidently patient-authored or not. Agreement was generally high (Cohen's $\kappa = 0.72$), disagreements primarily arising from very short updates. To improve classifier accuracy and address biases potentially introduced via non-random sampling of updates for annotation, we conducted several rounds of uncertainty sampling, resulting in a training set of 1,035 updates. This classifier achieved an accuracy of 92.5% on a held-out validation set of 258 updates, which we determined to be sufficient for the accurate identification of sites primarily authored by patients. During random sampling of sites for the human annotation described subsequently, we observed no sites that were not primarily patient-authored. After the exclusion of sites based on the authorship classifier, we selected 4,946 sites for subsequent analysis (described in Table 2) containing 158,597 updates.

Journal Updates	Median: 22 updates M=32.1; SD=43.7	
Site Visits	Median: 1017 visits M=2099.2; SD=4136.9	
Survival Time	Median: 8.2 months M=12.9; SD=13.3	

Breast	2752 (55.6%)	Leukemia	209 (4.2%)
Lymphoma	597 (12.1%)	Ovarian	169 (3.4%)
Other	380 (7.7%)	Lung	168 (3.4%)
Not Specified	257 (5.2%)	Myeloma	120 (2.4%)
Colorectal	225 (4.5%)	Brain	69 (1.4%)

Table 2: Descriptive info about the 4,946 selected CB sites. Survival time is the time elapsed between the first and last journal update on a site.

4 Operationalization

4.1 Operationalization Methods

We define operationalization as the construction of a structured taxonomy from description of themes in existing qualitative theory. Following Zhang et al. (2018), we suggest that not all themes may be useful in the target social media context; rather, the operationalization process creates a "shared vocabulary" that identifies conceptually coherent categories. Echoing Figueiredo et al. (2017), the qualitative

[2] github.com/levon003/icwsm-cancer-journeys

Phase	Occurrence	Disagreement	κ
PT	7.4%	5.5%	0.91
T	69.7%	7.4%	0.94
EOL	1.9%	0.2%	—
NED	6.4%	3.6%	0.95
Overall	99.62%	10.2%	0.93

Table 3: Annotated phase occurrence proportions and IRR. Disagreement is the percentage of a phase's occurrence in multi-annotated updates with disagreement. Cohen's κ is reported for two coders' annotations of 31 sites containing 619 updates; none of these sites contained EOL updates. Overall stats describe updates annotated with any phase.

framework is a lens— a "conceptual framework to recognize and compare"—to understand the relationship between patients' writing on CB and the taxonomic categories.

Tangibly, operationalization involves a mapping between indicators in the data and particular qualitative themes. These mappings define the categories in the taxonomy. We operationalize two taxonomies from the phase and responsibility frameworks discussed in Section 3.1. In a social media context, data indicators are units of text that relate to the qualitative framework. For example, we defined a particular responsibility to be present in an update if the author explicitly acknowledges having done a related task or having a need for a related task; a patient's description of a task *indicates* the presence of a responsibility. The taxonomy codebook describes which task descriptions indicate particular categories. We focus on indicators of responsibilities that require human but not specific-domain expertise to identify (Zhang et al. 2018); in particular, it's not at all clear what if any domain expertise could exist for responsibilities given that the indicators are non-medical.

For both phases and responsibilities, we created initial category descriptions directly from the theme descriptions in the corresponding qualitative work. We conducted multiple rounds of annotation followed by discussion to resolve disagreements, resulting in updates to the taxonomy in the form of examples and guidance for annotators. Such iterative processes are widely used in codebook development (Zhang et al. 2017; Geiger and Halfaker 2017; Adcock and Collier 2001). Annotators could assign as many responsibility labels to an update as evidence indicated, while phase labels were initially treated as mutually exclusive. Four researchers participated in codebook development and annotation, all familiar with CaringBridge data but not medically trained. Two of these researchers functioned as primary annotators, together annotating the majority of labeled data. Each round of annotation consisted of the primary annotators independently labeling 20 randomly sampled sites and computing Cohen's κ to assess the level of inter-rater reliability (IRR). After taxonomies were defined, we annotated additional sites to provide data for the training of classification models.

Responsibility	Occurrence	Disagreement	κ
CO	1.3%	2.3%	0.00
IF	7.5%	17.0%	0.06
CD	3.4%	6.1%	0.21
PR	14.4%	26.2%	0.22
ST	20.4%	32.9%	0.15
CS	9.2%	12.9%	**0.43**
SM	52.4%	16.7%	**0.57**
CP	46.6%	26.8%	**0.45**
MT	12.3%	22.9%	0.13
FM	1.8%	2.6%	**0.42**
CM	5.0%	7.4%	0.32
GB	2.6%	4.8%	**0.42**
BC	2.6%	4.4%	**0.44**
Overall	96.19%	85.2%	0.10

Table 4: Annotated responsibility occurrence proportions and IRR. Disagreement is the percentage of a responsibility's occurrence in multi-annotated updates with annotator disagreement. Cohen's κ is reported for two coders' annotations of 20 sites containing 471 updates; the six emphasized responsibilities are used for future classification. Overall stats describe annotated updates containing at least one responsibility, where κ evaluates agreement with the requirement that both annotators agree on all responsibilities for that journal.

4.2 Operationalization Results

We identified two challenges common to both phase and responsibility operationalization: interrogating thematic boundaries and mapping the conceptual to the observable.

Interrogating thematic boundaries We experienced challenges developing distinct boundaries between themes from the indicators in the text. For the phase taxonomy, we began our exploration using all five phases described by Hayes et al. (2008): screening and diagnosis, information seeking, acute care and treatment, no evidence of disease, and chronic care and disease management. We observed that "screening and diagnosis" and "information seeking" were intertwined; updates in the first few weeks of a site described experiences with no clear correspondence with exactly one of the phase themes. We merged these themes into a single "pretreatment" phase that encapsulates the qualitative descriptions of both, constructing a taxonomy with four categories: pretreatment (PT), treatment (T), no evidence of disease (NED), and chronic care and disease management (EOL). Hayes et al. included discussion of the valid transitions between phases (depicted in Figure 1 as arrows), which we found to cohere with the data patterns we observed i.e. we observed no transitions other than those indicated. To complete the phase taxonomy, additional rounds of annotation focused on clarifying the most relevant thematic boundaries—PT/T and T/NED—and adding examples to the annotation guidance e.g. identifying medical port insertion as a common transition from PT to T.

For the responsibility taxonomy, we observed two distinct types of indicators that referred to the CJF's Support Management responsibility. The patients' literal descriptions of coordinating support blended with the sharing of medical information by authoring the CB update. We split Support Management into two new responsibilities—Coordinating Support and Sharing Medical Info—each defined from subsets of the CJF's description of Support Management. This split enabled us to disentangle acknowledgements by the patient of support coordination apart from the act of writing updates on CB. Pooling could be used for later analyses, but we embraced the suggestive split in the data. With 13 responsibilities, we had many more boundaries to negotiate and discuss, finding that a single task indicator may correspond with multiple responsibilities in an ambiguous way.

Mapping conceptual to observable In mapping conceptual themes to observable units of data, some indicators were ambiguously linked to one or more categories. For the phase taxonomy, we observed updates that described transitions between phases or for which phase could not be confidently identified. To address this challenge, we allowed annotators to select up to two phases for a single update and introduced an "Unknown" checkbox to the annotation interface to indicate uncertainty.

For the responsibility taxonomy, we observed that many responsibilities were ambiguous within the data, consistently finding low IRR despite multiple rounds of iteration and discussion. In the final round of iteration, we adapted a method described by Schaekermann et al. (2018) to conduct a more detailed disagreement discussion process for the seven responsibilities for which we found IRR to be the lowest. This process consisted of (i) an evidence-finding phase in which an annotator was asked to highlight specific textual evidence for a particular responsibility's presence in an update, followed by (ii) a reconsideration phase in which annotators who had not indicated the presence of that responsibility were asked to consider the presence of that responsibility in light of the textual evidence provided by another annotator. 25.7% of 152 updates reconsidered in this discussion process resulted in irresolvable disagreement i.e. the primary annotators continued to disagree. Furthermore, after subsequent annotation of 20 sites to compute IRR (Table 4), three of the seven responsibilities involved in the disagreement discussion process achieved lower agreement compared to scores on a prior annotation set. The high amount of irresolvable disagreement indicates high ambiguity in those responsibility's themes. We return to this point in Section 6.

Complete taxonomies For phases, we defined taxonomic categories over two rounds of iteration, finding high annotator agreement as shown in Table 3. Patterns between the annotated phases are shown in Figure 1. For responsibilities, we defined taxonomic categories over five rounds of iteration. Table 4 shows low annotator agreement for many responsibilities. Low-agreement responsibilities like Preparation may not be useful in describing patient behavior in a social media context without further qualitative elucidation of those responsibilities; as it stands, the mapping from the conceptual to the observable is too ambiguous. As we turn to classification, we drop the lowest-agreement responsibilities and focus on the six responsibilities with Cohen's $\kappa > 0.4$. This division is arbitrary, but reflects commonly used guidelines indicating $\kappa > 0.4$ as moderate agreement (Landis and Koch 1977).

Sampling:	Random		Uncertainty		Death	
	S	U	S	U	S	U
Cancer Phases	109	2791	28	278	63	3852
Responsibilities	82	1891	23	34	—	—

Table 5: Human annotation counts in terms of number of sites (S) and number of journal updates (U).

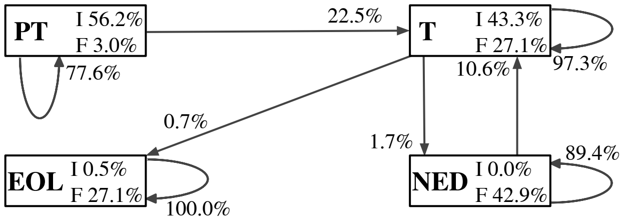


Figure 1: Phase transition probabilities based on human-annotation. Each phase indicates the percentage of sites with an initial (I) and final (F) update in this phase.

Total human annotations of sites and updates following the final iteration of both taxonomies are shown in Table 5. During random sampling, we observed only a single site that ended in the death of the patient, which ran counter to a finding from Ma et al. that 37% of cancer sites on CB do so (2017). We speculated that patient-centered narratives are less likely to provide clear indicators of patient death. To investigate end-of-life sites more carefully, we identified a high-precision filter to identify candidate sites that may contain such updates. We filtered to 63 sites using the conjunction of predictions from a death classifier developed by Ma et al. (2017), a keyword list[3], and sites that ended with a non-patient-authored update. After annotation of these sites for phases, we determined that 82.5% of them contained end-of-life updates.

5 Classification

5.1 Classification Methods

Using the complete taxonomies operationalized from qualitative work, we classify CB updates by taxonomic category. We train supervised ML models from the data annotated during the development of the taxonomies. We compare the ML classifiers to keyword classifiers that assign a category label to an update if it contains one of the words on a keyword list defined for that category.

ML classifier We formulate both phase and responsibility identification as multilabel classification problems. For

[3] Keywords used: hospice, funeral, death, passed away, obituary, wake, commemoration

phases, the prediction target is a 4x1 vector of labels corresponding to the four phases, whereas for responsibilities the prediction target is a 6x1 vector. To make use of correlations between the classes, we evaluate multilabel models rather than transforming the problem to independent binary classification problems (Read et al. 2009). We remove from consideration all updates with fewer than 50 characters of text content between the title and body text. All models were trained using Vowpal Wabbit (Langford, Li, and Strehl 2007). After evaluating several models, we achieved the best performance with cost-sensitive one-against-all (CSOAA) regression models, with human annotations converted to costs in the (0,1) range to be predicted by the regressions. We use hashed unigram, bigram, and skip bigram text features extracted from the title and body text of each journal. CSOAA is a binary logistic regression model per label with weighting applied to minimize false positives (Beygelzimer, Langford, and Zadrozny 2005). Performance could likely be improved through the use of a state-of-the-art NLP classification model e.g. (Howard and Ruder 2018) or through alternative problem formulations e.g. phases as sequences (MacLean et al. 2015); the models we present here represent a proof of concept and a reasonable ML baseline against which the keyword classifiers may be compared. We found that classification performance was not particularly sensitive to the choice of ML model among the linear classifiers we evaluated.

In the phase model, we make use of both annotator uncertainty and annotator disagreement to increase the cost of human-assigned phases by 0.2 when 'Unknown' is selected and 0.1 when two annotators disagree on an update. We also include contextual information from the two prior updates on the site, adding features from those updates' text and the number of seconds elapsed since those updates.

For training and validation, we used human-annotated journal updates obtained after the final taxonomic iteration. After training initial models, we utilized uncertainty sampling to identify additional updates for annotation (Table 5). To improve the phase model, we identified additional sites by averaging uncertainty metrics across all updates on a site.[4] We also selected sites that generated erroneous tags or erroneous transitions, e.g. an update tagged PT and EOL, or a transition from NED to PT. For the responsibility model, we sampled individual journal updates.[5]

We evaluate the performance of the two classifiers using means from fifty executions of 20-fold cross validation. To avoid leaking specific author information into the validation set, CV folds are generated at the site level, with all annotated updates from any specific site appearing in just the training or the validation set.

Keyword classifier A keyword-based classifier assigns a class label to an update if it contains one of the words on a keyword list defined for that class. While keyword lists are constructed in many ways, we invert the problem and ask: regardless of the keyword selection method, what is the best performance that can be achieved by an optimally-selected keyword list? We develop keyword classifiers to reflect two definitions of "best performance": maximum precision and maximum representativeness. Following a common requirement that keyword lists have near-perfect precision, we first identify a keyword list for each class that ensures perfect precision and the highest possible recall. While identifying the optimal keyword list is NP-hard, we represent the selection of the keyword list for each class as the maximum k-cover problem and approximate the optimal lists through a well-known greedy algorithm[6] (Feige 1998). In this formulation, for each class label c, we identify a set W_+ of words appearing only in updates assigned c and a keyword list containing k words in W_+ covering the maximum number of updates annotated with c. As keyword lists contain only words in W_+, each keyword list ensures 100% precision but unknown recall. For this evaluation, we allow the keyword lists' "words" to contain unigrams or bigrams and remove English stopwords from consideration. We evaluate the generalizability of these keyword lists via 10-fold cross validation.

We build a second set of keyword-based classifiers to represent situations where keyword lists are constructed from the words that are most "representative" of each category and for which perfect precision is not a requirement. We identify words that are most associated with each phase and responsibility using frequency-based odds ratio—a measure used in prior OHC work (MacLean et al. 2015). If $f_c(w)$ is the number of updates assigned class label c that contain word w, then $OR(w, c) = (f_c(w) \times f_{\bar{c}}(\bar{w}))/(f_c(\bar{w}) \times f_{\bar{c}}(w))$. For each class label, the keyword list contains the k non-stopword unigrams with the highest odds ratio that appear in at least 10% of updates assigned that class label. These lists contain the words that are most representative of the category relative to the other categories and may better reflect the possible output of an expert-driven keyword identification process. We evaluate the generalizability of these keyword lists by the mean performance over fifty executions of 10-fold cross validation.

5.2 Classification Results

Baseline model results To contextualize the subsequently presented ML model results, we report two baselines recommended for multilabel classification problems by Metz et al. (2012): (1) *Subset-Accuracy* ($\mathrm{B_{SA}}$)—a baseline that predicts the most common multi-label in the dataset, meaning {T} for phases and {SM,CP} for responsibilities; and (2) *F-Measure* ($\mathrm{B_{FM}}$)—a baseline that predicts the set of labels that maximizes F1 score. Results are shown in the final rows of Table 6. To reflect an interest in correctly identifying the most-common classes, all mean results in this paper are

[4] We used three uncertainty metrics defined by Li and Guo (2013): entropy, distance to the decision threshold, and maximum separation margin.

[5] We used two uncertainty metrics appropriate for multilabel classification defined by Li and Guo (2013): maximum separation margin and label cardinality inconsistency.

[6] The maximum k-cover problem's greedy algorithm has an approximation ratio of $1 - 1/e \approx 0.632$ assuming $P \neq NP$, a claim on which this paper takes no position. Thus, the identified keyword lists achieve a recall that is at worst 63% of the optimal recall.

Resp.	P	R	F1	Phase	P	R	F1
CS	0.75	0.83	0.80	PT	0.91	0.95	0.93
SM	0.93	0.98	0.95	T	0.96	0.99	0.97
CP	0.90	0.97	0.93	EOL	0.55	0.96	0.70
FM	0.47	0.92	0.58	NED	0.86	0.86	0.86
GB	0.19	0.87	0.68	—	—	—	—
BC	0.32	0.41	0.34	—	—	—	—
Mean	0.89	0.96	0.92	**Mean**	0.94	0.97	0.95
B_{SA}	0.70	0.86	0.77	B_{SA}	0.74	0.86	0.79
B_{FM}	0.72	0.99	0.80	B_{FM}	0.74	0.99	0.81

Table 6: ML classifier performance in terms of Precision (P), Recall (R), and F1 score, along with two baseline measures.

computed as weighted macro averages such that per-class performance is weighted by the prevalence of that class.

Machine learning model results Table 6 presents the performance of the phase and responsibility ML classifiers. Both models significantly outperform the baselines. Performance is better for the phases than the responsibilities, reflecting the challenges described during operationalization.

We analyzed patterns in the predictions generated by the models. For phases, 7,181 updates (4.7%) are given invalid phase assignments i.e. a combination of labels representing a transition not shown in Figure 1. We find a relationship between these erroneous outputs and two primary factors: 69.2% of the invalidly-labeled updates are either less than 500 characters or the first journal on a site. Discounting invalidly-labeled updates, 3.2% of sequential updates are labeled with invalid transitions.

For the responsibility model, we compared the number of responsibilities predicted present in the update to the number of responsibilities in the ground truth for that update. While humans annotated no updates containing all six responsibilities, 4.2% of updates are predicted to contain all six responsibilities. These likely-erroneous predictions are primarily assigned to short updates: 90.4% of updates predicted to contain all six responsibilities are shorter than 500 characters. The model predicts that a higher proportion of updates (+4 percentage points on average) contain each responsibility than the proportions identified by human annotators (Table 4). 7.7% of updates are predicted to contain no responsibilities, a decrease of 1.59 percentage points compared to the human-annotated updates.

Keyword classifier results Table 7 shows the performance of the max-precision keyword classifier for two values of k. Even when $k = 10$, the selected keyword lists overfit to the human-annotated data and perform worse than the ML models on held-out sites, demonstrating that these keyword lists fail to capture the salient information of each of the classes under consideration. Table 8 shows the performance of the maximally-representative keyword classifier. Note that when $k = 100$ recall is near-perfect in every category, which triggers a corresponding drop in precision and thus F1 score. Performance is significantly better than the max-precision keyword lists, at the cost of low precision. Generalization performance is higher relative to the max-precision keyword lists. Qualitative investigation of the keyword lists generated using both approaches reveals sensible selections. The keywords for the second classifier in particular seem appropriately representative.

	k=10				k=100			
Class	Train		Test		Train		Test	
Label	R	F1	R	F1	R	F1	R	F1
CS	.19	.32	.04	.08	.87	.93	.14	.16
SM	.34	.50	.30	.46	.90	.95	.73	.81
CP	.22	.36	.20	.32	.79	.88	.58	.68
FM	.47	.64	.07	.11	.95	.97	.09	.09
GB	.39	.56	.00	.00	.99	.99	.05	.08
BC	.30	.46	.02	.02	.99	.99	.03	.03
PT	.08	.15	.01	.02	.45	.62	.03	.04
T	.13	.23	.05	.09	.49	.66	.31	.46
EOL	.39	.56	.21	.31	.99	.99	.26	.31
NED	.11	.20	.00	.01	.52	.69	.03	.04

Table 7: Max-precision keyword list performance in terms of Recall (R) and F1 score. Precision is 1 on train.

Taken together, these results provide evidence that predictive models based on operationalizations from qualitative themes perform better using machine-learning-based approaches rather than keyword lists. However, when generalization to unseen data is not a concern, high-precision lists containing relatively few words can be constructed to achieve high recall, although inconsistent performance across categories may be challenging to identify. For exploratory modeling where precision is less important, small numbers of representative words (as may be revealed during the qualitative operationalization process) can achieve reasonable results and motivate additional data exploration. The use of keyword-based methods may also be seen as a trade-off between interpretability and robustness; the specifics of the modeling application and the need to communicate the prediction process—e.g. to designers—might motivate a preference for keywords over machine learning models. Keywords may also be appropriate when stronger assumptions about the text in a particular domain can be made (O'Connor, Bamman, and Smith 2011).

6 Model Analysis

6.1 Model Validation

To explore the expert validity of the phase and responsibility models, we invited an expert involved in the creation of the qualitative frameworks used in this paper (an author of the CJF without any affiliation or conflict of interest with this paper) to provide feedback on our operationalization. Across the elements of each taxonomy codebook, the expert rated the reasonableness of each definition on a 5-point Likert scale from strongly disagree (-2) to strongly agree (+2). Overall agreement was high for items in both the responsibility (M=1.74) and phase (M=1.83) codebooks. In the expert's qualitative feedback, several comments related to a

	k=10						k=100	
Class	Train			Test			Train	Test
Label	P	R	F1	P	R	F1	F1	F1
CS	.24	.88	.37	.23	.86	.36	.26	.26
SM	.86	.98	.92	.86	.98	.92	.93	.93
CP	.77	.99	.87	.77	.99	.87	.87	.87
FM	.22	.87	.35	.20	.77	.30	.06	.07
GB	.16	.65	.25	.12	.50	.19	.08	.08
BC	.14	.69	.23	.08	.42	.13	.08	.08
PT	.12	.72	.21	.12	.71	.20	.13	.14
T	.88	.92	.89	.88	.90	.88	.92	.92
EOL	.10	.73	.18	.11	.72	.18	.03	.03
NED	.06	.97	.12	.07	.97	.13	.11	.12

Table 8: Maximally-representative keyword list performance in terms of Precision (P), Recall (R) and F1 score.

divergence between what is observed in patient interviews and what patients self-report on CaringBridge, a gap that we leave for future work to better understand the motivations of patient sharing in OHCs.

How do we account for poor inter-annotator agreement for responsibilities despite high agreement for phases? The same annotators were involved in both models, which suggests that coder quality is not a primary cause. While annotator domain expertise may be a factor, phase indicators generally require more medical knowledge to identify than responsibility indicators. The assessment of the expert that the operationalizations are reasonable suggests there is no fundamental weakness in the iterative operationalization process used or the resulting taxonomy. Instead, we hypothesize that ambiguity in the identification and mapping of indicators to responsibilities is a critical factor. To probe the role of ambiguity in producing low IRR for the responsibilities, we conducted a qualitative analysis of the primary annotators' comments during the Schaekermann et al.-motivated discussion of disagreements (2018).

Looking at the annotator justification in cases of irresolvable disagreement reveals two preliminary themes: (1) disagreement about the directness of supporting evidence needed to assign a responsibility and (2) disagreement about which responsibility a piece of evidence indicates. These themes align with two significant dimensions of ambiguity identified by Chen et al. (2018): (a) data ambiguity, meaning multiple reasonable interpretations, often due to missing or unclear context, and (b) human subjectivity, meaning distinct interpretations resulting from "different levels of understanding or sets of experiences" among annotators. Chen et al. further utilize disagreement between coders as a proxy for ambiguity, with the lower IRR scores relative to the phases indicating a higher degree of ambiguity. The irresolvable cases suggest that data ambiguity is excacerbated by soft boundaries between responsibilities in the codebook, but the supportive external validation of the current codebook and consistently low IRR after five codebook iterations suggest an inherent ambiguity to the classification task. To reduce ambiguity, we view the next reasonable step as conducting additional qualitative work to elucidate the CJF in CB updates specifically (as opposed to additional qualitative inquiry outside the OHC context).

	Contains r?	Baseline rate of r	G^2 (df=30287)
CS	1.48 ± 0.06	1.031 ± 0.001	22122.01
SM	1.21 ± 0.03	1.011 ± 0.001	4460.91
CP	1.26 ± 0.03	1.011 ± 0.001	7171.64
FM	2.16 ± 0.50	1.053 ± 0.006	11078.42
GB	1.85 ± 0.22	1.043 ± 0.003	15279.27
BC	1.88 ± 0.24	1.047 ± 0.003	14357.79
Mean	1.64	1.033	—

Table 9: Within-week responsibility co-occurrence Poisson regression models. Incidence rate ratios with 95% confidence bounds and deviance (G^2) are given, demonstrating a greater proportion of site updates contain a responsibility r if another update published in the same week contains r. All model coefficients are significant at $p < 0.001$.

To further investigate the validity of the responsibility model specifically, we tested the expectation that an author mentioning a responsibility in an update is more likely to mention that responsibility in other updates authored in the same week, as most responsibilities in the CJF are more than momentary (Jacobs, Clawson, and Mynatt 2016). For each responsibility r, we fit a Poisson regression to predict the number of updates on site s in week w that contain r based on whether a randomly selected journal from s, w contains r. We consider only weeks with at least 2 updates and use the total number of updates authored that week on s as the exposure, additionally controlling for the baseline rate of updates on s predicted to contain r. Incidence rate ratios are shown in Table 9. When an update on a site is predicted to contain a responsibility r, other site updates in that week are predicted to contain r at a rate 1.64 times greater than if the update is predicted not to contain r. This confirms the hypothesized co-occurrence of responsibilities and provides additional evidence that the responsibility predictions are valid.

6.2 Model Integration

By classifying both phases and responsibilities for unannotated updates, we can explore temporal trends and integrate predictions to explore the relationship *between* phases and responsibilities. To mitigate noise introduced by the 4.7% of invalid predicted phases, we reassign the phase prediction of updates surrounded by single-phase updates to match the phase of its neighbors. After reassignment, 2.6% of adjacent updates predict a transition considered invalid in our phase model. Using these reassigned phase predictions, we consider responsibility predictions that co-occur with valid phase predictions to establish baseline responsibility occurrence proportions and the per-phase deviations from that baseline.

Phase model predictions over time Figure 2 traces proportions of the phases over time. Few sites have updates in the PT phase past the first 2 months. NED updates are more frequent over time, with a temporal variance that reflects

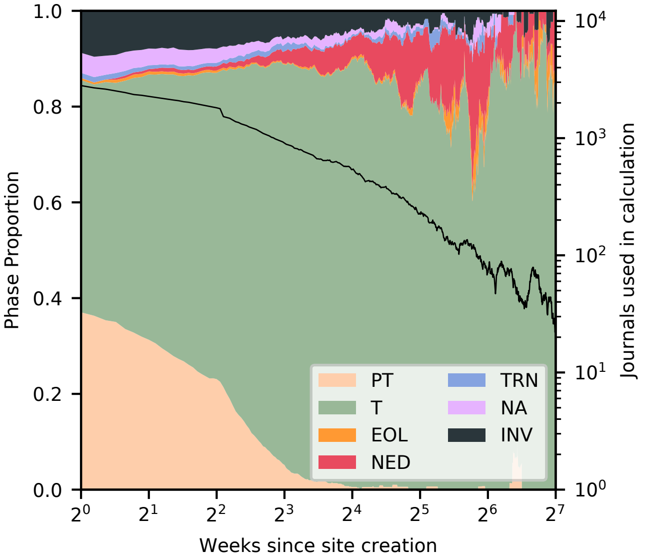


Figure 2: Predicted phases over time on a log scale. Updates across sites are binned to the day and proportions are computed based on a rolling average of 30 days (Paul, White, and Horvitz 2015). The right axis and plotted line indicate the number of updates used to compute the proportions; note fewer than 100 updates are available after 40 weeks. Proportions include updates assigned a single phase in addition to updates with multiple phases (TRN), no assigned phases (NA), or an invalid combination of multiple phases (INV).

our qualitative observation that NED updates are frequently written on consistent anniversaries (e.g. a spike around one year after initial diagnosis). Few patients continue posting updates in the EOL phase. The vast majority of updates on CB are written during the treatment phase.

Integrated model predictions Figure 3 shows the frequency of each responsibility relative to its occurrence in other phases. In contrast to the CJF's categorization of responsibilities into phases (Table 1), we find Coordinating Support, Sharing Medical Info, and Compliance appear *less* in treatment updates than in other phases, and Giving Back and Health Behavior Changes appear less in NED updates than in other phases.

7 Discussion

Bridging qualitative frameworks describing cancer patients to a user model of OHC behavior is an important step towards designing personalized digital services for cancer patients. Practically, we intend to use these models in the design of recommender systems to connect patients based on commonalities in cancer phase and expressed responsibilities, in support of an informed social network with knowledge about the cancer experience (Skeels 2010).

We experienced challenges operationalizing taxonomies from the qualitative frameworks we selected, finding phases easier to operationalize than responsibilities. To generalize this method beyond health-related qualitative frameworks, further study is needed; to facilitate application in other contexts, we discuss three aspects of the qualitative frameworks we selected that made operationalization challenging.

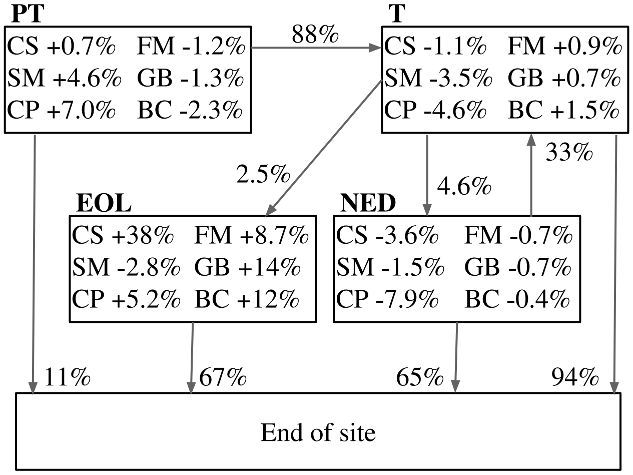


Figure 3: Predicted site phases and responsibilities. Responsibilities are the percentage point change in proportion relative to updates made in other phases. Phase transitions are labeled with the percentage of updates that follow that phase; invalid transitions e.g. NED→PT are not shown but are included in per-phase transition totals.

The first aspect is the type of mapping between observable data and conceptual theme. Taxonomic categories will be easier to define if the corresponding indicators in the data form a one-to-one map with the qualitative themes. As the number of indicators that refer to a single theme grows—the Preparation responsibility had many possible referents—the category will be increasingly hard to define and identify reliably. In contrast, the T phase had a limited set of medical indicators that could be reliably identified during operationalization. The second aspect is the temporal scale of behavioral themes. If themes describe behaviors that span lengths of time shorter than the update frequency of the available social media data, a windowed trace of user behavior makes reliable retrieval difficult. Cancer phase changes slowly and could be tracked across multiple updates, but frequent responsibilities were often alluded to without necessary context. The third aspect is the degree to which the qualitative themes are mutually exclusive. Despite periods of transition, cancer phases are largely singular and conflicting indicators within a single update rarely co-occur; responsibilities have no natural exclusivity and a single update may contain many indicators each mapping to many responsibilities.

A risk intrinsic to the bridging process we describe in this work is a perpetuation of the underlying qualitative framework's implicit lens. Our models reproduce the subjectivities of the CJF's source interviews even while mapping to a broader context of social media users. Thus, we risk magnifying or distorting aspects of the patient experience. We suggest that bridging can serve as a compliment to other methods, enabling researchers to triangulate their understandings through the inclusion of user behavior models informed by qualitative themes.

8 Conclusion

In this paper, we explored a process for bridging qualitative themes to social media user models. We built two models using taxonomic categories operationalized from two qualitative frameworks to classify unstructured text data and trace behavior over time in an OHC. We identified two primary challenges in the operationalization process along with strategies for managing them. We found that supervised ML outperforms common keyword-based approaches in classification performance. In our study of CB users, the model outputs describe the longitudinal behavior of cancer patients, which may enable the delivery of personalized digital health services. Future work includes developing more sophisticated methods for resolving challenges and understanding ambiguity in the operationalization process in order to broaden the potential scope of qualitative themes and social media contexts in which bridging can be applied.

Acknowledgements

We would like to thank the Minnesota Supercomputing Institute (MSI) at the University of Minnesota, our partners at CaringBridge, our colleagues in the GroupLens Research Laboratory, particularly Jasmine Jones and C. Estelle Smith, and the anonymous reviewers. This work was partially supported by NSF Grant No. 1464376.

References

Adcock, R., and Collier, D. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *The American Political Science Review* 95(3):529–546.

Beygelzimer, A.; Langford, J.; and Zadrozny, B. 2005. Weighted one-against-all. In *Proc. of AAAI'05*, 720–725.

Birnbaum, M. L.; Ernala, S. K.; Rizvi, A. F.; De Choudhury, M.; and Kane, J. M. 2017. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *J Med Internet Res* 19(8).

Bruckman, A. S.; Fiesler, C.; Hancock, J.; and Munteanu, C. 2017. CSCW Research Ethics Town Hall: Working Towards Community Norms. CSCW '17 Companion, 113–115.

Buis, L. R. 2008. Emotional and Informational Support Messages in an Online Hospice Support Community. *CIN: Computers, Informatics, Nursing* 26(6):358–367.

Chancellor, S.; Mitra, T.; and De Choudhury, M. 2016. Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media. In *Proc. of CHI'16*, CHI '16, 2111–2123. ACM.

Chen, N.-C.; Drouhard, M.; Kocielnik, R.; Suh, J.; and Aragon, C. R. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans. Interact. Intell. Syst.* 8(2):9:1–9:20.

Concannon, S. J.; Balaam, M.; Simpson, E.; and Comber, R. 2018. Applying Computational Analysis to Textual Data from the Wild: A Feminist Perspective. In *Proc. of CHI'18*, CHI '18, 226:1–226:13. New York, NY, USA: ACM.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting Depression via Social Media. *ICWSM* '13.

De Choudhury, M.; Sharma, S.; and Kiciman, E. 2016. Characterizing Dietary Choices, Nutrition, and Language in Food Deserts via Social Media. In *Proc. of CSCW '16*, 1155–1168. San Francisco, California, USA: ACM Press.

Eschler, J.; Dehlawi, Z.; and Pratt, W. 2015. Self-Characterized Illness Phase and Information Needs of Participants in an Online Cancer Forum. In *ICWSM*.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding Topic Signals in Large-Scale Text. 4647–4657. ACM Press.

Feige, U. 1998. A Threshold of Ln N for Approximating Set Cover. *J. ACM* 45(4):634–652.

Ferguson, R. D.; Massimi, M.; Crist, E. A.; and Moffatt, K. A. 2014. Craving, creating, and constructing comfort: insights and opportunities for technology in hospice. 1479–1490. ACM Press.

Figueiredo, M. C.; Caldeira, C.; Reynolds, T. L.; Victory, S.; Zheng, K.; and Chen, Y. 2017. Self-Tracking for Fertility Care: Collaborative Support for a Highly Personalized Problem. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW).

Geiger, R. S., and Halfaker, A. 2017. Operationalizing Conflict and Cooperation Between Automated Software Agents in Wikipedia: A Replication and Expansion of 'Even Good Bots Fight'. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW).

Gui, X.; Chen, Y.; Kou, Y.; Pine, K.; and Chen, Y. 2017. Investigating Support Seeking from Peers for Pregnancy in Online Health Communities. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW):50:1–50:19.

Hayes, G. R.; Abowd, G. D.; Davis, J. S.; Blount, M. L.; Ebling, M.; and Mynatt, E. D. 2008. Opportunities for Pervasive Computing in Chronic Cancer Care. In *Pervasive Computing*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 262–279.

Heyland, D. K.; Dodek, P.; Rocker, G.; Groll, D.; Gafni, A.; Pichora, D.; Shortt, S.; Tranmer, J.; Lazar, N.; Kutsogiannis, J.; and Lam, M. 2006. What matters most in end-of-life care: perceptions of seriously ill patients and their family members. *CMAJ* 174(5):627–633.

Hornbk, K.; Sander, S. S.; Bargas-Avila, J.; and Simonsen, J. G. 2014. Is Once Enough? On the Extent and Content of Replications in Human-Computer Interaction.

Howard, J., and Ruder, S. 2018. Universal Language Model Fine-tuning for Text Classification. *arXiv:1801.06146 [cs].*

Huang, T.; Elghafari, A.; Relia, K.; and Chunara, R. 2017. High-resolution Temporal Representations of Alcohol and Tobacco Behaviors from Social Media Data. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW):54:1–54:26.

Jacobs, M.; Clawson, J.; and Mynatt, E. D. 2014. Cancer navigation: opportunities and challenges for facilitating the breast cancer journey. 1467–1478. ACM.

Jacobs, M.; Clawson, J.; and Mynatt, E. D. 2016. A Cancer Journey Framework: Guiding the Design of Holistic Health Technology. In *Proc. of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '16, 114–121.

Kiciman, E.; Counts, S.; and Gasser, M. 2018. Using Longitudinal Social Media Analysis to Understand the Effects of Early College Alcohol Use. *ICWSM*.

Kuksenok, K.; Brooks, M.; Robinson, J. J.; Perry, D.; Torkildson, M. K.; and Aragon, C. 2012. Automating large-scale annotation for analysis of social media content. In *Proc. of TextVis '12*.

Kulkarni, V.; Kern, M. L.; Stillwell, D.; Kosinski, M.; Matz, S.; Ungar, L.; Skiena, S.; and Schwartz, H. A. 2018. Latent human traits in the language of social media: An open-vocabulary approach. *PLOS ONE* 13(11):e0201703.

Landis, J. R., and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159–174.

Langford, J.; Li, L.; and Strehl, A. 2007. Vowpal Wabbit online learning project.

Li, X., and Guo, Y. 2013. Active Learning with Multi-Label SVM Classification. In *Proc. of IJCAI*, 7.

Liu, J.; Weitzman, E. R.; and Chunara, R. 2017. Assessing Behavior Stage Progression From Social Media Data. In *Proc. of CSCW '17*, CSCW '17, 1320–1333. ACM.

Ma, H.; Smith, C. E.; He, L.; Narayanan, S.; Giaquinto, R. A.; Evans, R.; Hanson, L.; and Yarosh, S. 2017. Write for Life: Persisting in Online Health Communities Through Expressive Writing and Social Support. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW):73:1–73:24.

MacLean, D.; Gupta, S.; Lembke, A.; Manning, C.; and Heer, J. 2015. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In *Proc. of CSCW '15*, CSCW '15, 1511–1526. New York, NY, USA: ACM.

Massimi, M.; Dimond, J. P.; and Le Dantec, C. A. 2012. Finding a New Normal: The Role of Technology in Life Disruptions. In *Proc. of CSCW '12*, CSCW '12, 719–728.

Metz, J.; de Abreu, L. F. D.; Cherman, E. A.; and Monard, M. C. 2012. On the Estimation of Predictive Evaluation Measure Baselines for Multi-label Learning. In *IBERAMIA 2012*, Lecture Notes in Computer Science, 189–198.

Morgan, D. L. 1998. Practical Strategies for Combining Qualitative and Quantitative Methods: Applications to Health Research. *Qual Health Res* 8(3):362–376.

Muller, M.; Guha, S.; Baumer, E. P.; Mimno, D.; and Shami, N. S. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *Proc. of GROUP '16*, GROUP '16, 3–8. ACM.

Newman, M. W.; Lauterbach, D.; Munson, S. A.; Resnick, P.; and Morris, M. E. 2011. It's Not That I Don't Have Problems, I'm Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health. In *Proc. of CSCW*, CSCW '11, 341–350. ACM.

O'Brien, M.; Stricker, C. T.; Foster, J. D.; Ness, K.; Arlen, A. G.; and Schwartz, R. N. 2014. Navigating the Seasons of Survivorship in Community Oncology. *Clinical Journal of Oncology Nursing* 18(s1):9–14.

O'Connor, B.; Bamman, D.; and Smith, N. A. 2011. Computational text analysis for social science: Model assumptions and complexity. In *Proc. of the NeurIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Olteanu, A.; Varol, O.; and Kiciman, E. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. In *Proc. of CSCW '17*, 370–386.

Paul, M. J.; White, R. W.; and Horvitz, E. 2015. Diagnoses, Decisions, and Outcomes: Web Search as Decision Support for Cancer. In *Proc. of WWW '15*, WWW'15, 831–841.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language. use: our words, our selves. *Annu Rev Psychol* 54:547–577.

Prendergast, T. J., and Puntillo, K. A. 2002. Withdrawal of Life Support: Intensive Caring at the End of Life. *JAMA* 288(21):2732–2740.

Prochaska, J. O., and Velicer, W. F. 1997. The transtheoretical model of health behavior change. *American journal of health promotion* 12(1):38–48.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2009. Classifier Chains for Multi-label Classification. In *Machine Learning and Knowledge Discovery in Databases*, 254–269.

Sachdeva, N.; Kumaraguru, P.; and De Choudhury, M. 2016. Social Media for Safety: Characterizing Online Interactions Between Citizens and Police.

Salminen, J.; Almerekhi, H.; Milenkovi, M.; Jung, S.-g.; An, J.; Kwak, H.; and Jansen, B. J. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *ICWSM*.

Schaekermann, M.; Goh, J.; Larson, K.; and Law, E. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 2(CSCW):154:1–154:19.

Shah, S. K., and Corley, K. G. 2006. Building Better Theory by Bridging the Quantitative-Qualitative Divide. *Journal of Management Studies* 43(8):1821–1835.

Singer, P.; Lemmerich, F.; West, R.; Zia, L.; Wulczyn, E.; Strohmaier, M.; and Leskovec, J. 2017. Why We Read Wikipedia. In *Proc. of WWW '17*, 1591–1600. ACM.

Skeels, M. M.; Unruh, K. T.; Powell, C.; and Pratt, W. 2010. Catalyzing Social Support for Breast Cancer Patients. In *Proc. of CHI '10*, CHI '10, 173–182.

Skeels, M. M. 2010. *Sharing by Design: Understanding and Supporting Personal Health Information Sharing and Collaboration within Social Networks*. ProQuest LLC.

Star, S. L., and Strauss, A. 1999. Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work (CSCW)* 8(1):9–30.

Tamersoy, A.; De Choudhury, M.; and Chau, D. H. 2015. Characterizing Smoking and Drinking Abstinence from Social Media. In *Proc. of HT '15*, HT '15, 139–148. ACM.

Wen, M., and Rose, C. P. 2012. Understanding Participant Behavior Trajectories in Online Health Support Groups Using Automatic Extraction Methods. In *Proc. of GROUP '12*, GROUP '12, 179–188. New York, NY, USA: ACM.

Zhang, A. X., and Counts, S. 2015. Modeling Ideology and Predicting Policy Change with Social Media: Case of Same-Sex Marriage. In *Proc. of CHI '15*, 2603–2612.

Zhang, S.; Grave, E.; Sklar, E.; and Elhadad, N. 2017. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *Journal of Biomedical Informatics* 69:1–9.

Zhang, A. X.; Robbins, M.; Bice, E.; Hawke, S.; Karger, D.; Mina, A. X.; Ranganathan, A.; Metz, S. E.; Appling, S.; Sehat, C. M.; Gilmore, N.; Adams, N. B.; Vincent, E.; and Lee, J. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion of WWW '18*, 603–612. Lyon, France: ACM.

Zhang, A. X.; Culbertson, B.; and Paritosh, P. 2017. Characterizing Online Discussion Using Coarse Discourse Sequences. *ICWSM* 10.