Peer Recommendation Interventions for Health-related Social Support: a Feasibility Assessment

ZACHARY LEVONIAN, GroupLens, University of Minnesota, US
MATTHEW ZENT, GroupLens, University of Minnesota, US
NGAN NGUYEN, GroupLens, University of Minnesota, US
MATTHEW MCNAMARA, GroupLens, University of Minnesota, US
LOREN TERVEEN, GroupLens, University of Minnesota, US
SVETLANA YAROSH, GroupLens, University of Minnesota, US

Online health communities (OHCs) offer the promise of connecting with supportive peers. Forming these connections first requires finding relevant peers—a process that can be time-consuming. Peer recommendation systems are a computational approach to make finding peers easier during a health journey. By encouraging OHC users to alter their online social networks, peer recommendations could increase available support. But these benefits are hypothetical and based on mixed, observational evidence. To experimentally evaluate the effect of peer recommendations, we conceptualize these systems as health interventions designed to increase specific beneficial connection behaviors. In this paper, we designed a peer recommendation intervention to increase two behaviors: reading about peer experiences and interacting with peers. We conducted an initial feasibility assessment of this intervention by conducting a 12-week field study in which 79 users of CaringBridge.org received weekly peer recommendations via email. Our results support the usefulness and demand for peer recommendation and suggest benefits to evaluating larger peer recommendation interventions. Our contributions include practical guidance on the development and evaluation of peer recommendation interventions for OHCs.

CCS Concepts: • Human-centered computing → Empirical studies in collaborative and social computing; Empirical studies in HCI; User studies.

Additional Key Words and Phrases: recommendation, online health communities, social support, peer support

ACM Reference Format:
Zachary Levonian, Matthew Zent, Ngan Nguyen, Matthew McNamara, Loren Terveen, and Svetlana Yarosh. 2025. Peer Recommendation Interventions for Health-related Social Support: a Feasibility Assessment. Proc. ACM Hum.-Comput. Interact. 9, 2, Article CSCW146 (April 2025), 59 pages. https://doi.org/10.1145/3711044

1 Introduction
Social support helps people in health crises cope with stressful circumstances. Access to emotional, informational, and instrumental support is associated with increased quality of life [6], improved psychosocial health [136], and physical health [54]. Support from peers—people who have had similar health experiences—is particularly useful [1, 116]. However, a person’s existing offline support community may lack peers [131, 133]. Online communities provide a place for people to
support each other in ways that their existing offline support communities cannot by offering opportunities to connect with peers. While health support is exchanged online on diverse platforms [14, 19, 115], online health communities (OHCs) are specifically designed for health-related discussion and support [1, 152].

Even in OHCs, however, finding supportive peers can be time-consuming [31, 39]. Algorithmic matching systems could enable design features that help OHC users to find peers, but existing approaches are limited or have remained entirely theoretical [49, 75]. For example, a common approach to peer finding requires a user to explicitly search or filter for people or topics of interest, a process that support seekers and providers may find labor-intensive and discouraging in health-related contexts [39, 103]. Recommendation systems for peer finding can incorporate the user’s past behaviors to implicitly identify potential matches—people with valuable similar experiences [149]. However, while recommendation systems are commonly seen in real-world OHCs, no rigorous experimental evidence exists linking peer recommendations with increases in beneficial peer connection behaviors. Given the hypothesized benefits, why so little evidence?

The goal of this paper is to identify and address the conceptual and practical barriers to researching the impact of peer recommendations on connection behavior. We accomplish this by designing, developing, and evaluating an email-based peer recommendation system for users of the online community CaringBridge. Despite recent research arguing for the potential utility of peer recommendation for providing social support [10, 18, 22, 31, 39, 49, 75, 76, 84, 95, 96, 102, 149], no study yet describes use of peer recommendation in practice. We argue that peer recommendation should be conceptualized as a health intervention that can change specific user behaviors—behaviors that are linked to psychosocial or physical health. Inspired by public health research approaches, we conduct this initial study as a feasibility assessment to identify barriers to the implementation of a recommendation intervention for increasing social supportiveness and to determine how the intervention should be evaluated in a larger, more comprehensive study [15]. In the rest of this introduction, we summarize the proposed intervention, its evaluation in a 12-week field study, and the encouraging findings for our core contribution: evidence for the feasibility of using recommendations to connect peers in OHCs.

1.1 What is the proposed intervention?

The proposed intervention is displaying recommended peers to current OHC users. This design intervention has two aspects: an interface that displays details about recommendations and an algorithm that selects peers to recommend in the interface. The interface shows “profiles”—summaries of each individual peer being recommended [50]. By including previews of and links to a user’s recent activity within the OHC, we enable prospective peers to evaluate the relevance of the recommended user [2]. The algorithm identifies “relevant” peers for a specific user by incorporating their prior activity in the OHC and ranking potential peers based on their mutual activity.

We ground our feasibility assessment in the context of a large existing OHC—CaringBridge.org. CaringBridge users write blogs to describe their health experiences to their broader support networks. In the CaringBridge context, we will be recommending blogs written by peers to the authors of existing blogs. A recommendation intervention offers the potential for CaringBridge blog authors to form connections with peers that they may not be explicitly seeking but for whom they can give or receive meaningful support. We use a weekly email as the interface to display recommendations, copying an email alert design used by CaringBridge authors. We use a machine learning-based recommendation system to rank potential peers based on historical interaction behavior on CaringBridge.

As an intervention, recommending the blogs of peers to OHC users aims to directly manipulate those users’ natural social networks. Such manipulations are necessarily complex [126], so a focus
on specific user behaviors that the manipulation will induce is important. The intervention is designed to increase two behaviors: reading about the experiences of peers and interacting with peers. Both of these behaviors are associated with benefits, such as reduced stress, useful coping information, and a sense of community [104]. However, the experimental evidence linking peer connection with benefits is mixed [1], and includes risks such as increased distress [104]. These mixed outcomes motivate us to carefully evaluate the feasibility of an intervention to increase these behaviors and produce benefits for participants.

1.2 How did we assess feasibility of the intervention?

Feasibility refers to the ability to use an intervention to achieve desired behavior changes in reality. Content recommendations are in use on many social networking sites—including some OHCs1—with the goal of increasing user engagement [35, 62, 137, 155]. Most evidence for the causal effect of content recommenders is not public [62], and we’re aware of no evidence for the causal effect of recommenders on OHC users specifically. However online health communication is a sensitive context and people use OHCs differently than they use other social networking sites [95], so we need to be particularly careful about potential negative outcomes from the deployment of recommenders. Potential harms—including increased user distress—are a serious risk not only for users who receive recommendations but also the users recommended by these systems i.e. negative second-order effects. So the feasibility assessment we conduct is focused both on designing for the specifics of the OHC context and on evaluating the possibility of harms. One could proceed directly to a randomized controlled trial (RCT) focused directly on a construct like perceived social support rather than on-platform behavior, but without a reasonable power analysis or other minimal information about the acceptability or efficacy of peer recommendations the success of such a trial is unlikely. The feasibility assessment we conduct is designed to provide the evidence needed to run a successful RCT focused on increasing social support via peer recommendations. As feasibility is multifaceted, we adapt Bowen et al.’s useful framework for feasibility studies [15], collecting evidence of feasibility in five focus areas summarized in Table 1: Demand, Implementation, Practicality, Acceptability, and Efficacy. We designed a system and ran a field study to get converging lines of evidence for each of the feasibility focus areas.

The primary contribution of this paper is our determination that peer recommendation interventions designed to facilitate OHC user connections are ethically and practically feasible to evaluate in a larger RCT. Figure 1 outlines the paper and the primary questions we address by analyzing participants’ survey responses and behaviors on CaringBridge. During the 12-week field study, 79 participants received weekly peer recommendations via email, leading to hundreds of

---

1Online health communities include both dedicated sites and sub-communities on larger social networking sites e.g. Reddit and Facebook [152]. While all sub-communities on larger sites will include post or possibly follow recommendations, these recommenders are not specifically designed or evaluated for increasing social support.
Table 1. Feasibility assessment of a peer recommendation intervention. We collected evidence in five focus areas (originally described by Bowen et al. [15]).

| Feasibility Area | Description                                      | Evidence (Section)                               |
|------------------|--------------------------------------------------|--------------------------------------------------|
| Demand           | Interest in the intervention                     | Expressed interest (5.1.2), actual use (5.2.1)   |
| Implementation   | Tangible design and engineering to implement the intervention in a particular context | The system design: both interface (3.2) and model (3.3) |
| Practicality     | Requirements for administering the intervention  | Data and compute requirements (App. D)           |
| Acceptability    | How participants react to the intervention       | Explicit participant preferences (5.1.3) and feedback (5.2.2) |
| Efficacy         | How much the intervention affects the desired behaviors | Reading behavior (5.3.1), interaction behavior (5.3.2), second-order effects (5.3.3) |

repeat visits to blogs and hundreds of extra peer interactions. Participants clicked 5% of recommendations, although less than half of the participants clicked any recommendation and fewer still chose to visibly interact with recommended blogs. We find no evidence of significant second-order harms or benefits, and overall find an interest in and willingness to engage with blogs written by peer strangers. These results provide tangible guidance for the future development of peer recommendation systems for OHCs. We identify design trade-offs and implications for successful future deployments, discussing each of the five feasibility focus areas and arguing that peer recommendations can facilitate connections that authors may not be explicitly seeking—and that those connections can facilitate meaningful support.

2 Related Work
To design a peer recommendation system for users of online health communities (OHCs), we drew from existing work on OHCs, on peer matching, and on algorithmic recommendation.

2.1 Online health community use
People use OHCs in the hope that they will gain useful support. Both those experiencing a health condition themselves and the caregivers of others are motivated to use OHCs to overcome isolation by offsetting deficits in existing relationships and identifying people who have had similar experiences [104]. While the effects of using OHCs are somewhat unclear, feeling socially supported is a key determinant of health linked to OHC use [1, 54]. Meta-reviews reveal consistent associations between social support and a variety of health outcomes e.g. mortality [54, 134, 135]. OHCs have diverse interfaces and affordances, including forums, listservs, blogs, chatrooms, Q&A sites, and update feeds [104]. The social support available from OHCs is also diverse, including informational support from discussions of treatments and symptoms, emotional support from sympathetic others and being a part of a community, and other forms of support including the instrumental and the spiritual [1, 119]. Support can come from many types of users [149], but one particular benefit of OHCs is that they expose you to many peers.

Peers are people who have similar experiences [116]. Peer relationships differ from professional/patient and mentor/mentee relationships in that no formal role or expectation structures the relationship [116]. Compared to a person’s existing offline support networks, peers can provide more useful support [116, 131]. A variety of theoretical models support the potential benefits of peer...
connection [1, 116, 131, 133]. We avoid adopting a specific theoretical basis for our current study, as we do not operationalize or measure social support directly, instead focusing on behaviors—reading and interaction—that are compatible with multiple theoretical models. If interaction is the primary focus of online social support research [1], we view “reading” as an important but distinct way to experience online support [46, 132]. We motivate our focus on these two behaviors by describing prior work linking these behaviors to benefits.

2.1.1 Reading about peer experiences. Reading about the experiences of peers can be beneficial even in the absence of interaction [1]. In addition to learning from the valuable information contained in peers’ writing e.g. coping strategies [151], reading peer experiences can build a sense of community [111]. Further, reading can reduce loneliness [132], contribute to feelings of normalcy and hope [53, 132], reduce uncertainty and anxiety [151], and enable collective sensemaking about one’s journey [39]. In general, reading the experiences of others can benefit readers by enabling positive and normalizing social comparisons to the experiences of others [85, 118]. But, making social comparisons is not without risk: the negative experiences of others can produce a sense of helplessness or increase distress [10, 67, 85].

2.1.2 Interacting with peers. Interacting with peers offers many potential benefits—among them are membership in a community, acquisition of new information, normalization of one’s experiences, and relief from distress [104]. Online peer interaction can take three general forms: providing support to others, receiving support from others, and forming reciprocal relationships. While receiving support from others has obvious appeal, not all support is perceived as wanted or useful, and in general there is mixed causal evidence for the benefits of online interaction-based social support [1]. A gap between received and perceived support bedevils designers of social support interventions: increased received support is only weakly correlated with perceptions of that support [105, 131]. Providing support to peers, on the other hand, may be more beneficial to the provider than the receiver [110]. Support needs differ over the course of a health journey, and providing support presents an opportunity to “give back” and enact self-efficacy [60, 61]. Reciprocal peer relationships can offer the best of both worlds, but also present significant risks in health contexts: stress can increase if online contacts are doing poorly or doing well due to social comparisons [85, 104]. The sudden drop-out of a connection, due to churn or patient death, can also increase distress [7]. Further, peers might be unintentionally unsupportive due to differences in communication style [100]. Due to the risks of interacting with peers, interventions designed to increase interaction cannot be deployed without careful evaluation of the risks and benefits—which motivates us to conduct an initial feasibility assessment of peer recommendation specifically.

2.2 Social support interventions

The archetypal peer support intervention is the support group [25]. Online support groups offer similar approaches using a different medium, although generally still designed for and managed in a clinical setting [69, 141]. Other clinical approaches bridge the gap to OHCs: Haldar et al. designed an OHC for people in the same hospital [45]. Peer support interventions can have many goals: providing emotional support, providing health information or education, developing self-efficacy (e.g. by vicarious viewing of peer behavior [111]), adjusting social norms (e.g. encouraging treatment compliance), or even facilitating social movements for patient advocacy [116]. In 2004, Cohen expressed skepticism of existing peer support interventions, arguing that peer support groups produce minimal effects and that interventions should focus instead on forming weak ties and propping up existing support networks [25]. Twenty years later, the challenges associated with designing effective peer support interventions remain [1]. As an intervention into people’s online social networks, we examine recommender systems as a mechanism for encouraging initial
interactions that can blossom into weak tie relationships. Recommendation aims to improve the quality of received peer support by matching users according to some measure of “fit”.

2.3 Health peer matching

Health peer matching has occurred largely in the context of hospital-attached programs where mentors and mentees are matched by a 3rd-party broker, usually a nurse or program manager [2, 90, 129]. Consider “woman-to-woman”, a peer support program for women with gynecologic cancer: when a new participant expresses interest in the program, the program manager selects a match “of similar diagnosis and age” from a pool of volunteer mentors [89]. In contrast, online peer recommendation is not constrained to formal mentor/mentee pairings and can draw from a much larger pool of prospective “volunteers” at the cost of the clear expectations that come with structure and a human coordinator. In this study, we aim to seriously consider non-coordinated peer matching as a health-related social support intervention.

Little explicit guidance exists for peer matching [5, 116]. Prior work identifies a variety of peer characteristics that may be important for effective peer matching: we list all of the relevant papers we’re aware of in Appendix A. A general finding from HCI research is that a shared health condition is not a requirement for peer communication; people perceive value to learning from and communicating with others based on many factors that shift throughout a health journey as needs and communication goals change [10, 31]. While prior work identifies some relevant peer characteristics, minimal comparative work exists to identify the most important characteristics [116]. Hartzler et al. are a notable exception, running scenario-based sessions in which participants explicitly evaluated five potential peer mentors based on provided health information [49]. For this feasibility assessment, we make recommendations using machine learning approaches that learn valued peer characteristics from prior peer interactions.

2.4 Algorithmic recommendation for peer matching

Few published works explicitly discuss computational recommendation systems for online health communities. Hartzler et al. matched peer mentors on the basis of shared health interests, language style, and demographics—as extracted from prior posts made in the CancerConnect OHC [49]. The other notable example is described only in a PhD thesis: Yang developed and deployed a recommendation system in the American Cancer Society’s Cancer Survivor Network (CSN) forums “to direct participants to useful and informative threads that they might be interested in” [148]. They evaluated a model based on implicit feedback from prior commenting behavior by presenting recommended threads and users within the CSN interface, reporting greater thread click-through rate compared to a baseline model recommending recently popular threads. We evaluate with both click rate and more explicit metrics to understand the influence of recommendations on reading and interaction behavior.

Outside of health, a variety of problem formulations and modeling methods have been used for the problem of recommending people. Recommendation models are generally supervised machine learning models that optimize a loss function comparing the model’s output to “ground truth” labels: explicit or implicit feedback provided by users. Xu et al. present a useful review [147]. Use of implicit behavioral feedback is based on relevance assumptions, e.g. that clicked items are relevant while non-clicked items are not relevant [64], which may not hold true in practice [82, 145]. Given historical user/user interactions, one can then optimize a pointwise loss that rewards high scores for assumed-relevant user/user pairs. Input features vary from IDs for the users—which gives the classic matrix factorization approach to collaborative filtering [109]—to side information about the context where the recommendation was generated (e.g. the time and place) or the two users’ prior content (e.g. their comments) [41, 123].
Person-to-person recommendation is typically modeled as similar to the conventional user/item recommendation problem. Facebook’s deep learning recommendation model (DLRM) represents a common approach, using embeddings for categorical features (including user IDs) and multilayer perceptrons for creating dense representations of other features, then combining all representations with additional linear layers [93]. We use a simplified form of DLRM to design a recommendation system appropriate for the peer support context.

3 System Design

3.1 Observed prior use of CaringBridge for peer connection

CaringBridge.org is an online health community and blogging platform that has been the focus of prior HCI research e.g. [75, 83, 120]. In collaboration with the CaringBridge organization, we were given access to usage data from the CaringBridge website. Registered users on CaringBridge can create blogs called sites, on which they can publish blog posts called Journal updates. An author is a user who has published at least one Journal update. Much usage of CaringBridge is based around notification emails (an example is shown in Figure 3). Visitors can follow a site to be notified via email when an author of that site publishes a new Journal update.2 The majority of visitors to a site will be a patient’s existing support network [75]. Registered visitors can leave reactions to Journal updates (a reaction labeled “Heart” is the default) to show their support [121]. Alternately, they can

Recorded actions:
a. Visit
b. Follow
c. Journal update
d. Reaction
e. Comment
f. Guestbook

Fig. 2. The CaringBridge interface. We record six logged-in user actions: visits (to any of these site pages), follows (clicks on “Follow Site” and others, see sec. 3.1), Journal updates, reactions (on Journal updates, comments, or guestbooks), comments (on Journal updates or guestbooks), and guestbooks.

[Image: Site home page, as viewed by a logged-in author of that site. The study recruitment banner is visible at the top of the page. (a) Site home page, as viewed by a logged-in author of that site. The study recruitment banner is visible at the top of the page. (b) Journal page and entry, as viewed by a logged-out visitor. (c) The six logged-in user actions, above a guestbook on the Well Wishes page.]
Fig. 3. Existing design of email notifications on CaringBridge and the Site Suggestion email interface designed for this study. Data shown is a representative fabrication.

leave text-based comments on individual Journal updates or write guestbooks which are comments that appear on a special Well Wishes page. These interface components and the associated actions we track for this study are shown in Figure 2.

We are interested here in interactions (via reactions, comments, and guestbooks) between registered CaringBridge authors. An initiation is the first interaction between an author and a site on which they are not an author. In assessing demand for a peer recommendation intervention, we consider the volume of inter-author communication on CaringBridge, studied in detail by Levonian et al. [75]. Between 2010 and 2021, 275K authors initiated with other authors, more than 32.3% of all 852K authors active during that period; this interaction occurs despite minimal existing discovery features to facilitate finding peers. Further, there is associational evidence—detailed in Appendix C—that interactions from peer authors change author behavior. Sites that receive at least one interaction from a peer author will publish on CaringBridge for a median of 3.2 additional months (with 6 additional updates) compared to sites that receive only interactions from non-author visitors. Staying longer on CaringBridge is associated with benefits to authors [83]; a peer recommendation intervention aims to induce these positive author behavior changes by causing more peer interaction.

3.2 Adapting a recommendation system to CaringBridge

For a peer recommender intervention to be successful, the system must be adapted to the specifics of the context. First, CaringBridge offers only a limited public profile view, so only the content of the sites themselves provides a meaningful view into who an author is. For this reason, we optimize recommendations at the author level, but present those recommendations as site summaries. In other words, recommendations appear to users to be recommendations for sites but are actually

3While CaringBridge offers a search feature, authors use this tool to find specific authors, not for general searches for e.g. particular health conditions (see Appendix B).
recommendations for that site’s author. Second, most CaringBridge usage is motivated by email notifications, so we chose to deliver recommendations via email as well. Author notification emails (depicted in Figure 3b) are sent to site authors when a visitor interacts on a site or a co-author publishes a Journal update. As authors are familiar with this design, we use the same layout and style for our Site Suggestion emails (depicted in Figure 3c). Site Suggestion emails contain 5 bullet-pointed site recommendations, a request for feedback, a link to a feedback survey, a link to the study FAQ, and an unsubscribe link. Each site recommendation is presented with the site’s title (usually the patient’s name), a link to the site’s Journal page, and a preview of the most recent Journal update—inspired by Hartzler et al.’s finding that the most useful feature for evaluating a peer mentor is sample text [49].

3.3 Model development
To present recommended sites to a recommendation-seeking author, we include the top 5 sites as scored by a recommendation model. We adhered to two key design requirements: First, recommendations should be personalized, focusing on the right connection rather than a popular connection. We should not recommend any one site to a lot of people as that could create a negative experience for that site. Second, recommendations should be available even for authors who have never visited or interacted with another CaringBridge site. Because support-based reading and interaction is our goal, an author’s initial "cold start" recommendations should be of similar quality to recommendations for long-time authors.

To provide recommendations that meet these two criteria, we implemented a content-based recommendation system. We use implicit feedback from user activity and content-based (text) features to train a multilayer perceptron model that predicts historical peer connections and, ultimately, new site recommendations. To support the future development of similar systems, we provide a full accounting of the specific recommendation models, features, and computational resources we used in Appendix D. The offline evaluation we conducted used a train/validation/test split containing 7 years, 6 months, and 6 months of data respectively. We used standard recommendation system metrics appropriate for this context (discussed in detail in App. D.3), achieving a hit rate (HR@5) of 19% and a mean reciprocal rank (MRR) of 0.16, which we deemed more than acceptable. Recommendations were based on authors’ recent activity on CaringBridge, their prior author interactions, and their recent Journal updates. To preserve user privacy, future recommendation systems could make use of less data; we discuss the trade-offs involved in App. D.5.

4 Methods
To evaluate our recommendation system, we conducted a field study. The field study consisted of recruiting CaringBridge authors and sending them 11 weekly Site Suggestion emails. The full analysis timeline is shown in Figure 5. System implementation and analysis code are available on GitHub.4

4https://github.com/levon003/HealthBlogRec
4.1 Recruitment Survey

We recruited active CaringBridge authors by displaying a banner ad with a link to an opt-in survey. The banner appeared only to logged-in CaringBridge users on the home page of sites on which they are an author, as shown in Figure 2a. The recruitment banner and opt-in survey were active for three weeks in August 2021. The survey asked authors to opt-in to the study and included three optional questions on peer connection: on prior use of CaringBridge for connecting with strangers, on motivations for peer connection, and on characteristics that make a peer connection appealing. All survey texts are available in Appendix E.

4.1.1 Participant matching. Figure 6 shows the recruitment pipeline. Of the 100 survey completions, 96 opted-in and met the three study consent criteria: being at least 18 years old, being a current CaringBridge author, and consenting to provide the email address associated with their CaringBridge account in order to receive Site Suggestion emails. Electronic consent was stored in Qualtrics.

We matched survey responses to a specific CaringBridge account based on their provided email. For the 8 cases where we couldn’t find an associated account, we sent a follow-up email asking for profile information, matching 2 additional profiles. Finally, we excluded 11 participants who had published less than 3 Journal updates by September 1, 2021. Ultimately, 79 participants were sent Site Suggestion emails.

4.1.2 Observed prior use. The median participant had been writing Journal updates on CaringBridge for fewer than 6 months at the time of enrollment. But, consistent with observations by Levonian et al. [75], a majority of the 79 participants both visited and interacted with at least one fellow author’s site. That suggests at least some demand for peer connection, although the field study extends beyond known contacts to peer strangers. To quantify the differences between authors who chose to enroll in the study and other CaringBridge authors, we identified 30K authors who visited their own site while the banner survey was live and thus could have seen the banner, filtering down to 1,759 who had at least 3 Journal updates and thus could have received Site Suggestion emails. Table 2 compares participants to this set of non-enrolled authors, which we term a pseudo-control group, revealing that participants are generally newer to CaringBridge than other eligible authors.

---

5The recruitment banner being visible only on the home page of an authored site means that it will not be seen by users who authored a site in the past but are not regularly visiting their site(s). However, this recruitment method also excludes active authors who (a) visit a site sub-page like the Journal page (Fig. 2b) directly e.g. via a bookmark or (b) access CaringBridge using the mobile app.

6See App. D.1 for a discussion of why we require three updates to produce recommendations.
Table 2. Pre-study CaringBridge usage by participants enrolled in the field study and by a pseudo-control group of eligible non-enrolled authors. Author tenure is the number of days between a user’s first published Journal update and Sept. 1, 2021. *Indicates a significant difference at the 99.5% threshold for a Welch’s $t$-test on the mean difference and for a Mann-Whitney $U$ test (reported as the common language effect size).

|                        | Participants ($n_P=79$) | Pseudo-Control ($n_C=1759$) | $M_P - M_C$ | $U_P/(n_Pn_C)$ |
|------------------------|--------------------------|-----------------------------|-------------|----------------|
| Author tenure (days)    | 179 709.4 (1146.6)       | 3894 4078.5 (983.1)          | -3369.1*    | 3.2%*          |
| Journal updates         | 28 98.6 (262.4)          | 77 167.1 (307.7)             | -68.5       | 31.1%*         |
| Peer site visits        | 3 6.5 (12.3)             | 10 32.0 (84.0)               | -25.5*      | 23.9%*         |
| Peer site initiations   | 1 2.5 (4.4)              | 5 11.5 (35.2)                | -9.0*       | 23.8%*         |
| Peer site interactions  | 2 39.7 (94.4)            | 30 201.2 (1384.5)            | -161.5*     | 29.0%*         |

While the field study uses an uncontrolled design, we will use this pseudo-control group to estimate the potential impact of the peer recommendation intervention on author behavior.

### 4.2 Recommendation Emails

After we sent an initial Site Suggestion email on September 2, 2021, we conducted an initial assessment of interest and determined the study could proceed; we resumed sending emails on September 17, 2021 at a weekly pace. After 11 Site Suggestion emails, a “thank you” email was sent with a final request for feedback. To evaluate the effectiveness of the Site Suggestion emails as a recommendation interface, we analyzed the rate of clicks on recommendations and the explicit feedback we received from participants.

Click estimates are based on multiple data sources, although the primary source was via UTM tags embedded in each email’s links. Appendix F presents additional details. Explicit participant feedback was collected from four sources: direct responses to the emails, responses to a feedback survey linked in each email, an unsubscription survey linked in each email, and responses to the final “thank you” email sent on December 3, 2021. The Site Suggestion email was previously described in sec. 3.2. The “thank you” email was sent in two versions. To participants who clicked on none of the recommendations, we asked for feedback on whether they had seen the emails and why they chose not to visit any of the recommended sites. To participants who clicked on at least one recommendation, we listed up to three random sites they clicked and asked for reflections.

### 4.3 Observed behavior on CaringBridge

#### 4.3.1 Reading and interaction behavior.

To analyze the two primary behavioral outcomes—reading and interaction behavior—we identified measured proxies. We used repeated visits and site Follows as proxies for interest in reading a site. While a single site visit may still indicate value for the reader, we would need either explicit feedback or some measure of dwell time on the site. Thus, we focus on instances where a participant returns to a site at least once. Follow actions are harder to interpret (see 3.1), but at a minimum indicate an interest in continuing to read updates on the followed site. We use reactions, comments, and guestbooks left on a recommended site as evidence of interaction.

We analyzed the textual content of the comments and guestbooks created by participants. Three researchers conducted a qualitative analysis of the participant/site dyads where interaction occurred. Using Grounded Theory methods, we wrote axial codes and memos based on the Journal updates and comments in which participants and authors of recommended sites interacted, identifying emergent themes through code comparison [21].
4.3.2 Second-order effects of recommendations on behavior. In addition to the reading and interaction behavior of participants, receiving Site Suggestion emails and visiting strangers’ CaringBridge sites might have second-order effects: harms or benefits that accrue to both participants and the authors of the sites they visit. While we cannot draw firm causal conclusions from our non-experimental study, we can check for potential harms by estimating the effect of recommendations on two secondary behavior outcomes: (a) publishing Journal updates—the “primary” use of CaringBridge by authors—and (b) visits to and interactions with fellow authors’ sites—a hypothesized outcome of peer interaction. We estimate the effects of two “treatments”: for authors, we estimate the participation effect of receiving 11 weeks of Site Suggestion emails by comparing participants to the pseudo-control group of eligible unenrolled authors (introduced in sec. 4.1.2); for sites, we estimate the visit effect of receiving site visits from peer strangers by comparing visited recommended sites to both non-visited recommended sites and a pseudo-control group of non-recommended sites.\(^7\)

We quantify the difference between the “treated” authors/sites and the untreated authors/sites, producing three estimates: associational, model-adjusted, and causal. The associational difference for a behavioral outcome \(Y\) is \(E[Y|T=1] - E[Y|T=0]\): the raw observed difference in the mean between treated and untreated authors/sites, where \(E[Y|T=1]\) is the mean outcome for the treated group i.e. participant authors or visited sites. The model-adjusted estimate uses linear regression (OLS) to adjust for activity variables \(A\), adding assumptions about model misspecification to compute the quantity \(E[Y|T=1,A] - E[Y|T=0,A]\). The causal estimate requires us to make untestable assumptions about the modeled relationship between the treated and untreated groups [51]—see Appendix J for a detailed discussion of these assumptions. Using potential outcomes notation, we define \(E[Y_{t=1}]\) as the mean behavioral outcome that would have been observed if all authors in the participant and pseudo-control group had been participants in the study. The true causal effect \(E[Y_{t=1}] - E[Y_{t=0}]\) is the influence on \(Y\) that would occur if Site Suggestion emails were sent to all eligible authors. We use the Bang-Robins doubly robust estimator to compute the causal effect, a modeling approach which combines inverse probability weighting and standardization [9] (see also [51], Ch. 13).\(^8\)

The author and site outcomes are measured as number of actions in the 13 weeks post-study and post-visit respectively. As the non-visited comparison sites were not visited by participants, we fabricate a visit time by sampling a random visit time from among the sites that were visited in that same batch. The activity variables \(A\) are computed based on an equivalent time window before the event of interest—13 weeks before the study for participants and 5 weeks before the site visit respectively.\(^9\) The analysis is not sensitive to the time window over which the pre-study features and post-study outcomes were measured (see Appendix J). The associational and causal estimates require specification of a model that includes all relevant confounds. For example, as we saw in Table 2, pseudo-control authors have been active on CaringBridge longer than participants. Activity variables used are shown in Appendix Table 18, although we cannot measure important confounds such as “interest in receiving Site Suggestion emails”, which is unobserved and potentially unexplained by the activity variables we do observe.

4.3.3 Sample sizes needed for a powered RCT. We use observed participant behavior during our field study to estimate the effect sizes of the peer recommendation intervention’s impact on reading and interaction behavior. The structure of our data lets us consider two potential future interventions:

\(^7\)The non-recommended pseudo-control sites are the five highest-scoring sites for each participant and each batch after removing the sites we actually recommended—additional details and comparison to recommended sites in Appendix I.

\(^8\)A common alternative approach is propensity score matching, but PSM is sensitive to misspecifications of the propensity model [66, 68]. We leave application of more sophisticated modeling approaches to future work.

\(^9\)We include only 5 weeks of activity context due to a lack of available repeat visit data before August 2021.
Table 3. Recruitment survey responses.

| Has visited a stranger’s site? | # checked | %  |
|------------------------------|-----------|----|
| Yes                          | 40        | 40.0% |
| No                           | 60        | 60.0% |
| Skipped                      |           |      |

Motivations

| Motivation                          | # checked | %  |
|-------------------------------------|-----------|----|
| Learn from others                   | 75        | 79.8% |
| Communicate with peers              | 43        | 45.7% |
| Receive experienced support         | 41        | 43.6% |
| Mentor newer authors                | 27        | 28.7% |
| Not interested, but maybe in future | 6         | 6.4%  |
| Never interested                    | 2         | 2.1%  |
| Not interested, but maybe in past   | 0         | 0.0%  |
| Something else                      | 8         | 8.5%  |

Characteristics

| Characteristic                          | # checked | %  |
|-----------------------------------------|-----------|----|
| Similar diagnosis or symptoms           | 79        | 84.0% |
| Similar treatment                       | 51        | 54.3% |
| High-quality writing                    | 48        | 51.1% |
| Same caregiver relationship             | 30        | 31.9% |
| Lives near me                           | 23        | 24.5% |
| Similar cultural background             | 13        | 13.8% |
| Something else                          | 8         | 8.5%  |

Because of the risk associated with the sensitive nature of the intervention, we limited the sample size to a total of 79 participants. To assess the effects of the intervention, we compared the usage behavior of the intervention group with a control group. We estimated the standardized effect sizes at 80% power with \( \alpha = 0.05 \) using G*Power’s one-tailed point biserial model [32]. Additional details are available in Appendix K.

4.4 Ethical Considerations

Peer support during health journeys is a sensitive research context. To mitigate risk, we manually reviewed all recommendations before they were sent (sec. D.6), monitored for negative experiences during the study, and restricted the pool of recommended sites to include only sites with the lowest privacy settings i.e. those that are indexable by search engines and visible to all visitors. The University of Minnesota Institutional Review Board determined that this research was exempt from full review. All usage data was collected in compliance with the CaringBridge terms of service and privacy policy, and data was anonymized with placeholder IDs before transfer to an encrypted data repository accessible only by the research team. Throughout this paper, quotes from “public” Journal updates and comments are representative but not literal to reduce discoverability [86].

5 Results

We recruited authors via a survey (sec. 5.1). Then, we sent 79 participants weekly emails (sec. 5.2). Finally, we observed the subsequent usage of CaringBridge by both the participants and the authors of recommended sites (sec. 5.3).

5.1 Recruitment Survey

Quantitative survey responses are summarized in Table 3.
5.1.1 Self-reported prior use. 40% of participants said they visited the CaringBridge site of an author who they did not know personally.\textsuperscript{10} This prior usage provides evidence that the visits and interactions among peers includes strangers, although most interactions are likely between people who already know each other offline \cite{75}.

5.1.2 Interest & motivations. We asked participants why they might visit a fellow author’s CaringBridge site. The top motivation was to learn from others (80%), then to communicate with peers (46%) and to receive experienced support (44%). A smaller percentage (29%) were motivated by mentoring newer authors. 8 respondents (9%) indicated they were not interested in visiting stranger’s sites, although 6 suggested they could be interested in the future. Free response motivations centered on learning from others, including finding inspiration, hearing how others handled similar issues, and learning “ideas on how to engage readers” of their own site.\textsuperscript{11}

5.1.3 Peer characteristics. We asked participants what characteristics of peer sites would make them want to read and engage with that site. The most selected characteristic was a similar diagnosis or symptoms (84%), a theme echoed in more specific free responses e.g. “neurological conditions” and a participant calling shared diagnosis the most important characteristic. A majority of respondents indicated that similar treatments (54%) and high-quality writing (51%) were important. Free responses indicated a desire for “inspirational”, “honest” writing with specific details that show “positive and negative aspects”; a “glimpse into the future”. Fewer respondents (32%) indicated that having the same caregiver relationship with the patient was important, although that checkbox was only relevant for non-patient respondents. 25% of respondents indicated that geography was important; one specified sharing the same hospital as a relevant characteristic. Few selected similar cultural background (14%): a divergence from prior work \cite{30}, although the question may have been phrased too euphemistically to elicit specific preferences about e.g. age, gender, and ethnicity.

5.2 Recommendation Emails

In total, we sent 4,190 recommendations to 79 participants, with 526 unique sites recommended. To evaluate the effectiveness of the Site Suggestion emails as a recommendation interface, we analyzed the rate of clicks on recommendations (sec. 5.2.1) and the explicit feedback we received from participants (sec. 5.2.2).

5.2.1 Click rate. Over the 11 weekly batches of Site Suggestion emails, participants clicked 220 (5.3%) of the 4,190 recommendations.\textsuperscript{12} Figure 7 breaks down observed clicks by email batch and by participant. Clicks are approximately log-normal; only 30 of 79 (38%) participants clicked any recommendations, and the most active participant, who we designate P1, was responsible for 54 (24.5%) of the recommendation clicks—and the majority of the interactions, as we will see in sec. 5.3.

5.2.2 Explicit feedback. Only 8 participants provided explicit feedback on the recommendations, across 13 responses. No participants provided recommendation feedback after batch 7, so responses reflect initial impressions. Table 4 summarizes the quantitative feedback received: only 4 of 10 responses found the recommendations interesting in general, while 46% of responses to specific

\textsuperscript{10}If we assume that participants visited CaringBridge sites with the account they used to fill the survey, we can compare self-reported use to actual use: 76% of participants had made at least one logged-in visit to a CaringBridge they do not author—considering only the 79 participants we could link to an existing CaringBridge account (see sec. 4.1.1). So, among authors who have visited one or more CaringBridge sites, 65% report visiting the site of a stranger.

\textsuperscript{11}We report and discuss a variety of participant quotes in Appendix G.

\textsuperscript{12}We would like to estimate click rate conditional on opening the email (i.e. the click-through rate) independently from the base click rate, but for logistical and ethical reasons we did not use email trackers in the email HTML so we assume every email was received and opened.
5.3 Observed behavior on CaringBridge

We describe the observed impact of the recommendation intervention on behaviors: reading (sec. 5.3.1), interaction (sec. 5.3.2), and second-order effects on other behaviors (sec. 5.3.3). Participant reading and interaction behaviors are the primary outcomes and are summarized in Table 5.

---

13As the sample is small and the textual feedback is more informative, we conducted no further statistical analysis on recommendation relevance feedback.
Table 5. Observed participant behavior in response to Site Suggestion emails, from the start of the study (September 2021) to the end of data collection (April 2022).

| Behavior           | Recommendations n | Recommendations % (of 4190 total) | Participants n | Participants % (of 79 total) | Recced sites n | Recced sites % (of 526 total) |
|--------------------|-------------------|----------------------------------|----------------|-------------------------------|----------------|--------------------------------|
| First Visits/Clicks| 220               | 5.5%                             | 30             | 38.0%                         | 158            | 30.0%                          |
| Second Visits      | 86                | 2.1%                             | 17             | 21.5%                         | 76             | 14.4%                          |
| Repeat Visits      | 589               | -                                | 17             | 21.5%                         | 76             | 14.4%                          |
| Follows            | 24                | 0.1%                             | 5              | 6.3%                          | 23             | 4.4%                           |
| Initiations        | 36                | 0.9%                             | 9              | 11.4%                         | 33             | 6.3%                           |
| Interactions       | 948               | -                                | 9              | 11.4%                         | 33             | 6.3%                           |
| Text Interactions  | 268               | -                                | 4              | 5.1%                          | 20             | 3.8%                           |
| Relationships      | 1                 | 0.0%                             | 1              | 1.3%                          | 1              | 0.2%                           |

5.3.1 Reading behavior. Follows were rare; only five participants ever followed a site. Repeat visits were more common: 86 (39.1%) of 220 clicked recommendations were visited a second time and collectively accrued 589 repeat visits during and after the study.

5.3.2 Interaction behavior. Nine participants interacted with recommended sites, which is 30% of those who visited at least one recommended site. Collectively, 33 sites received 948 additional interactions as a result of this study, although the majority of these accrue to just a few sites: median interactions with a single site was 6 (M=28.7; SD=56.9). Participants initiated with 38 non-recommended sites during the course of the study as well—likely with sites authored by people they already knew [75]. Thus, participation was associated with a 94.7% increase in total initiations during the study period. At an average of 0.37 initiations per author during the study period, participants also initiated more often than the 0.05 initiation average for non-participating active authors. For context, participant initiations comprised 0.2% of all the author initiations that occurred on CaringBridge during the study period, from among 39.5K active authors.

Only four participants interacted using comments or guestbooks, resulting in 20 participant/site dyads in which interaction occurred. Our qualitative analysis identified two areas of interest: first-contact strategies and potential norm violations. We observed a diversity of first-contact strategies, varying from formally introducing the self (“I am managing a CaringBridge site for my sister ... Your site was mentioned in an email from CaringBridge, I took interest in your story.”[^16^]), to general expressions of support (“Hope you get some rest soon”), to establishing common ground by sharing their personal health experiences (“I kind of know the road you are traveling. [personal health history]”). A larger study could investigate first-contact strategies that are particularly effective, either at encouraging a response or at providing useful support. In addition, we observed potential norm violations: situations where participant comments diverge from other visitor comments. These divergences include specific behaviors: first-time sympathetic comments on posts announcing the death of the patient, being the first and only commenter on a Journal update, bringing in potentially-unwanted religious messaging (“God bless you”), and asking explicit questions of the update author. The commenting norms perceived by ‘power user’ peers may diverge from most visitors, enabling them to provide support in situations others may not but running the risk of

[^14^]: This proportion is similar to the percentage of reaction interactions for all users (71.6%) and all participant interactions with non-recommended sites (71.2%) during the same period.

[^15^]: Three participants described the Site Suggestion emails as coming “from CaringBridge”.

[^16^]: Three participants described the Site Suggestion emails as coming “from CaringBridge”.
producing uniquely negative experiences. P1 provided the only comment on a death announcement update: "I am so very, very sorry for the passing of your loved one." We have no evidence of how these comments were received by authors of recced sites; we observed no objections as comment replies or indirectly via CaringBridge help/support lines.

The author of a recced site responded to a participant comment in only a single instance,\(^{17}\) and their subsequent interactions expanded into a relationship. While one relationship is insufficient for generalization, we sketch it here as a rich example. P1’s first contact with R1 shared personal experiences and offered support. "You don’t know me but you sound like you’re handling this all very well. .... I know what it’s like to get chemo and I know what it’s like after." Within weeks, R1 left supportive comments on P1’s recent Journal updates ("I can’t even begin to imagine how strong you are"). Support exchange continued, with evidence that off-CaringBridge communication had been established. Two months into their relationship, R1 explicitly referenced P1 in a Journal update ("I have a friend, I think she’d agree with that title. She somehow found me through Caring Bridge, and has been a constant support to me and my family since.") As a frequent commenter, P1 occasionally generated further discussion (Visitor: "I agree with [P1].") and on one occasion was thanked for their support by R1’s relative. While we cannot quantify the expected number of relationships per recommendation from this single example, we take the creation of a new relationship as an important ‘existence proof’ for the expected effects of peer recommendation. P1 is much more active than other authors, being on the 97th percentile by number of Journal updates and the 99th percentile by number of initiations and interactions. Will authors that are less active still form peer relationships given the opportunity? We estimate the sample sizes needed for a larger trial in sec. 5.3.4.

5.3.3 Second-order effects on participant and recommended site behavior. Estimates of the participation effect are shown in Figure 8.\(^ {18}\) While the associational effect of recommendation indicates that participants published more Journal updates than non-participants, we suspect this difference

\(^{17}\)The observed 5% reciprocation rate to participant text initiations diverges from the 12% reciprocation rate observed by Levonian et al. [75].

\(^{18}\)Additional outcomes and analysis for both participants and clicked sites are provided in Appendix J.
is primarily due to participants’ shorter tenure, and the effect disappears after adjusting for author tenure. We present a selection of outcomes related to visiting and interacting with others’ sites (specifically excluding any sites recommended during the study); none of these estimates suggest a positive or negative effect at the 95% significance level. We conclude that receiving recommendation emails had a small or high-variance effect on participant behavior—and no clear harms.

Estimates of the visit effect are shown in Figure 9—derived by comparing clicked sites to both non-clicked recommendations and to lower-scoring sites that were not recommended (see Appendix J). The impact of an individual visit should be small, so we expect and observe small, statistically indistinguishable differences. We observe that peer visits from clicking participants are associated with additional author interactions on their own site and on peer sites, although these effects disappear after adjustment.\footnote{A notable exception in Figure 9 is that visited sites have on average 2 additional self-interactions per week compared to the pseudo-control sites, although pseudo-control sites already had fewer self-interactions pre-study.}

Absent clear evidence of a harmful effect on associated behaviors and given the successful primary behavior manipulation, we recommend proceeding to a randomized controlled trial (RCT) for the peer recommendation intervention.

5.3.4 Sample sizes needed for a powered RCT. Figure 10 shows the sample sizes needed for an uncontrolled replication of our field study, based on the observed visit and interaction behavior. If sending a one-time recommendation email, at least 314 authors should be included in order to obtain a reliable estimate of total peer interactions with (and visits to) recommended sites. Designing a trial to estimate the effect of recommendations on relationships is more challenging, as no relationships formed due to the first batch of Site Suggestion emails and only one relationship formed during the entire study. Based on that one relationship, a recurring email study would need 478 participants to detect at least 1 relationship per 79 participants ($d=0.11$, $\alpha = 0.05$, $\beta=0.2$), although 10 times that number of participants would be needed if the proportion of participants that form relationships is closer to 1 in 1000 than 1 in 100—and tens of thousands are needed if relationship formation is uniformly probable across participant/recommendation pairs (as only 1 in...
Sample sizes needed to detect beneficial effect (at 80% power and 5% sig. threshold)

|                     | One-time email | Recurring email |
|---------------------|---------------|-----------------|
| Unique repeat visits| 75 (d=0.28)   | 81 (d=0.26)     |
| Total Interactions  | 314 (d=0.14)  | 444 (d=0.12)    |

Fig. 10. Sample sizes needed to detect effects of the magnitude we observed during the field study. Field study effect sizes are shown parenthetically as Cohen’s $d$.

4190 recommendations led to a relationship. We also estimated effect sizes relative to the pseudo-control group (see full results in Appendix K). Based on unadjusted effect size differences between those groups, a 300-participant recurring-email RCT—with 50% not receiving recommendation emails—should be sufficient to estimate the relative impact of recommendation on total peer visits and interactions. Assuming the same click rate we observed in this study (5.3%), such an RCT would also be sufficiently powered to investigate any potential negative impact of peer visits on recommended site update frequency.

6 Discussion: Feasibility of a peer recommendation intervention

In this study, we implemented a peer recommendation system in order to assess its feasibility as a behavior-change intervention. We collected evidence for feasibility in five areas (Table 1): Demand, Implementation, Practicality, Acceptability, and Efficacy. Here, we summarize our results as evidence for each area. A top-line summary of results is shown in Table 6.

6.1 Demand

CaringBridge users are motivated to read and interact with peer authors—and do so in practice. Demand refers to interest in the intervention. Prior use indicates both a large number of peer author interactions and a positive correlation between interaction with peer authors and retention (sec. 3.1). Peer interaction includes interactions with peer strangers as well: 40% of participants reported previously visiting the site of a stranger (sec. 5.1.2). Participants indicated a motivation to connect with peers, although less interest in interaction specifically; 80% reported an interest in learning from the experiences of others, compared to 46% reporting an interest in communication "with a peer who understands". These motivations suggest that participants primarily approach peer recommendations as an opportunity to receive social support by reading the experiences of others rather than as an opportunity to form reciprocal relationships or to provide support to others. Expressed demand was matched in practice with reasonable recommendation click rates—5% overall, 14% in the first batch (sec. 5.2.1).

6.2 Implementation

A text-based email interface with recommendations from an implicit feedback model was effective. Implementation refers to the tangible design and engineering required to implement the intervention. Choosing an appropriate recommendation interface is important, and the text-centric email interface we chose may be inappropriate on other platforms (sec. 3.2). Future studies should adapt the interface design to the platform, integrating recommendations into OHC interfaces as appropriate. Representing peer profiles in an interface remains an important open question for future work [50]; text previews were effective, but their focus on recency is a trade-off relative to curated profiles or previews that attempt to highlight the expected utility of the recommended site to the viewer e.g. via explanations [149]. The model underlying the interface was based on implicit

[20]While comparisons are tricky, these click rates are broadly consistent with other academic studies of email recommendations e.g. [12, 55].
Table 6. Summary of quantitative results.

| Feasibility Area | Result                                                                 | Section |
|------------------|------------------------------------------------------------------------|---------|
| Demand           | 1 in 3 authors interacted with at least one peer author.               | s3.1    |
| Demand           | Interactions from peer authors increase retention relative to interac-
|                  | tions from non-author visitors.                                       | s3.1    |
| Demand           | 40% of surveyed authors reported visiting a stranger’s site.           | s5.1.1  |
| Demand           | 80% of surveyed authors reported an interest in learning from the      | s5.1.2  |
|                  | experiences of others.                                                 |         |
| Demand           | Surveyed authors were less motivated to interact (46% interested in   | s5.1.2  |
|                  | communicating with peers, 44% interested in receiving experienced     |         |
|                  | support.)                                                              |         |
| Demand           | 14% click rate on the first batch of recommendations, 5% overall.     | s5.2.1  |
| Demand           | 62% (49) participants never clicked a recommendation.                 | s5.2.1  |
| Acceptability    | A majority of surveyed authors identified a similar diagnosis (84%),   | s5.1.3  |
|                  | treatment (54%), and engaging writing (51%) as important character-
|                  | istics for potential peer connections.                                |         |
| Acceptability    | Only 10% of participants (8) gave explicit feedback. In that limited  | s5.2.2  |
|                  | feedback, only 4 of 11 batches and 46% (26 of 56) of specific rec-
|                  | ommended sites were deemed interesting.                               |         |
| Efficacy         | 57% of participants visited at least one recommended site twice. 39%  | s5.3.1  |
|                  | of clicked sites were visited at least twice.                          |         |
| Efficacy         | 30% of participants interacted with at least one recommended site.    | s5.3.2  |
|                  | Participants made 948 interactions on recommended sites. 72% of       |         |
|                  | interactions were non-textual post reactions.                         |         |
| Efficacy         | 4 participants (5%) left textual comments, and 1 relationship formed. | s5.3.2  |
| Efficacy         | Study participation had non-significant impacts on author engage-
|                  | ment behavior 3 months post-study.                                    | s5.3.3  |
| Efficacy         | Visits from study participants had a nonsignificant impact on the    | s5.3.3  |
|                  | behavior of visited-site authors.                                     |         |
| Efficacy         | An RCT aiming to increase peer interaction via a one-time recom-
|                  | mendation intervention ought to recruit at least 314 OHC users.       | s5.3.4  |

feedback from historical peer interactions (sec. 3.3). Future peer recommender systems should carefully consider the available implicit feedback and the implications of optimizing for historical patterns when new interaction patterns are desired. For example, learning from the experiences of others was the most common participant motivation, but learning was likely not the primary motivation for authors’ historical interactions—collecting explicit feedback could supplement or replace this implicit feedback to better align the training objective and user motivations.

6.3 Practicality

Prior user interaction data are useful for training recommenders that recreate historical interactions. Practicality refers to requirements for administering the intervention in practice. During model development, none of the non-personalized approaches produced recommendations that effectively captured historical peer interaction dynamics, but such approaches may still be appropriate in other contexts that want to avoid using sensitive text-based disclosures as the
basis for recommendations (Appendix D). For example, if only activity data is available, it may be reasonable to form peer cohorts based on sign-up time and activity level [47]. We did not attempt to compare multiple models during our field study; future investigations will need to link the specific peer connection benefits sought with the type of modeling approach used and conduct appropriate online evaluations. Given our interest in interaction, we recommended only recent sites in order to make candidate scoring more practical—≈13K authors active within a week vs 1 million total authors—but a system focused on recommending historical information or completed journeys might consider other candidate-generation approaches [79].

6.4 Acceptability

Recommendations were inflexible—and often rejected by participants. Acceptability refers to how participants react to the intervention. Pre-study, the characteristics identified as most important for potential peer connections were a similar diagnosis, treatment, and engaging writing (sec. 5.1.3). These were the most frequently mentioned characteristics in recommendation feedback as well, with the addition of similar age. Only 8 participants gave explicit feedback (sec. 5.2.2). To those participants, recommendations were only modestly acceptable; 39% of specific recommendations were deemed irrelevant, frequently due to a lack of common ground between the participant and the recommended site. Future peer recommenders might consider including model features that capture the specific aspects of similarity deemed most important to users. For example, users might volunteer information about health condition, or it might be inferred from existing disclosures [78].

We received three explicit requests from users to alter the types of recommendations they were receiving, and one request from a user for the ability to filter the recommendations. Functionally, these requests speak to the coordination role played by a recommendation system. In clinical contexts, a human coordinator can incorporate this feedback to alter their peer matching approach on an individual level [129]. An algorithmic recommendation system faces a much greater challenge soliciting useful feedback [130]. By designing feedback interfaces, recommendation users could self-coordinate, at the risk of trade-offs between user learning and ease of use [37].

6.5 Efficacy

Receiving recommendations increased peers’ reading and interaction behavior—and did not reduce other usage. Efficacy refers to how much the intervention affects the desired behaviors. Given the small sample and the uncontrolled design, we can only evaluate efficacy in a limited way, although our results are promising. 39% of visited sites were visited a second time and a majority (57%) of participants who clicked at least once went on to visit at least one recommended site twice, suggesting that participants were interested in reading about ongoing health journeys (sec. 5.3.1). A smaller percentage (30%) of participants interacted with at least one recommended site, and only one reciprocal relationship formed as a result of the recommendations (sec. 5.3.2). These low percentages indicate likely barriers to forming deep connections that the current design is ineffective at overcoming. Unclear social norms around interaction with strangers may present a barrier to greater interaction, such as whether an author is open to receiving unsolicited support [75]. Future design work could explore soliciting indicators of openness to interactions from strangers and providing first-contact writing assistance for peer supporters who are not sure what to say [100, 119]. We observed no evidence of second-order engagement harms due to recommendations. These results suggest that efficacy should be evaluated in a larger, controlled trial.

All of the interactions we observed in this study—excepting the one relationship—saw our participants providing support to others. Being a peer supporter may by more attractive and accessible than being a support recipient [129]. On one hand, a recommender system that encourages a supporter role runs the risk of generating self-fulfilling prophecies and preventing exploration
of other roles [39]—leading users to deprioritise their own needs [1]—as well as contributing to exclusion of new users and reinforcement of existing cliques [91]. On the other hand, exposure to peer recommendations might lead to more exploration than what users choose without this scaffolding [127], and supporting others might serve as the endpoint for a personal transition to working in broader service of a health community [60, 76]. In general, the impacts of recommendation on community social dynamics are hard to predict [92]—and it is for that reason that future trials are necessary. In future studies, recommendation and visit dynamics should be carefully monitored for harmful social dynamics or intervention-generated inequalities [138].

7 Discussion: Designing for peer recommendations

7.1 Design implications

Give recommendations to OHC users. Despite negative reception in the limited explicit feedback we received, we observed no quantifiable harms to user engagement. Taken together with evidence of positive benefits—specifically, an increase in reading and peer interaction—our results suggest that designers should try giving recommendations to OHC users. In other words: recommendations probably won’t hurt and might help. Some users will ignore or dislike recommendations, as with the participant who found reading other sites “depressing”; others will use recommendations to discover sites they find valuable. We also see opportunities for disrupting “rich get richer” social network effects, using recommendations to promote equitable support exchange e.g. by recommending new sites that might derive more benefit from incoming support [125]. On CaringBridge, while only 1 in 40 of non-recommended initiations are on sites without any prior peer interactions, 1 in 10 of the recommendations we gave during the field study are for those “siloed” sites.

Let users shape the recommendations. A theme in participant feedback was requests for more or less of a particular “type” of recommendation. Explicit feedback enables users to shape the type of recommendations they receive, such as by changing the training data or specifying a specific model to use based on their motivation for viewing peer recommendations. Prior work demonstrates that people will explore multiple recommender options if given the opportunity, and some people will even opt for non-personalized recommenders e.g. only sites with a specific health condition [29]. Designers could add controls that highlight particular kinds of filtering, e.g. for sought characteristics like particular treatments or geographic proximity. Based on a provided motivation, implicit feedback models could rely on different training signals and assumptions. Our model assumed that interaction was the primary goal, but a model could be trained for sites that users want to read (trained on repeat visits), for sites that could benefit from additional comments (trained on first comments), or for sites on which a relationship is more likely to occur (trained on reciprocal interactions). Users could select recommendations appropriate for their motivations, either as a support seeker or a support provider.

Provide “first-contact” guidance. The email recommendation interface we used in this feasibility assessment provided minimal guidance on intended usage. We observed a variety of approaches to writing initial comments on a stranger’s site, as well as multiple potential social norm violations. Prior work in CSCW has identified providing design support for navigating communicative norms as a priority, especially when engaging with potential peer supporters [3]. In other online communities, a visitor’s first interaction is an important time to intervene, potentially by demonstrating appropriate norms [20]. When composing an initial comment, the interface might highlight comments written on previous Journal updates or provide some other form of writing guidance such as a prompt [75]. A recommendation interface can lead to the initial visit, but designing to reduce the friction of providing context-appropriate interaction will make it easier to provide meaningful support and could increase the likelihood of forming deep relationships. Due to the siloed nature
of CaringBridge sites, we observed minimal cross-site community building during this study, but on more collective OHCs the stakes for encouraging appropriate initial participation might be higher. Recommendations alone are evidently not sufficient to cultivate relationships or broader community, but future qualitative and design work could focus on the flow from recommendation to first contact to meaningful and supportive relationships.

7.2 Limitations & Future Work

Our feasibility assessment of peer recommendations has significant limitations, although we are excited by the prospect of additional research developing the systems and interventions we considered. We developed and evaluated only a single interface and a single model. Our interface followed the existing design of CaringBridge notification emails, but a more iterative design process could produce more useful representations of recommendations [49]. The deep learning model we use is not “state-of-the-art”, but represents a reasonable modern approach. Future studies should conduct more extensive offline and online model comparisons [11]—including with models that are focused on specifically facilitating reciprocal matching [97].

We looked at the potential impact of the recommendations for 12 weeks during the field study and 13 weeks post-study. We saw no influence of participants’ visits on the journaling behavior of recommended sites in the 13 weeks post-study, but our dataset prevents us from analyzing any longer-term influence on the authors of recommended sites. In particular, authors that perceive a change in their audience after visits or interactions from peers might change the content of their Journal updates. Writing updates involves self-disclosure, and self-disclosure can lead to vulnerability and potential disappointment in the absence of reciprocation [124]. For an audience of peers, authors may be less likely to disclose negative information [74], which could decrease the potential benefits of expressive writing and self-reflection [83] and, unintentionally, lead to posts that are less useful for both peer and non-peer readers. Future studies should collect explicit feedback from authors to understand authors’ perception of peer visitors.

As an intervention, the peer recommender system we evaluated was intended to target behaviors correlated with social support benefits. We proposed an RCT focused on quantifying the increase in reading and interaction as a result of exposure to peer recommendations. However, due to the gap between received and perceived support [104], an increase in these behaviors may not produce corresponding increases in perceptions of social support. Future experimental designs could include pre-post self-report measures, either for social support directly (e.g. using the social connectedness scale [73]) or for desired impact on downstream health (e.g. using the perceived stress scale [72] or measures for health-related quality of life [34]). Ultimately, the purpose of this feasibility study is to support the future execution of an RCT that moves beyond engagement metrics. An RCT for a peer recommendation intervention should evaluate not only that perceived social support increases for impacted users, but also that the increases in social support induced by peer recommendation are positively associated with more general self-reported health measures.

We designed and evaluated the peer recommender system only with users of a single OHC. We argue these results generalize most to OHCs with similar affordances to CaringBridge. Rains argues that communication technologies for social connection have four primary affordances: visibility, availability, control, and reach [104]. Control, for example, is the potential to manage interactions, while availability is the potential to interact at particular times or places, so other text-based asynchronous communities provide similar potential. Visibility is the potential to make one’s self known to others or to observe others’ behavior; on CaringBridge, visibility is linked to specific blogs and communities that form around those blogs. Forums or listservs offer very different trade-offs around visibility of user-generated content, as do OHCs with rich user profiles [39, 49]. Reach is the potential to contact specific individuals, groups, or communities: CaringBridge provides...
reach only to specific individuals known to the user by name [75]. Peer recommendation might function differently in an ecosystem where other social discovery features are already present [148]. Future work should explore peer recommendation in OHCs with diverging affordances from those provided by CaringBridge—such as Q&A communities.

We primarily collected quantitative feedback in this study, and the limited text feedback we did collect raised questions that would best be answered with future qualitative work. Qualitative interviews might identify authors’ motivations for seeking peer recommendations, authors’ perceptions of their “imagined audience” [80], and the impact of peer visits on those perceptions. In-lab usability studies—representing a midpoint between online and offline evaluations [11, 153]—might identify what makes a peer recommendation good. In particular, participants disliked some of our recommendations: in-depth qualitative elicitation methods could identify why these recommendations were ineffective and what kinds of recommendations would meet their peer connection needs.

8 Conclusion

This paper conceptualized peer recommendations as an intervention to increase reading and interaction behavior in online health communities. We implemented and evaluated a peer recommendation system, demonstrating both the feasibility of and challenges associated with running a larger experimental study. Our system meets a demand for learning from the experiences of others and demonstrates the practicality of content-based models implemented in an email interface. During a 12-week field study, peer recommendation emails successfully led participants to read about and interact with peers—with no evidence of harmful second-order effects. We hope that our specific implementation serves to promote future investigation of peer recommendation: including models, interfaces, and outcomes for both individuals and communities.

Acknowledgments

We would like to thank Daniel Kluver, Juan F. Maestre, Stevie Chancellor, Martin Michalowski, Haiwei Ma, and C. Estelle Smith for their feedback. Special thanks to our collaborators at CaringBridge, especially Dennis Still, Brigid Bonner, Patricia McMorrow, Dale Durham, and Jon Linn. The Minnesota Supercomputing Institute at the University of Minnesota provided computing support. This research was funded in part by the University of Minnesota Doctoral Dissertation Fellowship.

References

[1] Kimberley R. Allison, Pandora Patterson, Daniel Guilbert, Melissa Noke, and Olga Husson. 2021. Logging On, Reaching Out, and Getting By: A Review of Self-reported Psychosocial Impacts of Online Peer Support for People Impacted by Cancer. Proceedings of the ACM on Human-Computer Interaction, 5, CSCW1 (April 2021), 95:1–95:35. doi:10.1145/3449169

[2] Nazanin Andalibi and Madison K. Flood. 2021. Considerations in Designing Digital Peer Support for Mental Health: Interview Study Among Users of a Digital Support System (Buddy Project). JMIR Mental Health 8, 1 (2021), e21819. doi:10.2196/21819 Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

[3] Nazanin Andalibi and Andrea Forte. 2018. Responding to Sensitive Disclosures on Social Media: A Decision-Making Framework. ACM Trans. Comput.-Hum. Interact. 25, 6 (Dec. 2018), 31:1–31:29. doi:10.1145/3241044

[4] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2018. Social Support, Reciprocity, and Anonymity in Responses to Sexual Abuse Disclosures on Social Media. ACM Trans. Comput.-Hum. Interact. 25, 5 (Oct. 2018), 28:1–28:35. doi:10.1145/3234942

[5] Anna M. Anderson, Elizabeth C. Lavender, Esther DuSabe-Richards, Teumzghi F. Mebrahtu, Linda McGowan, Philip G. Conaghan, Sarah R. Kingsbury, Gerry Richardson, Deborah Antcliff, and Gretl A. McHugh. 2021. Peer mentorship to improve self-management of hip and knee osteoarthritis: a randomised feasibility trial. BMJ Open 11, 7 (July 2021), e045389. doi:10.1136/bmjopen-2020-045389 Publisher: British Medical Journal Publishing Group Section: Rheumatology.
[24] Andrea Civan, David W. McDonald, Kenton T. Unruh, and Wanda Pratt. 2009. Locating patient expertise in everyday life. In *Proceedings of the ACM 2009 international conference on Supporting group work (GROUP ’09)*. Association for Computing Machinery, New York, NY, USA, 291–300. doi:10.1145/1531674.1531718

[25] Sheldon Cohen. 2004. Social relationships and health. *American psychologist* 59, 8 (2004), 676.

[26] Rafael del Pino-Casado, Antonio Frias-Osuna, Pedro A. Palomino-Moral, María Ruzafa-Martínez, and Antonio J. Ramos-Morcillo. 2018. Social support and subjective burden in caregivers of adults and older adults: A meta-analysis. *PLOS ONE* 13, 1 (Jan. 2018), e0189874. doi:10.1371/journal.pone.0189874 Publisher: Public Library of Science.

[27] C-L Dennis, E Hodnett, L Kenton, J Weston, J Zupancic, D E Stewart, and A Kiss. 2009. Effect of peer support on prevention of postnatal depression among high risk women: multisite randomised controlled trial. *BMJ* 338, jan15 2 (Jan. 2009), a3064–a3064. doi:10.1136/bmj.a3064

[28] Jingtao Ding, Yuhan Quan, Quanming Yao, Yong Li, and Depeng Jin. 2020. Simplify and Robustify Negative Sampling for Implicit Collaborative Filtering. *Advances in Neural Information Processing Systems* 33 (2020), 1094–1105. https://proceedings.neurips.cc/paper/2020/hash/0c7119c3a6a2209da6a5b90e5b75bd-Abstract.html

[29] Michael D. Ekstrand, Daniel Klver, F. Maxwell Harper, and Joseph A. Konstan. 2015. Letting Users Choose Recommender Algorithms: An Experimental Study. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys ’15)*. Association for Computing Machinery, New York, NY, USA, 11–18. doi:10.1145/2798283.2800195

[30] Gayathri Embuldeniya, Paula Veinot, Emma Bell, Mary Bell, Joyce Nyhof-Young, Joanna E. M. Sale, and Nicky Britten. 2013. The experience and impact of chronic disease peer support interventions: A qualitative synthesis. *Patient Education and Counseling* 92, 1 (July 2013), 3–12. doi:10.1016/j.pec.2013.02.002

[31] Jordan Eschler and Wanda Pratt. 2017. “I’m So Glad I Met You”: Designing Dynamic Collaborative Support for Young Adult Cancer Survivors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW ’17)*. ACM, New York, NY, USA, 1763–1774. doi:10.1145/2998181.2998326

[32] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G’Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods 39*, 2 (May 2007), 175–191. doi:10.3758/BF03193146

[33] Casey Fiesler and Nicholas Proferes. 2018. “Participant” Perceptions of Twitter Research Ethics:. *Social Media + Society* (March 2018). doi:10.1177/2056305118763366 Publisher: SAGE PublicationsSage UK: London, England.

[34] Centers for Disease Control and Prevention. 2000. *Measuring Healthy Days*. Technical Report. CDC, Atlanta, GA. 44 pages.

[35] Jill Freyne and Michael Massagli. 2008. Social Uses of Personal Health Information Within PatientsLikeMe, an Online Patient Community: What Can Happen When Patients Have Access to One Another’s Data. *Journal of Medical Internet Research* 10, 3 (2008), e15. doi:10.2196/jmir.1053

[36] K E. Goldschmitt and Nick Seaver. 2019. Shaping the Stream: Techniques and Troubles of Algorithmic Recommendation. In *Proceedings of the ACM 12th Conference on Recommender Systems*. ACM, Vancouver British Columbia Canada, 541–542. doi:10.1145/3240323.3241622

[37] Doğa Gatos, Ashi Günday, Güncel Kirlangıç, Kemal Kuscu, and Asim Evren Yantac. 2021. How HCI Bridges Health and Design in Online Health Communities: A Systematic Review. In *Designing Interactive Systems Conference 2021 (DIS ’21)*. Association for Computing Machinery, New York, NY, USA, 970–983. doi:10.1145/3461778.3462100

[38] K. E. Goldschmitt and Nick Seaver. 2019. Shaping the Stream: Techniques and Troubles of Algorithmic Recommendation. In *The Cambridge Companion to Music in Digital Culture*, David Trippett, Monique M. Ingalls, and Nicholas Cook (Eds.). Cambridge University Press, Cambridge, 63–81. doi:10.1017/9781316676639.006

[39] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. WTF: the who to follow service at Twitter. In *Proceedings of the 22nd international conference on World Wide Web (WWW ’13)*. Association for Computing Machinery, New York, NY, USA, 505–514. doi:10.1145/2488388.2488433

[40] Ido Guy, Inbal Ronen, and Eric Wilcox. 2009. Do You Know? Recommending People to Invite into Your Social Network. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*. ACM, Sanibel Island, Florida, USA, 10.

[41] Ido Guy, Sigalit Ur, Inbal Ronen, Adam Perer, and Michal Jacovi. 2011. Do you want to know? recommending strangers in the enterprise. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work (CSCW ’11)*.
Association for Computing Machinery, New York, NY, USA, 285–294. doi:10.1145/1958824.1958867

Shefali Haldar, Yoojung Kim, Sonali R. Mishra, Andrea L. Hartzler, Ari H. Pollack, and Wanda Pratt. 2020. The Patient Advice System: A Technology Probe Study to Enable Peer Support in the Hospital. Proceedings of the ACM on Human-Computer Interaction 4, CSCW2 (Oct. 2020), 112:1–112:23. doi:10.1145/3415183

Jeong Yeob Han, Jung-Hyun Kim, Hye Jin Yoon, Minsun Shim, Fiona M. McTavish, and David H. Gustafson. 2012. Social and Psychological Determinants of Levels of Engagement with an Online Breast Cancer Support Group: Posters, lurkers, and Non-Users. Journal of health communication 17, 3 (2012), 356–371. doi:10.1080/10810730.2011.585696

F. Maxwell Harper, Shilad Sen, and Dan Frankowski. 2007. Supporting social recommendations with activity-balanced clustering. In Proceedings of the 2007 ACM conference on Recommender systems (RecSys ’07). Association for Computing Machinery, New York, NY, USA, 165–168. doi:10.1145/1297231.1297262

Charles R. Harris, K. Jarrod Millman, Stéfán J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Martin H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Rio, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. Nature 585, 7825 (Sept. 2020), 357–362. doi:10.1038/s41586-020-2649-2 Publisher: Springer Science and Business Media LLC.

Andrea L. Hartzler, Megan N Taylor, Albert Park, Troy Griffiths, Uba Backonja, David W McDonald, Sam Wahbeh, Cory Brown, and Wanda Pratt. 2016. Leveraging cues from person-generated health data for peer matching in online communities. Journal of the American Medical Informatics Association : JAMIA 23, 3 (May 2016), 496–507. doi:10.1093/jamia/ocv175

Andrea L. Hartzler, Bridget Weis, Carly Cahill, Wanda Pratt, Albert Park, Uba Backonja, and David W. McDonald. 2016. Design and Usability of Interactive User Profiles for Online Health Communities. ACM Transactions on Computer-Human Interaction 23, 3 (June 2016), 15:1–15:33. doi:10.1145/2930517

Miguel A. Hernán and J. M. Robins. 2020. Causal Inference: What If? (1 ed.). Chapman & Hall/CRC., Boca Raton. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Jeremy Heyer, Zachary Schmitt, Lynn Dombrowski, and Svetlana Yarosh. 2020. Opportunities for Enhancing Access and Efficacy of Peer Sponsorship in Substance Use Disorder Recovery. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–14. doi:10.1145/3313831.3376241

Sarah Holbrey and Neil S. Coulson. 2013. A qualitative investigation of the impact of peer to peer online support for women living with Polycystic Ovary Syndrome. BMC Women’s Health 13, 1 (Dec. 2013), 51. doi:10.1186/1472-6874-13-51

Julianne Holt-Lunstad, Timothy B. Smith, and J. Bradley Layton. 2010. Social Relationships and Mortality Risk: A Meta-analytic Review. PLOS Medicine 7, 7 (July 2010), e1000316. doi:10.1371/journal.pmed.1000316

Shih-Wen Huang, Jonathan Bragg, Isaac Cowhey, Oren Etzioni, and Daniel S. Weld. 2016. Toward Automatic Bootstrapping of Online Communities Using Decision-Theoretic Optimization. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work &amp; Social Computing (CSCW ’16). Association for Computing Machinery, New York, NY, USA, 582–594. doi:10.1145/2818048.2820024 event-place: San Francisco, California, USA.

Jane Hughes, Eleri Wood, and Gaynor Smith. 2009. Exploring kidney patients’ experiences of receiving individual peer support. Health Expectations 12, 4 (Dec. 2009), 396–406. doi:10.1111/j.1369-7625.2009.00568.x Publisher: John Wiley & Sons, Ltd.

J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. Computing in Science & Engineering 9, 3 (2007), 90–95. doi:10.1109/MCSE.2007.55 Publisher: IEEE COMPUTER SOC.

Omayma Husain, Naomie Salim, Rose Alinda Alias, Samah Abdelsalam, and Alzubair Hassan. 2019. Expert Finding in Online Social and Psychological Determinants of Levels of Engagement with an Online Breast Cancer Support Group: Posters, lurkers, and Non-Users. Journal of health communication

S.c. Ieropoli, V.m. White, M. Jefford, and D. Akkerman. 2011. What models of peer support do people with colorectal cancer prefer? European Journal of Cancer Care 20, 4 (2011), 455–465. doi:10.1111/j.1365-2354.2010.01218.x_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2354.2010.01218.x

Joshua Introne, Isabel Munoz, and Bryan Semaan. 2021. The Narrative Tapestry Design Process: Weaving Online Social Support from Stories of Stigma. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (Oct. 2021), 389:1–389:35. doi:10.1145/3479533

Maia Jacobs, James Clawson, and Elizabeth D. Mynatt. 2016. A Cancer Journey Framework: Guiding the Design of Holistic Health Technology. In Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth ’16). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 114–121. http://dl.acm.org/citation.cfm?id=3021319.3021337

Dietmar Jannach and Michael Jugovac. 2019. Measuring the Business Value of Recommender Systems. ACM Trans. Manage. Inf. Syst. 10, 4 (Dec. 2019), 16:1–16:23. doi:10.1145/3370082
[63] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2020. A Re-visit of the Popularity Baseline in Recommender Systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (July 2020), 1749–1752. doi:10.1145/3397271.3401233 arXiv: 2005.13829.

[64] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ’02)*. Association for Computing Machinery, New York, NY, USA, 133–142. doi:10.1145/775047.775067

[65] Thorsten Joachims, Ben London, Yi Su, Adith Swaminathan, and Legun Wang. 2021. Recommendations as Treatments. *AI Magazine* 42, 3 (Nov. 2021), 19–30. doi:10.1609/aimag.v42i3.18141 Number: 3.

[66] Joseph D. Y. Kang and Joseph L. Schafer. 2007. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statist. Sci.* 22, 4 (Nov. 2007), 523–539. doi:10.1214/07-STS227 Publisher: Institute of Mathematical Statistics.

[67] Naveena Karusala, David Odhiambo Seeh, Cyrus Mugo, Brandon Guthrie, Megan A Moreno, Grace John-Stewart, Irene Inwani, Richard Anderson, and Keshet Ronen. 2021. “That courage to encourage”: Participation and Aspirations in Chat-based Peer Support for Youth Living with HIV. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI ’21)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3411764.3445313

[68] Gary King and Richard Nielsen. 2019. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis* 27, 4 (2019), 435–454.

[69] Sarah D. Kowitt, Katrina R. Ellis, Veronica Carlisle, Nivedita L Bhushan, Kristin Z. Black, Kaitlyn Brodar, Nicole M. Cranley, Kia L. Davis, Eugenia Eng, Michelle Y. Martin, Jared McGuirt, Rebecca L. Sokol, Patrick Y. Tang, Anissa I. Vines, Jennifer S. Walker, and Edwin B. Fisher. 2019. Peer support opportunities across the cancer care continuum: a systematic scoping review of recent peer-reviewed literature. *Supportive Care in Cancer* 27, 1 (Jan. 2019), 97–108. doi:10.1007/s00520-018-4479-4

[70] Brian Lakey and Catherine J. Lutz. 1996. Social Support and Preventive and Therapeutic Interventions. In *Handbook of Social Support and the Family*, Gregory R. Pierce, Barbara R. Sarason, and Irwin G. Sarason (Eds.). Springer US, Boston, MA, 435–465. doi:10.1007/978-1-4899-1388-3_18

[71] Elizabeth C. Lavender, Esther Dusabe-Richards, Anna M. Anderson, Deborah Antcliff, Linda McGowan, Philip G. Conaghan, Sarah R. Kingsbury, and Grett A. McHugh. 2021. Exploring the feasibility, acceptability and value of volunteer peer mentors in supporting self-management of osteoarthritis: a qualitative evaluation. *Disability and Rehabilitation* 0, 0 (Sept. 2021), 1–11. doi:10.1080/09638288.2021.1964625 Publisher: Taylor & Francis eprint: https://doi.org/10.1080/09638288.2021.1964625.

[72] Eun-Hyun Lee. 2012. Review of the Psychometric Evidence of the Perceived Stress Scale. *Asian Nursing Research* 6, 4 (Dec. 2012), 121–127. doi:10.1016/j.ajnam.2012.08.004

[73] Richard M. Lee and Steven B. Robbins. 1995. Measuring belongingness: The Social Connectedness and the Social Assurance scales. *Journal of Counseling Psychology* 42, 2 (1995), 232–241. doi:10.1037/0022-0167.42.2.232

[74] Stephen J. Lepore, Joanne S. Buzaglo, Morton A. Lieberman, Mitch Golant, Judith R. Greener, and Adam Davey. 2014. Eun-Hyun Lee. 2012. Review of the Psychometric Evidence of the Perceived Stress Scale. *Asian Nursing Research* 6, 4 (Dec. 2012), 121–127. doi:10.1016/j.ajnam.2012.08.004

[75] Zachary Levonian, Marco Dow, Drew Erikson, Sourjot Ghosh, Hannah Miller Hillberg, Saumik Narayanan, Loren Terveen, and Svetlana Yarosh. 2021. Patterns of Patient and Caregiver Mutual Support Connections in an Online Health Community. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 275:1–275:46. doi:10.1145/3434184

[76] Zachary Levonian, Drew Richard Erikson, Wenqi Luo, Saumik Narayanan, Sabirat Rubya, Prateek Vachher, Loren Terveen, and Svetlana Yarosh. 2020. Bridging Qualitative and Quantitative Methods for User Modeling: Tracking Cancer Patient Behavior in an Online Health Community. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, Vol. 14. AAAI, Atlanta, GA, 12. https://www.aaai.org/ojs/index.php/ICWSM/article/view/7310

[77] Zachary Levonian, Chia-Jung Lee, Vanessa Murdoch, and F. Maxwell Harper. 2022. Trade-offs in Sampling and Search for Early-stage Interactive Text Classification. In *27th International Conference on Intelligent User Interfaces (IUI ’22)*. Association for Computing Machinery, New York, NY, USA, 566–583. doi:10.1145/3490099.3511134

[78] Changye Li, Zachary Levonian, Haiwei Ma, and Legun Wang. 2021. Recommendations as Treatments. *AI Magazine* 42, 3 (Nov. 2021), 19–30. doi:10.1609/aimag.v42i3.18141 Number: 3.

[79] Joseph D. Y. Kang and Joseph L. Schafer. 2007. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statist. Sci.* 22, 4 (Nov. 2007), 523–539. doi:10.1214/07-STS227 Publisher: Institute of Mathematical Statistics.

[80] Zachary Levonian, Chia-Jung Lee, Vanessa Murdock, and F. Maxwell Harper. 2022. Trade-offs in Sampling and Search for Early-stage Interactive Text Classification. In *27th International Conference on Intelligent User Interfaces (IUI ’22)*. Association for Computing Machinery, New York, NY, USA, 281–284. doi:10.1145/3272973.3274075

[81] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained Transformers for Text Ranking: BERT and Beyond. arXiv:2010.06467 [cs] (Oct. 2020). http://arxiv.org/abs/2010.06467 arXiv: 2010.06467.

[82] Eden Litt and Eszter Hargittai. 2016. The Imagined Audience on Social Network Sites. *Social Media + Society* 2, 1 (Jan. 2016), 2056305116633482. doi:10.1177/2056305116633482 Publisher: SAGE Publications Ltd.
[81] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs] (July 2019). http://arxiv.org/abs/1907.11692 arXiv: 1907.11692.

[82] Hongyu Lu, Min Zhang, and Shaoqing Ma. 2018. Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, Ann Arbor MI USA, 435–444. doi:10.1145/3209978.3210007

[83] Haiwei Ma, C. Estelle Smith, Lu He, Saurik Narayanan, Robert A. Giaquinto, Roni Evans, Linda Hanson, and Svetlana Yarosh. 2017. Write for Life: Persisting in Online Health Communities Through Expressive Writing and Social Support. Proc. ACM Hum.-Comput. Interact. 1, CSCW (Dec. 2017), 73:1–73:24. doi:10.1145/3134708

[84] Haley MacLeod, Grace Bastin, Leslie S. Liu, Katie Siek, and Kay Connelly. 2017. “Be Grateful You Don’t Have a Real Disease”: Understanding Rare Disease Relationships. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI ’17). Association for Computing Machinery, New York, NY, USA, 1660–1673. doi:10.1145/3025453.3025796 event-place: Denver, Colorado, USA.

[85] Sumaira H. Malik and Neil S. Coulson. 2008. Computer-mediated infertility support groups: An exploratory study of online experiences. Patient Education and Counseling 73, 1 (Oct. 2008), 105–113. doi:10.1016/j.pec.2008.05.024

[86] Annette Markham. 2012. Fabrication as Ethical Practice. Information, Communication & Society 15, 3 (April 2012), 334–353. doi:10.1080/1369118X.2011.641993 Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2011.641993.

[87] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Stéfan van der Walt and Jarrod Millman (Eds.), 56 – 61. doi:10.25000/Majora-92bf1922-00a

[88] Hannah Kang Moran, Joanna Veakey Brooks, and Lori Spozokak. 2021. “There’s somebody like me”: perspectives of a peer-to-peer gynecologic cancer mentorship program. Supportive Care in Cancer 29, 12 (Dec. 2021), 7679–7686. doi:10.1007/s00520-021-06348-w

[89] Arden Moulton. 2016. Woman to Woman: a hospital-based support program for women with gynecologic cancer and their families. In Social Work in Health Settings (4 ed.). Routledge, 12. Num Pages: 12.

[90] Arden Moulton, Amy Balbierz, Stephanie Eisenman, Elizabeth Neustein, Virginia Walther, and Irwin Epstein. 2013. Woman to Woman: A Peer to Peer Support Program for Women With Gynecologic Cancer. Social Work in Health Care 52, 10 (Nov. 2013), 913–929. doi:10.1080/00981389.2013.834031 Publisher: Routledge _eprint: https://doi.org/10.1080/00981389.2013.834031.

[91] Drashko Nakikj and Lena Mamykina. 2017. A Park or A Highway: Overcoming Tensions in Designing for Socio-emotional and Informational Needs in Online Health Communities. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW ’17). ACM, New York, NY, USA, 1304–1319. doi:10.1145/2998181.2998339

[92] Drashko Nakikj and Lena Mamykina. 2018. Lost in Migration: Information Management and Community Building in an Online Health Community. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18). Association for Computing Machinery, Montreal QC, Canada, 1–14. doi:10.1145/3173574.3173720

[93] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Uditi Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodimiry Kondratenko, Stephanie Pereira, Xianjie Chen, Wenzlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. arXiv:1906.00091 [cs] (May 2019). http://arxiv.org/abs/1906.00091 arXiv: 1906.00091.

[94] Zahra Nazari, Praveen Chandar, Ghazal Fazelnia, Catherine M. Edwards, Benjamin Carterette, and Mounia Lalmas. 2022. Choice of Implicit Signal Matters: Accounting for User Aspirations in Podcast Recommendations. In Proceedings of the ACM Web Conference 2022 (WWW ’22). Association for Computing Machinery, New York, NY, USA, 2433–2441. doi:10.1145/3485447.3512115

[95] Mark W. Newman, Debra Lauterbach, Sean A. Munson, Paul Resnick, and Margaret E. Morris. 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 Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW ’11). ACM, New York, NY, USA, 341–350. doi:10.1145/1958824.1958876

[96] Kathleen O’Leary, Arpita Bhattacharya, Sean A. Munson, Jacob O. Wobbrock, and Wanda Pratt. 2017. Design Opportunities for Mental Health Peer Support Technologies. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW ’17). ACM, New York, NY, USA, 1470–1484. doi:10.1145/2998181.2998349 event-place: Portland, Oregon, USA.

[97] Iván Palomares, James Neve, Carlos Porcel, Luiz Pizzato, Ido Guy, and Enrique Herrera-Viedma. 2021. Reciprocal Recommender Systems: Analysis of state-of-art literature, challenges and opportunities towards social recommendation. Information Fusion 69 (May 2021), 103–127. doi:10.1016/j.inffus.2020.12.001
[98] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[99] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830.

[100] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI ’20). Association for Computing Machinery, Honolulu, HI, USA, 1–15. doi:10.1145/3313833.3376695

[101] Luiz Pizzato, Tomek Rej, Thomas Chung, Irena Koprinska, and Judy Kay. 2010. RECON: a reciprocal recommender for online dating. In Proceedings of the 2010 ACM conference on Recomender systems (RecSys ’10). Association for Computing Machinery, New York, NY, USA, 207–214. doi:10.1145/1864708.1864747

[102] Anthony Poon, Matthew Luebke, Julia Loughman, Ann Lee, Lourdes Guerrero, Madeline Sterling, and Nicola Dell. 2023. Computer-Mediated Sharing Circles for Intersectional Peer Support with Home Care Workers. Proceedings of the ACM on Human-Computer Interaction 7, CSCW1 (April 2023), 1–35. doi:10.1145/3579472

[103] Claudette Pretorius, Darragh McCashin, Naoise Kavanagh, and David Coyle. 2020. Searching for Mental Health: A Mixed-Methods Study of Young People’s Online Help-seeking. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313833.3376328

[104] Stephen A. Rains. 2018. Coping with Illness Digitally. MIT Press. Google-Books-ID: EhPkuQEACAAJ.

[105] Stephen A. Rains and David M. Keating. 2011. The Social Dimension of Blogging about Health: Health Blogging, Social Support, and Well-being. Communication Monographs 78, 4 (Dec. 2011), 511–534. doi:10.1080/03637751.2011.618142

[106] Kathryn Rehberg, Adina Fleischmann, Elana Silber, Suzanne C’Neil, Frances Marcus Lewis, and Kenneth P Tercyak. 2021. Facilitators of peer coaching/support engagement and dissemination among women at risk for and surviving with breast cancer. Translational Behavioral Medicine 11, 1 (Jan. 2021), 153–160. doi:10.1093/tbm/ibz186

[107] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410

[108] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI ’09). AUAI Press, Arlington, Virginia, USA, 452–461.

[109] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural Collaborative Filtering vs. Matrix Factorization Revisited. In Fourteenth ACM Conference on Recommender Systems (RecSys ’20). Association for Computing Machinery, New York, NY, USA, 240–248. doi:10.1145/3383313.3412488

[110] Frank Riessman. 1965. The “Helper” Therapy Principle. Social Work 10, 2 (April 1965), 27–32. doi:10.1093/sw/10.2.27

[111] Herman Saksono. 2021. Algorithmic Patient Matching in Peer Support Systems for Hospital Inpatients. http://francisconunes.me/RealizingAIinHealthcareWS/papers/Saksono2021.pdf

[112] Herman Saksono, Vivien Morris, Andrea G Parker, and Krzysztof Z Gajos. 2023. Evaluating Similarity Variables for Peer Matching in Digital Health Storytelling. Proceedings of the ACM on Human-Computer Interaction 7, CSCW2 (Oct. 2023). doi:10.1145/3610060

[113] Sharron Sandhu, Paula Veinot, Gayathri Embuldeniya, Sydney Brooks, Joanna Sale, Sicong Huang, Alex Zhao, Dawn Richards, and Mary J. Bell. 2013. Peer-to-peer mentoring for individuals with early inflammatory arthritis: feasibility pilot. BMJ Open 3, 3 (Jan. 2013), e002267. doi:10.1136/bmjopen-2012-002267 Publisher: British Medical Journal Publishing Group Section: Rheumatology.

[114] Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference.

[115] Eva Sharma and Mummun De Choudhury. 2018. Mental Health Support and Its Relationship to Linguistic Accommodation in Online Communities. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18). ACM, New York, NY, USA, 641:1–641:13. doi:10.1145/3173574.3174215 event-place: Montreal QC, Canada.

[116] Jane M. Simoni, Julie C. Franks, Keren Lehavot, and Samantha S. Yard. 2011. Peer interventions to promote health: Conceptual considerations. American Journal of Orthopsychiatry 81, 3 (2011), 351–359. doi:10.1111/j.1939-0025.2011.01103.x
Handbook, Edwin B. Fisher, Linda D. Cameron, Alan J. Christensen, Ulrike Ehler, Yan Guo, Brian Oldenburg, and Frank J. Snoek (Eds.). Springer New York, New York, NY, 341–372. doi:10.1007/978-0-387-93826-4_12

[136] Debra Umberson and Jennifer Karas Montez. 2010. Social Relationships and Health: A Flashpoint for Health Policy. *Journal of Health and Social Behavior* 51, 1_suppl (March 2010), S54–S66. doi:10.1177/0022146510383501

[137] Elizabeth Van Couvering. 2007. Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality. *Journal of Computer-Mediated Communication* 12, 3 (April 2007), 866–887. doi:10.1111/j.1083-6107.2007.00354.x

[138] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (Aug. 2018), 1080–1088. doi:10.1093/jamia/ocy052

[139] Sruthi Viswanathan, Behrooz Omidvar-Tehrani, and Jean-Michel Renders. 2022. What is Your Current Mindset?. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–17. doi:10.1145/3491102.3501912

[140] Nina Wacholder. 2011. Interactive query formulation. *Annual Review of Information Science and Technology* 45, 1 (2011), 157–196. doi:10.1002/aris.2011.144050111

[141] Catherine Walsh, Diyi Yang, Robert Croft, Suzanne Skevington, Mari Lloyd-Williams, Gunn Grande, and Guillermo Perez Algorta. 2020. Peer support to maintain psychological wellbeing in people with advanced cancer: findings from a feasibility study for a randomised controlled trial. *BMJ Palliative Care* 19, 1 (Aug. 2020), 129. doi:10.1186/s12940-020-00651-x

[142] Ruyuan Wan, Zachary Levonian, and Svetlana Yarosh. 2020. How much is a “like” worth? Engagement and Retention in an Online Health Community. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, 413–417. http://doi.org/10.1145/3406865.3418320

[143] Catherine Walsh, Diane Roberts, Lynn Calman, Lynda Appleton, Robert Croft, Suzanne Skevington, Mari Lloyd-Williams, Gunn Grande, and Guillermo Perez Algorta. 2020. Peer support to maintain psychological wellbeing in people with advanced cancer: findings from a feasibility study for a randomised controlled trial. *BMJ Palliative Care* 19, 1 (Aug. 2020), 129. doi:10.1186/s12940-020-00651-x

[144] Sanford Weisberg. 2013. *Applied Linear Regression* (3rd edition ed.). Wiley, Hoboken, NJ. https://onlinelibrary.wiley.com/book/10.1002/0471704091

[145] Honygi Wen, Longqi Yang, and Deborah Estrin. 2019. Leveraging post-click feedback for content recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys ’19)*. Association for Computing Machinery, New York, NY, USA, 278–286. doi:10.1145/3298689.3347037

[146] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[147] Jun Xu, Xiangnan He, and Hang Li. 2020. Deep Learning for Matching in Search and Recommendation. *Foundations and Trends® in Information Retrieval* 14, 2–3 (July 2020), 102–288. doi:10.1561/1500000076 Publisher: Now Publishers, Inc..

[148] Diyi Yang. 2019. *Computational Social Roles*. PhD Thesis. Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

[149] Diyi Yang, Robert Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’19)*. http://kraut.hciresearch.org/sites/kraut.hciresearch.org/files/articles/Yang19-SocialRolesInOnlineHealthCommunities.pdf

[150] Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H. Chi. 2020. Mixed Negative Sampling for Learning Two-tower Neural Networks in Recommendations. In *Companion Proceedings of the Web Conference 2020*. ACM, Taipei Taiwan, 441–447. doi:10.1145/3366424.3368619

[151] Tiina Yli-Uotila, Anja Rantanen, and Tarja Suominen. 2014. Online Social Support Received by Patients With Cancer. *CIN: Computers, Informatics, Nursing* 32, 3 (March 2014), 118–126. doi:10.1097/CIN.0000000000000028

[152] Alyson L. Young and Andrew D. Miller. 2019. “This Girl is on Fire”: Sensemaking in an Online Health Community for Vulvodynia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19)*. ACM, New York, NY, USA, 129–129.13. doi:10.1145/3290605.3300359 event-place: Glasgow, Scotland Uk.

[153] Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *Comput. Surveys* (July 2022). doi:10.1145/3556536 Just Accepted.

[154] Qian Zhao, F. Maxwell Harper, Gediminas Adomavicius, and Joseph A. Konstan. 2018. Explicit or implicit feedback? engagement or satisfaction?: a field experiment on machine-learning-based recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, Pau, France, 1331–1340. doi:10.1145/3167132.3167275
Existing literature suggests a wide range of potential characteristics to incorporate in a peer recommender system. Table 7 lists peer characteristics identified in prior work as salient or important for effective peer matching. We distinguish these characteristics as either proposed as an implication of a particular study, used in practice to match peers in a study or support program, or expressed by participants as preferences for or barriers to effective peer support. There are characteristics we don’t represent in the table, such as abilities/skills [84], specific needs [31], interaction medium preferences (e.g. email) [24], existing social connections [24], etc.

In addition to the comparative works identified in sec. 2.3, Boyes surveyed cancer patients about the importance of specific shared characteristics such as gender, age, and cancer type, although this data is currently unpublished [16]. While less relevant to our work with the CaringBridge OHC, Saksono et al. surveyed single mothers to compare the importance of demographic, life, and lifestyle characteristics for encouraging physical activity [112].

A.1 Alternatives to recommendation

Recommendation is not the only available mechanism for facilitating online peer matching. Two notable alternatives are improving search and filter tools and designing enriched profile pages to make it easier to represent one’s diagnosis, expertise, and support needs [39]. Search is challenging in situations where a user’s needs are known only implicitly to the user or are challenging to express in terms the system will understand [13]. We suggest that peer support finding is an exploratory [8] search task (e.g. see Pretorius et al.’s discussion of person-centered help-seekers [103]). Even with rich peer profiles available, it is challenging for searchers to formulate a query that captures their needs and intent [139, 140]. Other search systems for finding people—such as expertise-finding systems—were created based on interfaces designed to capture users’ needs in a domain-specific query [58]. Mindsets is a recent example of the design work needed to capture domain-specific intents during query formulation [139]; additional research is needed in the peer support context to capture support seekers’ and providers’ intents. In contrast to search, recommendation offers opportunities to engage with potential peers without explicitly articulating a person’s current needs.

A.2 Peer matching without deep learning

Social networking sites have used approaches based on neighborhoods and similarity of interactions to connect with strangers specifically. Twitter’s “who to follow” recommendations used an alternative approach similar to PageRank that uses only the existing follow network to make recommendations [42]. Guy et al. explicitly attempted to recommend strangers in an enterprise setting based on number of shared interests and memberships [44]. The modeling problem closest to peer recommendation may be romantic relationship recommendation, a context that aims to encourage interaction between users and values reciprocity [101, 128].

B Use of search feature on CaringBridge

Do CaringBridge users use search to attempt to find information or supporters? To address this question, we collected a dataset of user-initiated search queries on CaringBridge. These queries were extracted from internal logs collected between July 4, 2021 and July 10, 2021. Our dataset
Table 7. Peer characteristics identified in prior work as important for effective peer matching. We included the starred characteristics (*) in our preference survey.

| Peer Characteristics          | Proposed | Used | Expressed Preference |
|------------------------------|----------|------|----------------------|
| **Health**                   |          |      |                      |
| Diagnosis*                   | [2, 5, 16, 31, 75, 84, 132] | [36, 88–90, 142] | [24, 30, 49, 129] |
| Treatment*                   | [16, 31] | [36, 59] | [24, 49, 129] |
| Symptoms*                    | [96]     | [24] | [113] |
| Severity                     | [31, 132]|      |                      |
| Timeline                     | [76]     |      | [10, 24, 49] |
| Health role                  | [75]     |      | [49] |
| Relevant knowledge           | [84]     | [56] | [10, 49, 52] |
| **Demographics**             |          |      |                      |
| Age                          | [5, 16, 31, 132] | [2, 49, 56, 59, 88–90] | [10, 24, 49] |
| Gender                       | [4, 5, 16, 31] | [49, 56, 59, 71, 112, 142] | [49, 113] |
| Ethnicity                    | [27, 56, 112] |      | [10, 17] |
| Sexual orientation           | [113]    |      |                      |
| Nationality                  | [49]     |      |                      |
| **Life**                     |          |      |                      |
| Geography/location*          | [5, 75, 132] | [27, 36, 71, 112, 142] | [18, 49] |
| Cultural values/background*   | [16, 106] | [89, 90] | [17, 24, 30] |
| Employment                   | [56]     |      | [24] |
| Religion                     | [90]     |      | [10] |
| Politics                     |          |      |                      |
| Socio-economic status        | [113]    |      |                      |
| Education level              | [129]    |      |                      |
| Social role*                 | [116, 132, 149] | [89] | [49] |
| Marital status               | [31]     | [56, 59, 112] | [10, 49] |
| Has children?                | [59, 88, 90, 112] | [49] | |
| Language                     | [89]     |      |                      |
| **Other**                    |          |      |                      |
| Communication style*         | [95]     | [49] | [129] |
| Lifestyle                    | [71, 112] |      | [18, 24, 30] |
| Interests                    | [2, 36]  |      | [24] |
| Personality                  | [71]     |      | [52, 129] |
| Commitment to support & recovery | [96]     |      | [30, 52] |

contained 103,830 searches comprising 32,722 unique query strings. We preprocessed the query strings by splitting them into tokens based on whitespace.

Based on a random sample of 100 queries with 1, 2, or 3 tokens and visual inspection of additional samples, all queries corresponded to either person names or existing site URL strings. Thus, we conclude that a very high percentage of searches are for a specific CaringBridge site. Queries with 4 or more tokens comprise < 1% of the query strings. Visual inspection suggests that the majority of queries with 4+ tokens are help requests or open-domain queries including spam. We could identify
Table 8. Associational differences between sites that receive interactions (ints) from at least one peer author and sites that only receive interactions from visitors. Site tenure is the number of months between the first and last Journal update on a site. # updates is the number of Journal updates published more than 30 days after the first update. The common language effect size (CLES) is reported where appropriate. All comparisons are significant at the 99.5% significance level.

|                      | 1+ peer int within 30 days | Non-peer ints only | Difference | CLES  |
|----------------------|----------------------------|--------------------|------------|-------|
| Site Count           | 92,352 sites               | 100,424 sites      | -8,072 sites | -     |
| Site Tenure (Median) | 3.9 months                 | 0.8 months         | +3.2 months | 62.4% |
| # updates (M; SD)    | 20.5 (47.6)                | 12.4 (35.7)        | +8.1 updates | 62.1% |
| # updates (Median)   | 8 updates                  | 2 updates          | +6 updates  | -     |
| % sites with 2+ updates | 78.1%                  | 59.5%              | +18.7pp    | -     |

no instances of users searching for e.g. a particular health condition, treatment, or symptom; if search is used in this way, it occurs at a low prevalence. Results were identical analyzing queries made specifically by author users.

C  Associational evidence of peer interactions on author behavior

Based on sec. 2.1, we would expect that peer interaction is associated with behavior change for authors. Prior work demonstrates that receiving interactions from visitors—peers and non-peers—is associated with retention on CaringBridge [143], but are peer interactions more impactful than non-peer interactions? Table 8 shows associational differences between CaringBridge sites based on the interactions (reactions, comments, and guestbooks) received from visitors within 30 days of a site’s first published Journal update. Sites for which at least one visitor interaction is left by a peer author will publish on CaringBridge for a median of 3.2 additional months (with 6 additional updates) compared to sites that receive only interactions from visitors. This analysis was conducted on 192,776 sites created between January 1, 2014 and September 1, 2021 (the start of the study) that received at least 1 visitor interaction in the first 30 days. These general results hold when adjusting for publishing rate, number of interactions received, number of visitors, and number of peer visitors. We omit full modeling tables (Poisson for # updates and Cox’s proportional hazards model for site tenure).

D  Modeling and optimization details

This section provides additional details beyond those provided in sec. 3.3. Modeling and analysis code make primary use of Python’s scikit-learn [99], statsmodels [114], transformers [146], NumPy [48], pandas [87], and Matplotlib [57] packages. The recommendation model was trained using PyTorch [98].

D.1  Implicit feedback

We processed all historical CaringBridge author interactions since January 1st, 2010. We train our recommendation models on implicit feedback—the behavioral signals that indicate an author is interested in reading a site or interacting with the author of that site [108]. Three potential sources of implicit feedback are available: (1) first visits—when an author visits another author’s site for the first time, (2) initiations—when an author interacts with a site for the first time, or (3) reciprocations—when an author initiates with a site and an author of that site subsequently interacts.
on the initiator’s site.\textsuperscript{21} Choice of implicit feedback signal can have a significant impact both on what recommendations are shown to users and how those users engage with those recommendations.\textsuperscript{94} We choose initiations as our implicit feedback signal, as it is less noisy than site visits (i.e. a visit may not indicate interest) while more plentiful than reciprocations (i.e. reciprocations are less common than unreciprocated initiations\textsuperscript{75}, which means less data available for model training). By selecting initiations, we assume that leaving a reaction, comment, or guestbook on another author’s site indicates a preference for reading that site and interacting with that author relative to other sites. Optimizing for recommendations that increase initiations is likely to increase actual initiations\textsuperscript{154}, but we acknowledge a semantic gap between the implicit feedback metric and metrics of interest; a construct like "perceived social support" may not increase despite receiving more peer interaction\textsuperscript{26, 70}.

For each initiation between a source author and a target site, we generate training data. Using initiations presents three complexities: (1) a user may initiate with a site before they become an author, (2) authors may write multiple blogs, and (3) authors may co-write with other authors on a single blog. We address these complexities by ranking \textit{author/site pairs} rather than sites alone. For each initiation, we generate one positive sample for each author on the target site (target author/site pairs) and each site written by the source author (source author/site pairs). An author/site pair is \textit{eligible} if the author has published at least three Journal updates on that site—a minimum activity threshold that ensures sufficient data is available during feature extraction. We include in the training data only initiations between eligible author/site pairs.\textsuperscript{22} At prediction time, given a recommendation-seeking source author, we score all candidate author/site pairs: eligible author/site pairs that have been active on CaringBridge in the last week and have not previously been interacted with by the source author.\textsuperscript{23} As we ultimately recommend sites to authors, we convert the ranking of candidate author/site pairs into a ranking of sites. If a site appears among candidate pairs twice (because that site has two authors), we remove all but the highest score from that site. If a recommendation-seeking author is an author of multiple sites, we merge the scores for each source author/site pair by averaging the scores for each candidate pair.\textsuperscript{24}

For each positive sample generated by initiation extraction, we sample an assumed-negative author/site pair to add to the training data.\textsuperscript{25} The negative author/site pair is randomly selected from among the candidates. As a history of previous initiations are not required for eligibility, we avoid selection bias issues that arise in models that require user feedback on negative samples.\textsuperscript{150}

\subsection*{D.2 Model features}

Given two author/site pairs, we use the recent behavior of the authors on CaringBridge to construct a feature representation of the two, summarized in Figure 11. Different models might use subset of these features, but we describe here the full set included in the model deployed during the field study.

\textsuperscript{21}A fourth option is \textit{follows}. Twitter, for example, used reciprocal follows as an implicit signal for some user recommendations\textsuperscript{126}. Due to the unusual nature of Follows on CaringBridge (see sec. 3.1) and lack of clarity around this feature’s usage, we omit it from consideration.

\textsuperscript{22}Eligibility is required at the time of the initiation; if an initiating author has not (yet) published three Journal updates on a site, no training samples will be produced for that initiation.

\textsuperscript{23}An active author is one that has created a Journal update, comment, guestbook, or reaction on any site within the last week. By focusing only on recommending active authors, we ensure that a recommendation-seeking author could receive a response from a site’s author if they leave a comment.

\textsuperscript{24}These merging strategies reflect the intuition that a good match with any one author of a site makes the site relevant. In practice, these strategies have minimal impact on rankings.

\textsuperscript{25}While more complex approaches to negative sampling have been recently proposed\textsuperscript{28}, we adopt uniform sampling of a single negative as a widely-used baseline.
Peer Recommendation Interventions for Health-related Social Support: a Feasibility Assessment

Fig. 11. Features available for recommendation, including 780 features for the recommendation-seeking source author/site pair and the same for each candidate author/site pair. Three dyadic features capture the relationship between the source and candidate within the initiation network. Total features: 1563

Fig. 12. Dataset splits and associated author initiation totals. All reported metrics are from the test period.

- **Text** (768 × 2 features): We incorporate context from the three most recent Journal updates written on a site. For each Journal update, we use the pre-trained RoBERTa [81] model available in the HuggingFace Transformers package [146] to compute size-768 contextualized word embeddings. Then, we mean pool the token embeddings and then the update embeddings to produce a single vector representation, an effective general approach to using word embeddings [77, 79, 107].

- **Activity** (9 × 2 features): For each of Journal updates, reactions, comments, and guestbooks, we include the count of that action within the last week and the time elapsed to the most recent action (in hours). In addition, we include the author’s current tenure (time elapsed to the author’s first published journal update).

- **Network** (6×2+3 features): During initiation extraction, we maintain the interaction network between authors as described by Levonian et al., in which new edges are created between authors when an initiation occurs [75]. For each author/site pair, we include: indegree, outdegree, and weakly-connected component size. In addition, we include three dyadic features: whether the two authors are weakly connected, whether the candidate author is the friend-of-a-friend of the source author (i.e. this initiation would create triadic closure), and whether the candidate author has previously initiated with the source author (i.e. this initiation would be a reciprocation to a prior initiation).

D.3 Offline evaluation

Given an initiation, the goal for our recommender system is to rank the site that was actually initiated with as high as possible: ideally in rank 1, above other candidates. We evaluate various modeling approaches by splitting initiations chronologically into a training, validation, and test set—see Figure 12 for initiation counts. Offline evaluations can diverge substantially from the usage preferences expressed by recommendation consumers [11]; we conduct one here to compare recommendation algorithms in terms of accuracy, coverage, and diversity of predictions on historical initiations.

During evaluation of each initiation, features are captured at the millisecond before the initiation actually occurred: for the source author/site pair (who did the initiation), for the target (who is being initiated with), and for the candidates (who are eligible, active author/site pairs that the source had not previously initiated with). This evaluation approach intends to reward models

26Analysis during model development suggested minimal impact of the pooling strategy (mean vs max vs concatenation) on model performance.
that rank a site well if the source author/site pair actually initiated with that site at the time the recommendation was made. The number of eligible, active candidates varies according to the time of the initiation; the median number of eligible, active authors during the test period was 13,252.

We used two primary evaluation metrics: mean reciprocal rank (MRR) and hit rate (HR). Mean reciprocal rank is computed based on the rank \( r \) assigned to the target site, where 1 is the best rank given to the highest score.\(^{27}\) MRR is the mean of the reciprocal ranks \( (1/r) \) for every test initiation. Hit rate is the proportion of the time that \( r \) is less than some threshold. As we provide 5 recommendations in a Site Suggestion email, we report HR@5 (how often would the model recommend the target site in a five-site set) and HR@1 (how often would the model rank the target site first). When we compare multiple hyperparameter configurations for a model, we select the best on the basis of validation MRR and report test metrics as the median from 3 random seeds.

Beyond accuracy-style metrics, we consider coverage as an important secondary goal \([40]\). Coverage has two aspects: (1) how many users can receive recommendations or have their sites recommended and (2) the diversity of sites that are recommended in practice. The first aspect of coverage is model agnostic and heavily affected by our decision to require three Journal updates to be eligible. Of 124,051 new authors in 2020, only 51.4% will publish three updates. Further, of those authors that do publish 3 updates, recommendations cannot be delivered until the publication of the 3rd update—a median wait of 2.5 days, but at least 72.4 days for the slowest-publishing 10% of authors. Beyond authors, a hypothetical system that served recommendations to any visitor with an interaction could reach an additional 1.3M new visitors who first interacted in 2020.

The second aspect of coverage—the diversity of recommended sites—is model dependent. To evaluate the coverage of recommended sites, we randomly sampled 1000 eligible, active authors at the end of the training period\(^{28}\) and produced recommendation sets with the 5 highest-scoring sites for each author. We then compared the sites that were actually recommended (R) to the sites that were not recommended (N), among the 12,432 candidate sites available at noon UTC on January 1, 2021. We consider three plausible goals for diverse recommendations, with an associated metric for each goal. The first goal is that a large number of unique sites are recommended rather than a few sites recommended many times; we compute the percentage of unique recommendations given \(#R/#5000\). The second goal is that newer sites are recommended, a period during which support may be particularly impactful \([43]\); we compute the minimum site tenure for each recommendation set and report the mean. The third goal is that authors without previous interactions from other authors are recommended; we compute the proportion of recommended sites that have no prior connections \(#SR/#R\) for “siloed” sites \(S_R\) and report the ratio to the proportion for non-recommended sites \(#SN/#N\). If the ratio is less than 1, that model recommends fewer siloed sites than would be expected by chance.\(^{29}\)

### D.4 Model comparison

We describe the recommendation models we implemented and compared, starting with the model we used during the field study. Table 9 presents the performance of these models in the order we describe them.

**MLP.** We used a multi-layer perceptron (MLP) with two hidden layers as a parsimonious yet effective deep recommender model. All source, candidate, and dyadic features are concatenated

---

\(^{27}\)Note that we punish ties by using competition ranking e.g. two sites scored the same both get assigned rank 2 and no site is assigned rank 1. At prediction time, we break ties randomly.

\(^{28}\)Specifically, recommendations are generated at the end of the training period + 12 hours. This approach captures a situation where a model is trained nightly and recommendations are generated the next day, as in the field study we conducted.

\(^{29}\)Note that existing initiations are with siloed sites only 2.8% of the time during the test period.
Table 9. Offline test performance for various accuracy and coverage metrics. Coverage is reported in terms of number of unique recommended sites ($|R|$), the percentage of unique recommendations made (out of 5000 total recommendations), and the mean of the minimum site tenure in each recommendation set (MMST). The final column displays the percentage of reced sites that have no prior connections (i.e. are “siloed”, $|S_U|/|R|$), the same percentage for non-recced sites, and the ratio of those percentages. The bolded model was used during the field study.

| Model                  | MRR   | HR@1  | HR@5  | |R|   | %Unique | MMST  | $|S_U|/|R|$/$|S_U|/|U|$ |
|------------------------|-------|-------|-------|-----|-----|---------|-------|-----------------|
| MLP\text{\textsubscript{Tuned}} | 0.173 | 13.37%| 20.27%| 665 | 13.3%| 6.2 weeks| 10.8% / 24.6% = 0.44 |
| MLP\text{\textsubscript{Study}}   | 0.163 | 12.76%| 19.00%| 735 | 14.7%| 5.6 weeks| 10.6% / 24.7% = 0.43 |
| PeopleYouKnow          | 0.102 | 7.86% | 13.07%| 3972| 79.44%| 29.2 weeks| 14.9% / 28.1% = 0.53 |
| CosSim                 | 0.002 | 0.05% | 0.25% | 3403| 68.1%| 27.1 weeks| 25.2% / 23.3% = 1.08 |
| MF                     | 0.002 | 3.05% | 15.23%| 13  | 0.26%| 40.3 weeks| 0.0% / 23.9% = 0.00 |
| RecentInits            | 0.036 | 1.27% | 4.81% | 6   | 0.12%| 0.4 weeks | 0.0% / 23.9% = 0.00 |
| MostInits              | 0.035 | 1.32% | 4.90% | 12  | 0.24%| 112.9 weeks| 0.0% / 23.9% = 0.00 |
| NewestAuthor           | 0.025 | 0.97% | 3.19% | 6   | 0.12%| 0.1 weeks | 83.3% / 23.8% = 3.50 |
| RecentJournals         | 0.016 | 0.19% | 1.58% | 6   | 0.12%| 3.1 weeks | 16.7% / 23.9% = 0.70 |
| MostInteractive        | 0.007 | 0.17% | 0.68% | 5   | 0.10%| 0.9 weeks | 0.0% / 23.9% = 0.00 |
| MostJournals           | 0.004 | 0.13% | 0.38% | 5   | 0.10%| 0.4 weeks | 60.0% / 23.8% = 2.52 |
| Random                 | 0.001 | <0.0% | 0.04% | 4159| 83.18%| 17.6 weeks| 23.7% / 23.9% = 0.99 |

We compare the MLP model to two personalized baselines. \text{PeopleYouKnow} uses only the dyadic network features, scoring highest the sites on which authors have already initiated with the source, then sites that are friends-of-friends with the source, then sites that are in the same network component as the source, and finally all other sites. \text{CosSim} scores author/site pairs by computing the cosine similarity between source and candidate. Cosine similarity was used by Hartzler et al. to match peers by health interest [49].

We focus on two non-personalized baselines. \text{MostInits} scores each site by the number of initiations received by that site within the last week.\textsuperscript{30} In other recommendation contexts, popularity baselines like MostInits are strong contenders; not so on CaringBridge, where popularity follows a flatter distribution. MostInits has the best MRR of several other plausible non-personalized activity baselines (defined in App. D). \text{Random} ranks sites randomly, and is included as a useful point of comparison.

Table 9 compares the models by accuracy and coverage metrics. The best model (MLP\textsubscript{Tuned}) would recommend the site that was actually initiated with more than 20% of the time; the study model (MLP\textsubscript{Study}) performs slightly worse. Both MLP models are more likely to recommend newer sites than the baseline models, but less likely to recommend a variety of sites and sites that had

\textsuperscript{30}Following the recommendation of Ji et al., we use a popularity baseline that takes into account the time point when a user interacts with the system [63].
Table 10. Offline performance of an MLP model trained with combinations of the activity (A), network (N) and text (T) feature sets. The model using all feature sets (A+N+T) is MLP\textsubscript{Tuned} from Table 9.

| Feature sets | MRR  | HR@1 | HR@5  | | | | Site Age | | |
|--------------|-----|------|------|---|---|---|------|---|---|
| A+N         | 0.204 | 16.12% | 23.62% | 598 | 12.0% | 0.9 weeks | 12.0% / 24.4% = 0.49 |
| A+N+T       | 0.173 | 13.37% | 20.27% | 665 | 13.3% | 6.2 weeks | 10.8% / 24.6% = 0.44 |
| N+T         | 0.144 | 11.75% | 16.58% | 803 | 16.1% | 5.4 weeks | 12.5% / 24.6% = 0.51 |
| N (Network) | 0.136 | 11.70% | 15.42% | 842 | 16.8% | 56.4 weeks | 11.6% / 24.7% = 0.47 |
| A (Activity) | 0.058 | 3.23% | 6.99% | 11 | 0.2% | 0.8 weeks | 9.1% / 23.9% = 0.38 |
| A+T         | 0.043 | 1.73% | 5.25% | 181 | 3.6% | 0.8 weeks | 16.0% / 24.0% = 0.67 |
| T (Text)    | 0.017 | 0.58% | 1.85% | 98 | 2.0% | 0.6 weeks | 13.3% / 23.9% = 0.55 |

Fig. 13. Site Suggestion email generation during the field study. Recommendations were typically sent around noon, 36 hours after the database snapshot was taken.

D.5 Feature ablations

Providing personalized recommendations requires some form of sensitive data collection from users [130], but privacy and other ethical concerns necessitate collecting as little sensitive data as is required to produce useful recommendations (or in some cases opting not to provide recommendations at all). Pervasive data collection contributes to perceptions of recommendation services as invasive “little brothers” [41], changing behavior and undermining the potential benefits of recommendation. Therefore, we compared the relative importance of the different feature sets in order to understand the utility of collecting particular types of sensitive data—results are shown in Table 10. Surprisingly, the model trained without text features performs better on all accuracy metrics and similarly in terms of coverage, and it is also more likely to recommend newer sites. This analysis suggests that the collection of textual data may not be necessary and peer recommendation may still be practical if interaction network data is available: reasonable recommendations could be generated purely based on usage metadata. We still chose to include text features in the model we evaluated during the field study, as otherwise new authors with minimal activity and no peer interactions would all receive the same recommendations, conflicting with our design goal of personalization even in the cold-start setting. We discuss these results and this decision further in Appendix D.9.
D.6 Model deployment & required resources

Figure 13 shows the steps required to generate a batch of Site Suggestion emails. The most time-intensive step was the anonymization and manual transfer of a nightly database snapshot, which led to a 36-hour lag time between the snapshot and the associated email batch.\textsuperscript{31} We created the Site Suggestion emails by retraining the recommender model on the most recent snapshot and scoring all eligible, active candidate sites. Rather than recommending the top 5 sites by score, we decided to limit the total number of times a site was recommended in a single batch to at most 10 times. We made this decision for two reasons: first, as the perception of stranger visits by site authors is unknown, a large influx of strangers might create a negative experience for the recommended site’s authors; second, this restriction ensures that a diversity of sites are recommended each week. To achieve this limit, we randomly drafted sites by score until each participant had five recommendations.\textsuperscript{32} For ethical reasons, the first author manually inspected all recommended sites, manually excluding sites that would be inappropriate to send to participants. Two sites were removed this way: one for spam and one for COVID-19-related health misinformation. This small percentage of potentially harmful recommendations suggests that after some initial validation of quality, individual recommendations need not be inspected in practice.

For future studies, the actual resource requirements could be reduced by computing recommendations in-house. Disk usage for the weekly snapshot and the feature databases was approximately 96GB. The most RAM-intensive processing is maintaining the author interaction network during feature extraction; all other processing tasks are parallelizable and generally IO-bound. The CPU-based training and inference of the recommendation model could be accelerated with the use of GPUs, although at 1.5s per participant we found inference to be fast enough to serve even a much larger participant population.

D.7 MLP model details

Both MLP\textsubscript{Study} and MLP\textsubscript{Tuned} use two hidden layers, ReLU activation functions, and dropout. Optimization occurs over 1000 epochs with a one-cycle learning rate scheduler proposed by Smith and Topin \cite{smith2017cyclical} and the Adam optimizer with default hyperparameters. Following best practices to prevent overfitting, we used for prediction the weights from the epoch with the lowest hold-out

\textsuperscript{31}This 36-hour lag time could mean that the Journal update previewed in the email was no longer the most recent when the emails were sent, but we deemed this lag acceptable as it gave us time to do weekly robustness checks on the trained model and the identified recommendations.

\textsuperscript{32}In the first batch, we applied the limiting only to the bottom quarter of recommended sites by total visit count. See details of the drafting process and an analysis of the small impact on recommendation quality in Appendix I.
loss. Hold-out loss was computed over a random 1% of training initiations. We conducted random search experiments (not reported) using the learning rate and Adam hyperparameter distributions given by Sivaprasad et al., finding minimal differences [117]. MLP\textsubscript{Study} uses 100 hidden units per layer with a dropout $p$ of 0.1. MLP\textsubscript{Tuned} uses 300 hidden units per layer with a dropout $p$ of 0.5 and adds a weight decay of 0.0001. Hyperparameter tuning occurred using grid search to set hidden units $\in \{100, 300, 500\}$, dropout $p \in \{0.1, 0.5, 0.9\}$, weight decay $\in \{0, 0.0001, 0.01\}$, and maximum learning rate $\in \{0.008, 0.016\}$. We fit models from 3 random seeds at each hyperparameter combination and selected parameters based on the median of the 3 models’ MRR. We did not explore additional model architectures in depth, and offline metrics could be improved through the use of a model closer to the state-of-the-art or through a larger hyperparameter sweep.

One downside to our chronological validation and testing sets is that this setup deviates from the common modeling practice of retraining on regular intervals. Fortunately, Figure 14 indicates that this affect has a minimal impact on our offline evaluation.

D.8 Other Baselines

This section defines less relevant baselines and provides more details on the baselines shown in Table 9. We explored CosSim with other feature sets; they all perform worse in terms of MRR and Coverage metrics than using all features. MF is the conventional matrix factorization approach to collaborative filtering, using the dot product to compute similarity and selecting hyperparameters as described by Rendle et al. [109]:³³ note the high hit-rate but low MRR, indicating a strong popularity bias. The new non-personalized baselines are RecentInits, which ranks sites by the amount of time since the last initiation with that site, MostJournals/RecentJournals, which mirror MostInits and RecentInits but count published Journal updates rather than initiations, NewestAuthor, which ranks newest sites first, and MostInteractive, which ranks sites by the number of recent interactions made by authors on that site.

D.9 Feature ablation discussion

The results presented in Table 10 suggest that RoBERTa text features are not important for peer recommendation, and that recommendations based solely on text data would require a different approach than the one we use here. We suspect that the relative unimportance of text data reflects a bias in our implicit feedback signal and offline evaluation metrics toward authors already known to the source. Chen et al. observed that social network information was more effective for discovering known contacts, while content similarity was more effective for discovering new connections [22]—a dynamic that may be recurring here, as most historical initiations are between known contacts [75]. Even if text-based recommendations are less useful, users may perceive them to be less invasive than network-based recommendations (or vice-versa). Careful assessment of the perceived acceptability of this data collection for peer recommendation is necessary [33].

We chose to deploy the model that included the text data (MLP\textsubscript{Study}) in part because it will still provide personalized recommendations, even to authors who have not yet interacted with peers. We can observe this effect in the coverage predictions: while recommended site diversity is higher for authors who had previously initiated (342 unique sites for 439 source users, each site recommended on average 6.4 times) compared to authors who had no previous interactions (303 unique sites for 561 source users, each site recommended on average 9.3 times), these differences are relatively small. Further, these recommendations are of a similar quality: MRR on test initiations is actually

³³All matrix factorization models were trained for 100 epochs with the same optimizer as the MLP model; the best model (by validation MRR) used embeddings of size 128 and weight decay of 0.0001. Embeddings were trained for the 41,567 authors and 79,588 sites that appear at least twice in training-period initiations; a single embedding each was reserved for previously unseen authors and sites.
higher for people with no previous initiations than previous initiators (0.190 vs 0.147). Thus, we were satisfied that deploying MLPStudy satisfied our cold-start design requirement.

E  Survey Materials
All questions other than the eligibility, consent, and email questions are optional. All surveys were hosted on the Qualtrics platform.

E.1  Preference Survey
All registered CaringBridge authors (18+ years old) are invited to take this 1-minute survey, which is part of a research collaboration by CaringBridge and a team of technology researchers at the University of Minnesota.

The purpose of CaringBridge is to help people get the support they need during health journeys—and support comes from many places! Many authors choose to make their CaringBridge sites open for anyone to read. We want to send you emails with links to CaringBridge sites that we think you’ll want to follow.

If you are interested in participating in our study, we’ll send you personalized emails with links to CaringBridge sites. (As always, [privacy comes first] on CaringBridge: no one can read your site unless you want them to.) Opting in and sharing your opinion will help CaringBridge and the research community design features that make giving support easy and make a difference in the lives of the patients and caregivers you care about. Complete information and an FAQ for this study are available by clicking here. [link]

Click the arrow in the lower right corner of your screen to take the survey.

—Page Break—

E.1.1  Is Adult. I am 18 years or older.

• Yes, I am 18 years or older.
• No, I am less than 18 years old.

E.1.2  Is Registered. Do you have a registered CaringBridge account?

• Yes, I have a CaringBridge account.
• No, I don’t have a CaringBridge account.

E.1.3  Is Author. Are you can author or co-author of a CaringBridge site? You’re an author or co-author if you’ve ever written a [Journal] update on a CaringBridge site.

• Yes, I am an author or co-author on a CaringBridge site.
• No, I am a visitor to CaringBridge and I haven’t helped author a site.

E.1.4  Opt-in. Do you agree to receive personalized follow-up emails with links to CaringBridge sites that we suggest?

• Yes, I want to participate in the study and receive follow-up emails with site suggestions.
• No, I am not interested in participating in the study.

—Page Break—

E.1.5  Email. We just need the email address that you use with CaringBridge so that we can send you personal site suggestions.

We won’t share this email address with anyone outside of CaringBridge or the research team; we’ll only use this email address to contact you and to connect to your CaringBridge user account.

[Link to https://www.caringbridge.org/what-we-offer/the-privacy-you-choose]
[Link to https://www.caringbridge.org/what-we-offer/a-journal-for-your-journey]
• The email address I use with CaringBridge is: [Free Response]
• I’m not sure which email address is associated with my CaringBridge account, but the email I use most is: [Free Response]
• I changed my mind about participating in the study.

E.1.6 Wants study result follow-up. **Are you interested in reading about the results of this study when they become available in the future?** If you select yes, we will send you an email update with information about our results.

• Yes, please email me with the results of this study.
• No, thanks.

—Page Break—

Thanks for providing your info, you’ll get emails with site suggestions from us soon. The following completely optional questions will help us create better site suggestions for you.

E.1.7 Previous visit to stranger site. (Optional) Have you ever visited the CaringBridge site of an author who you did not know personally?

• Yes, I have visited the CaringBridge site of someone I didn’t know.
• No, I have never visited a stranger’s CaringBridge site.

E.1.8 Motivations. (Optional) Which of the following might motivate you to visit a fellow author’s CaringBridge site, even if you didn’t personally know them? Check all that apply.

• To learn from the journeys of other CaringBridge authors.
• To help mentor or support newer CaringBridge authors.
• To receive advice or support from more experienced authors.
• To communicate with a peer who understands.
• I’m not interested in visiting other authors’ CaringBridge sites right now, but I might want to in the future.
• I’m not interested in visiting other authors’ CaringBridge sites right now, but I would have wanted to in the past.
• I’m never interested in visiting other authors’ CaringBridge sites.
• Something else: [Free Response]

E.1.9 Characteristics. (Optional) What characteristics of an author or their site would make you want to read & engage with that person’s CaringBridge site? Check all that apply.

• High-quality writing or photos
• Similar diagnosis or symptoms to you or the loved one you care for
• Similar treatments to you or the loved one you care for
• Lives near me
• Similar cultural background to you or the loved one you care for
• For caregivers: Sharing the same relationship (e.g. spouse, child) to the person they care for
• Something else: [Free response]

E.1.10 Free Response – General. (Optional) Anything else you want to share with us about visiting the CaringBridge site of a fellow author? [Free Response]

E.2 Feedback Survey
Thank you for providing feedback. All questions on this page are optional: submit your feedback by clicking the right arrow at the very bottom of this survey form.
Not sure why you received this email & survey? You agreed to receive site suggestion emails in a survey you took in August. See the FAQ [link]. If you don’t want to receive these emails anymore, you can unsubscribe [link].

**E.2.1 Overall Interest.** Overall, did the suggested sites seem interesting to you?
- Yes, the sites were generally interesting to me.
- Unsure or neutral.
- No, the sites were generally uninteresting to me.

**E.2.2 Free Response – Interesting.** Briefly, what seemed interesting to you about the suggested sites? [Free Response]

**E.2.3 Free Response – Uninteresting.** Briefly, what seemed uninteresting to you about the suggested sites? [Free Response]

**E.2.4 Specific Recommendation Relevance.** Specifically, how relevant did you find each of the suggested sites? [5-item response matrix with levels: Very Relevant; Somewhat Relevant; Unsure or Neutral; Somewhat Irrelevant; Very Irrelevant, Offensive, or Spam]
- 1st suggested site
- 2nd suggested site
- 3rd suggested site
- 4th suggested site
- 5th suggested site

**E.2.5 Free Response – General.** Any other thoughts you want to share about these site suggestions or your experience so far in this study? [Free Response]

**E.3 Unsubscribe Survey**

**E.3.1 Page 1.** To unsubscribe: just enter your email address on the line below and click the right arrow.

Or, maybe you’re looking for a feedback form [link] or the study FAQ [link] (including contact details for study coordinators) instead.

**E.3.2 Email.**
- Unsubscribe me from additional emails from cb-suggestions@umn.edu. My email address is: [Free response]

—Page Break—

**E.3.3 Free Response – Unsubscribe.**
- Tell us why you unsubscribed? [Free response]
- Anything else you’d like to share with us? Thank you for participating in our study! [Free response]

**F Click data**

We identify recommendation clicks in the Site Suggestion emails via three data sources: (1) Google Analytics counts, (2) CloudFront logs, and (3) logged-in user visits. Links to the recommended sites contain UTM tracking information, including the site, the batch ID, and a unique participant ID.

Each data source has limitations. The Google Analytics summary can only provide a total count and won’t be incremented if JavaScript is disabled. An event won’t be captured in the CloudFront...
log if UTM tracking tags are stripped and in unknown other cases.\textsuperscript{36} Logged-in visits will only be recorded at or near the time of a recommendation click if the participant is already logged in or logs in while on the site.

Google Analytics reports 270 total recommendation clicks, while we have timestamped CloudFront log entries for only 232 clicks, suggesting that we are missing timestamps, participant, and site information for 14.1\% of clicks. Of the 232 CloudFront clicks, 198 correspond to unique participant/site clicks. We can recover 22 clicks missing from the CloudFront request logs via the logged-in user visits, for a total of 220 total unique participant/site clicks. For subsequent statistical analyses, we assume that any additional clicks are missing at random, although in practice we are more likely to be missing clicks from participants who do not log in. If we assume that the Google Analytics count is authoritative and 14.1\% of clicks are missing at random from the CloudFront data, then the expected number of missing participant/site clicks is 10. Thus, we expect only a small impact on our statistical analyses. Logged-in visit data are extracted from daily snapshots of the CaringBridge database. Thus, we only captured the most recent logged-in visit within a 24-hour period, sufficient to identify daily repeat visits but insufficient for fine-grained analysis of browsing behavior.

\section*{G Participant Perspectives}

\subsection*{G.1 Free-response: participant motivations}

“I would like to see how other CaringBridge authors articulate their reactions and feelings as they undergo the medical treatments for and rehabilitation from whatever medical conditions they experience.” Two free responses described motivations for not engaging with others’ sites, both describing it as a distractor during a busy time. “it’s hard to think about joining in someone else’s journey. I can focus on writing this blog discussing our journey because it is letting friends and family know what is happening so it is narrow enough that it doesn’t take away from the other things needed to be accomplished during the rest of the day.” These responses indicate that demand for peer connection is goal-driven and subject to constraints.

\subsection*{G.2 Free response: peer characteristics}

Three participants specifically identified age of the author or patient as a relevant detail. Two participants listed shared social context as important characteristics, such as already knowing the person or having “common friendships”. One participant said they sought Spanish-language updates, while another said they were looking for sites authored by healthcare professionals. One participant wanted to see “multiple posts all the way through death. I wanted to see what I would probably be writing as time progressed. A glimpse into the future if you will.”

\subsection*{G.3 Free response: explicit recommendation feedback}

\textbf{Negative feedback:} One participant requested the ability to filter sites “by the number of times Christ is mentioned”; these objections to Christian religious content may reflect belief misalignment between reader and author \cite{119}. Less severe objections focused on aspects of the writing: “boring”, “wordy”, “no reflection”, “too much philosophizing”, “bragging”. One participant desired sites that “go beyond simply updating the reader”, describing their own efforts as an author to inspire readers and to include “silver linings”.

\textsuperscript{36}We recommend that you use the logs to understand the nature of the requests for your content, not as a complete accounting of all requests. CloudFront delivers access logs on a best-effort basis. The log entry for a particular request might be delivered long after the request was actually processed and, in rare cases, a log entry might not be delivered at all. \url{https://web.archive.org/web/20211206215502/https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/AccessLogs.html}
Table 11. Prevalence of identified content categories in the Journal update previews. Only the presence of Expressive Writing was significantly associated with clicks (47.1% of first-batch recommendations with Expressive Writing were clicked vs 28.7% without, \( p = 0.017 \)).

| Preview category                        | First batch prevalence |
|-----------------------------------------|------------------------|
| Reporting health status                 | 85.2%                  |
| Neutral disclosures                     | 31.5%                  |
| Positive disclosures only               | 25.8%                  |
| Negative disclosures only               | 22.2%                  |
| Positive & negative disclosures         | 5.8%                   |
| Expressive Writing                      | 31.2%                  |
| Managing Author/Audience Relationship   | 17.5%                  |
| Expressions of Appreciation             | 5.8%                   |

**Positive feedback:** Two participants valued the recency of the displayed updates. “I’m finding myself following a lot of these sites. I enjoy the sites that update you often ... I feel like I’m abreast of what’s going on and take a personal interest into the person and their health battle.” Participants valued common ground—such as a patient being treated at the same hospital—and familiarity with issues faced. “I was more interested in the ones I was kind of familiar with. Understood more.” One patient gave us a summary of how they engaged with recommendations: “Even with all I’m going through ... I’ve come to care about these people who sites I am following. And I leave a comment every time I login on 95% of them because I know what it’s like when nobody comments. So I’m really really grateful you guys have [sent me site recommendations] because it is just helps me personally to take my mind off of things when I can go and pray for some other poor souls problems. So good job!”

G.4 “Thank you” email feedback

The non-clicking participant who replied said: “CaringBridge filled its purpose and functioned as expected. ... I received a lot of care and support from friends.” The three participants who clicked at least once emphasized the importance of common ground. One participant connected first with a site due to common ground—the patient and participant had previously lived in the same small suburb—and then kept following the site due to its use of poetry and song lyrics to process grief. One participant described reading others’ sites as “helpful and informative” due both to feeling comforted seeing a similar person navigate treatment and for learning strategies when writing their own Journal updates. One participant decided not to follow some of the recommended sites because they had a lot of existing followers. “Having a Caringbridge site I know what it’s like when you don’t have a lot of supporters. ... So I try to help by leaving comments 99% of the time on the sites I follow to try and help the caregivers stay positive and give them kudos for all they are handling.”

H Analysis of Journal update preview content

The primary information available to participants while they were deciding to click on a recommendation was the text preview of a recent Journal update. To capture the preview characteristics that our participants could see and respond to, we conducted a content analysis of these previews. Two researchers generated open and axial codes independently, then used an affinity mapping process based on the Grounded Theory method to identify three high-level themes [21]. To produce quantitative prevalence estimates, we integrated our three themes with taxonomic descriptions from prior work to identify a set of four categories [83, 120]. We conducted a regression analysis to identify which categories were associated with recommendation clicks.
Table 12. Summary of themes in recommended Journal update previews

| Theme                                      | Description                                         | Example(s)                                                                 |
|--------------------------------------------|-----------------------------------------------------|--------------------------------------------------------------------------|
| **Reporting patient/caregiver status**     |                                                     |                                                                          |
| Past vs Future News                        | An axis that differentiates past occurrence and future plans. | This last week was oh so busy with all of my tests and appointments. Friday is a huge milestone for me. |
| Events                                     | Those describing events, specifically related to a singular point in time. | Bryan met with his neurosurgeon yesterday.                                 |
| Patient Symptoms, Status, and Health Processes | Those detailing how the patient is doing, and health processes related to a prolonged interval of time. | Well, Sid is getting stronger by the day.                                 |
| Beginnings, Endings, and Transitions       | Those describing singular health events that transition the patient into or out of a health process. | Kristen is finally off the ventilator.                                   |
| Status sentiment                            | An axis that defines the sentiment of reported news (as “good” or “bad”). | Today Ashley had her second day of PT and she did amazing in every way. We waited all night for and unfortunately he’ll have to go through another round of chemo. |
| Emotion-laden Reporting                    | Those explicitly stating the author’s attitudes towards an update using emotive language. | Tears of joy as I write this. We are sooo happy to announce Andrea is out of the ICU. It’s hard to believe it was only a year ago today that Jen started chemo. |
| Reflection                                  | Those reflecting on the author’s or patient’s experiences. |                                                                          |
|                                           |                                                     |                                                                          |
| **Managing author/audience relationship**   |                                                     |                                                                          |
| Expressions of Gratitude                   | Those that positively acknowledge a specific type of support. | We are so thankful for your prayers.                                     |
| Update Context                             | Those that explicitly provide contextual information about the post or its contents for the audience. | Today is going to be a little different. I don’t have any medical updates, I just need to let some stuff out. |
| Comments on Update Frequency               | Those acknowledging the time between multiple CaringBridge updates. | First, I need to apologize for how long it’s been since our last update. I’ve really struggled to put my feelings into words in these posts. |
| Reflection on Writing Process              | Those reflecting on the author’s experiences as a Caring Bridge author. |                                                                          |

H.1 Recommendation preview themes

Our analysis of recommendation previews identified three high-level themes. We report the first-batch prevalence of the categories identified from these themes in Table 11.

**Disclosing health status, symptoms, and treatment.** Previews that address the question, “what or how is the patient doing?” Authors report on both recent health updates and future plans or expectations. “Friday is a huge milestone for me. My chemo port is going to be removed.” Disclosures vary from reporting on processes ("Sally is getting better day by day"), to discrete events ("Bryan met with his neurosurgeon yesterday."), to transitions ("Kristen is finally off the ventilator.").
Communicating emotion and reflection. Previews that express the author’s attitudes or reflect on author or patient experiences. Emotions are either linked to specific experiences during the health journey (“we are immensely happy to finally be home”) or characterize the author’s current mental state (“I wish you could experience this wonderful euphoria I am feeling”). Reflection is often used as the introduction to a narrative (“It’s been 12 weeks since Beth’s accident. I cannot believe it’s been that long…”).

Managing author/audience relationship. Previews that engage with the reader. Includes requests, expressions of gratitude, and context about the update, author, or writing process. Requests acknowledge the reader as an active member in the patient’s health journey. "Please send all your prayers as I start my journey through chemotherapy. I am nervous, but with your help I know I can get through this.” Expressions of gratitude acknowledge received support from readers. Previews that provide additional context make the author’s writing work visible to the reader e.g. by discussing update frequency. “Sorry it’s been a while since I last posted an update. I’ve been finding it hard to work up the energy to write, but I’ll do my best to recap the last few weeks.”

Beyond the three high-level themes identified in our content analysis, the full set of themes with descriptions and examples is shown in Table 12. From these themes, we identified four categories for quantitative analysis:

1. Reporting Health Status, Symptoms, and Treatment. Includes previews that update on specific health events, symptoms, or emotions of the patient. We include two sub-categories for positive and negative disclosures e.g. as used by Yang et al. [149]. Previews can contain both positive and negative disclosures if both are provided or if the author qualifies the news (“in horrible pain, but finally a reason for hope”).

2. Managing Author/Audience Relationship. The author is visible through descriptions of their role as a writer or their relationship with their blog’s readers. See thematic description above. “writing these posts is somewhat cleansing.”

3. Expressions of Appreciation (EOA). This theoretical construct was previously used by Smith et al., defined as explicit expressions of thanks, gratitude, blessing, or happiness that recognize the support received by the author or patient [120].

4. Expressive Writing and Reflection. We adopt the definition used by Ma et al.: Expressive Writing is “completed using functionalities afforded by online communities to disclose users’ thoughts and feelings about personal experiences” [83]. We adapt this definition to the context of Journal previews by including personal health-related reflection as a component of Expressive Writing.

Using these categories, two researchers separately annotated 658 update previews in three rounds of coding. The researchers met between each round to discuss disagreements and update the codebook. We report pre-discussion IRR scores in Table 13.

H.2 Quantitative content analysis methods

We used the resulting annotations to model the relationship between these categories and clicks. This problem is the “post-presentation reward prediction” problem [38]. Unfortunately, clicks are noisy and it is not possible to directly assess if the presence of a particular category is associated with a greater propensity to click. This impossibility results from confounding: the rank of the recommendation in question, the other recommendations in the same email, when the email was generated and opened, and the specific participant are all confounding factors. Instead, we build multiple models with varied sets of assumptions and look for convergence. Specifically, we assume that clicks are independent—a common assumption in click models [23]—and report analysis for three subsets of recommendations: (a) Clicked Batches Only includes only recommendations in
Table 13. Inter-rater reliability as Cohen’s $\kappa$ and percent agreement (%A) between the two annotators in the final two coding rounds. Post-discussion codebook updates resulted in greater agreement during round 3.

| Preview Category             | Round 2 ($n=143$) | Round 3 ($n=376$) |
|-----------------------------|-------------------|-------------------|
|                             | $\kappa$    | %A   | $\kappa$    | %A   |
| Reporting Health            | 0.56        | 83.9%| 0.62        | 87.8%|
| Positive Disclosures        | 0.72        | 87.4%| 0.78        | 89.4%|
| Negative Disclosures        | 0.61        | 88.1%| 0.76        | 92.8%|
| Managing Audience Relationship | 0.66      | 90.2%| 0.74        | 91.0%|
| Expression of Appreciation  | 0.78        | 96.5%| 0.90        | 98.1%|
| Expressive Writing          | 0.38        | 69.2%| 0.48        | 73.7%|
| All                         | 0.33        | 37.8%| 0.45        | 48.9%|

emails for which some but not all of the recommendations were clicked, so aims to acquire the clearest view of choice among competing options. (b) B1 Only includes only recommendations sent in the first batch of emails, eliminating temporal confounding and reducing bias introduced by varying levels of participant activity. (c) Clickers Only includes only recommendations sent to participants who clicked at least once during the study, reducing bias introduced by never-seen recommendations. We fit logistic regression models to predict individual recommendation clicks. While we report only recommendation-level logistic regression models here, email-level models, multi-level models for participant and batch, models including subsets of the features, and feature transforms revealed similar patterns.

H.3 Quantitative content analysis results

We summarize the relationship between all four categories and click behavior in Table 14. None of the models with any subset of the variables outperforms the null model (F-test $p > 0.05$)—even the rank-only model. Expressive Writing was associated with a greater probability of being clicked among Clicked Batches (50% of recommendations with Expressive Writing clicked vs 36% of recommendations without) and in B1 (47% of recommendations with Expressive Writing clicked vs 29% of recommendations without), but not when considering all recommendations sent to clicking participants. Expressions of Appreciation were associated with a decreased probability of being clicked among participants who clicked at least once (7% of recommendations with EOA clicked vs 12% of recommendations without) but not in the other subsets. Both of these results should be taken with a large grain of salt, and in general these results suggest that participants were making clicking decisions based on factors we did not annotate.

I Site Recommendation Analysis

In sec. H.1, we describe characteristics of the recommended set of sites. In this section, we provide additional pre-click details and use them to create a pseudo-control comparison set of non-recommended sites.

I.1 Site filtering

In sec. D.6 we discuss a decision limiting the total number of times a site was recommended in a single batch to at most 10 times. Specifically, we conducted a five-round draft. In each round, a random participant ordering is chosen and participants take turns selecting their highest-scored

37Recommendations listed first were most likely to be clicked, but this difference was statistically insignificant.
Table 14. Logistic regression models predicting recommendation clicks from preview content for the recommendation subsets defined in sec. H.2. Preview contents are poor predictors of clicks, as evidenced by near-chance ROC AUC scores (estimated using leave-one-out cross-validation). Note: *p<0.05; **p<0.01; ***p<0.001

|                           | Clicked Batches Only | B1 Only   | Clickers Only   |
|---------------------------|----------------------|-----------|-----------------|
| Intercept                 | -0.467               | -1.943*** | -1.714***       |
|                           | (0.325)              | (0.488)   | (0.206)         |
| Rank within email         | 0.096                | -0.023    | 0.034           |
|                           | (0.085)              | (0.115)   | (0.052)         |
| EOA                       | -0.581               | 0.156     | -0.546*         |
| 5.8% of B1 recommendations| (0.396)              | (0.621)   | (0.277)         |
| Expressive Writing        | 0.638*               | 0.762*    | 0.050           |
| 31.2% of B1 recommendations| (0.250)              | (0.353)   | (0.152)         |
| Managing Audience Relation| -0.293               | 0.239     | -0.358          |
| 17.5% of B1 recommendations| (0.328)              | (0.409)   | (0.192)         |
| Reporting health status (categorical) | | | |
| Base level: None (14.8% of B1) | | | |
| Neutral disclosure        | -0.171               | -0.235    | -0.013          |
| 31.5% of B1 recommendations| (0.338)              | (0.498)   | (0.211)         |
| Positive disclosure only  | -0.615               | 0.125     | -0.078          |
| 25.8% of B1 recommendations| (0.340)              | (0.494)   | (0.205)         |
| Negative disclosure only  | -0.395               | -0.540    | -0.091          |
| 22.2% of B1 recommendations| (0.396)              | (0.546)   | (0.247)         |
| Pos & neg disclosures     | -1.070               | -0.541    | -0.706          |
| 5.8% of B1 recommendations| (0.551)              | (0.768)   | (0.368)         |

Observations: 310 365 1,590
Clicks: 120 51 220
Log-Likelihood: -199.69 -142.94 -632.92
ROC AUC: 0.554 0.469 0.488

site that has been picked fewer than 10 times. We analyze the impact of this decision in Figure 15 by plotting the mean of the rank and score distributions for hypothetical recommendation sets generated without filtering. We observe that the maximum model score varies between batches as the model was retrained weekly, and that the filtered sites were similar in score to the sites that would have been recommended without filtering—all filtered sites were still in the top 0.1% of the model’s scores. The use of filtering resulted in a 32.5% increase in the number of unique recommended sites, a fair trade-off for the modest decrease in score and rank induced by filtering.

I.2 Pseudo-control set of non-recommended sites

The pseudo-control set is composed of the top five sites for each participant within a batch of recommendations that were never recommended during the duration of the study. We visualize this set’s ranks in Figure 15 along with the resulting effect on the corresponding score assigned to each set by the recommendation algorithm. In Table 15, we see that the set of recommended author/site pairs were significantly different from the group of sites that could have been recommended (pseudo-control). Values were calculated at the time a recommendation was clicked; in the case of
Fig. 15. Distribution rank and the corresponding scores assigned by the recommendation model for unfiltered (blue, solid), actually recommended (red, dashed), and pseudo-control (green, dotted) sets by weekly batch. Across all batches, the unfiltered group contains \(n=397\) unique sites, the actual contains \(n=526\) unique sites, and the pseudo-control contains \(n=511\) unique sites. Each line shows the mean rank and score respectively, while the shaded region indicates the max and min value within each set.

Table 15. Observed recommended pseudo-control USP site behavior from the 35 day window before the point they were clicked (or could have been clicked). Site tenure is in days, Total # of authors is since the sites creation, # of authors and # days visiting peers is for the entire window period, while all other rows are the number of weekly actions.

|                              | Recommended \((n_p=4190)\) | Pseudo-Control \((n_C=4190)\) | \(M_p - M_C\) | \(U_p / (n_p n_C)\) |
|------------------------------|-----------------------------|-------------------------------|----------------|-----------------------|
| Site tenure (days)           | 113 \(227.4 (519.2)\)      | 152 \(262.7 (425.2)\)        | -35.3*         | 42.4%*                |
| Journal updates              | 1 \(1.9 (2.0)\)            | 1 \(1.6 (2.3)\)              | 0.3*           | 40.4%*                |
| # of authors                 | 1 \(1.3 (0.7)\)            | 1 \(1.2 (0.5)\)              | 0.2*           | 44.9%*                |
| Total # of authors           | 2 \(1.8 (1.0)\)            | 2 \(1.7 (0.8)\)              | 0.2*           | 46.1%*                |
| Peer visits                  | 0 \(0.3 (1.2)\)            | 0 \(0.1 (0.2)\)              | 0.2*           | 26.6%*                |
| Repeat user visits           | 0 \(2.1 (6.3)\)            | 0 \(0.9 (1.5)\)              | 1.2*           | 49.5%*                |
| Peer initiations             | 0 \(0.8 (1.7)\)            | 0 \(0.4 (0.5)\)              | 0.4*           | 47.2%*                |
| Peer interactions            | 0 \(3.6 (8.0)\)            | 0 \(1.8 (3.9)\)              | 1.8*           | 47.3%*                |
| # days visiting peers        | 7 \(11.6 (11.7)\)          | 7 \(9.3 (9.0)\)              | 2.3*           | 47.4%*                |
| Site author interactions     | 0 \(0.0 (0.2)\)            | 0 \(0.0 (0.2)\)              | 0.0            | 50.0%                 |
| Site author initiations      | 0 \(0.0 (0.0)\)            | 0 \(0.0 (0.0)\)              | 0.0            | 50.0%                 |
| Site author self interactions| 6 \(15.1 (24.9)\)          | 3 \(8.4 (15.4)\)             | 6.8*           | 40.8%*                |

non-clicked and pseudo-control sites, a click time was randomly chosen from the set of actual click times from the same recommendation batch.

In Tables 16 and 17, we extend this analysis to include the subset of clicked recommended author/site pairs. Again, we see many significant differences between the set of clicked recommended author/site pairs and the pseudo-control set. However, we find little significant differences between clicked and non-clicked recommended author/site pairs.
Fig. 16. Estimated impacts of receiving Site Suggestion emails on author behavior, both during the study (left column, the vertical dotted line indicates the date of the last Site Suggestion email) and after the study (right column). The estimators shown are as described in sec. 5.3.3: raw (solid, blue), OLS (dashed, orange), and DR (dotted, green). Figure 8 captures the vertical slice at post-study week 13. 95% confidence intervals are computed via bootstrapping (1000 iterations).
Table 16. Observed clicked and non-clicked recommended site behavior from the 35 day window before the point they were clicked (or could have been clicked). Site tenure is in days, Total # of authors is since the site’s creation, # of authors and # days visiting peers is for the entire window period, while all other rows are the number of weekly actions.

|                     | Clicked \((n_p=220)\) | Non-Clicked \((n_C=3970)\) | \(M_p - M_C\) | \(U_p / (n_P n_C)\) |
|---------------------|------------------------|-----------------------------|----------------|----------------------|
| Site tenure (days)  | 107 267.4 (609.1)      | 113 225.2 (513.8)           | 42.2           | 49.5%                |
| Journal updates     | 1 2.1 (2.2)            | 1 1.9 (2.0)                 | 0.2            | 46.6%                |
| # of authors        | 1 1.4 (0.7)            | 1 1.3 (0.7)                 | 0.1            | 45.7%*               |
| Total # of authors  | 2 1.9 (1.0)            | 2 1.8 (1.0)                 | 0.1            | 47.4%                |
| Peer visits         | 0 0.5 (1.7)            | 0 0.3 (1.1)                 | 0.2            | 43.9%*               |
| Repeat user visits  | 1 3.1 (8.1)            | 0 2.0 (6.2)                 | 1.1            | 43.6%*               |
| Peer initiations    | 0 1.1 (2.1)            | 0 0.8 (1.7)                 | 0.3            | 45.2%                |
| Peer interactions   | 0 4.6 (9.5)            | 0 3.6 (7.9)                 | 1.0            | 45.7%                |
| # days visiting peers | 11 13.0 (11.7)       | 7 11.5 (11.7)               | 1.5            | 45.0%                |
| Site author interactions | 0 0.1 (0.4)       | 0 0.0 (0.2)                 | 0.0            | 49.2%                |
| Site author initiations | 0 0.0 (0.1)       | 0 0.0 (0.0)                 | 0.0            | 49.2%                |
| Site author self interactions | 5 14.0 (25.8) | 6 15.2 (24.9) | -1.2 | 48.0%                |

Table 17. Observed clicked and pseudo-control site behavior from the 35 day window before the point they were clicked (or could have been clicked). Site tenure is in days, Total # of authors is since the site’s creation, # of authors and # days visiting peers is for the entire window period, while all other rows are the number of weekly actions.

|                     | Clicked \((n_p=220)\) | Pseudo-Control \((n_C=4190)\) | \(M_p - M_C\) | \(U_p / (n_P n_C)\) |
|---------------------|------------------------|-----------------------------|----------------|----------------------|
| Site tenure (days)  | 107 267.4 (609.1)      | 152 262.7 (425.2)           | 4.7            | 43.0%*               |
| Journal updates     | 1 2.1 (2.2)            | 1 1.6 (2.3)                 | 0.5*           | 37.3%*               |
| # of authors        | 1 1.4 (0.7)            | 1 1.2 (0.7)                 | 0.2*           | 40.7%*               |
| Total # of authors  | 2 1.9 (1.0)            | 2 1.7 (0.8)                 | 0.2*           | 43.4%*               |
| Peer visits         | 0 0.5 (1.7)            | 0 0.1 (0.2)                 | 0.4*           | 23.1%*               |
| Repeat user visits  | 1 3.1 (8.1)            | 0 0.9 (1.5)                 | 2.3*           | 43.4%*               |
| Peer initiations    | 0 1.1 (2.1)            | 0 0.4 (0.5)                 | 0.6*           | 47.2%                |
| Peer interactions   | 0 4.6 (9.5)            | 0 1.8 (3.9)                 | 2.8*           | 47.5%                |
| # days visiting peers | 11 13.0 (11.7)       | 7 9.3 (9.0)                 | 3.7*           | 42.3%*               |
| Site author interactions | 0 0.0 (0.2)       | 0 0.0 (0.1)                 | 0.0            | 49.8%                |
| Site author initiations | 0 0.0 (0.0)       | 0 0.0 (0.0)                 | 0.0            | 49.8%                |
| Site author self interactions | 5 14.0 (25.8) | 3 8.4 (15.4) | 5.6* | 42.5%*               |

J Estimates of study impact on behavior for authors and sites

In sec. 5.3.3, we estimated the impact of recommendation on second-order behaviors. We estimate this impact by comparison to a pseudo-control group of non-enrolled authors and non-visited
Table 18. Activity variables included in the participant author model and the visited site model. We removed some variables to avoid collinearity, but otherwise opted to incorporate as many potentially relevant behaviors as possible [144].

| Author model variables                                      | Site model variables                              |
|-------------------------------------------------------------|---------------------------------------------------|
| # Journal updates                                           | # Journal updates                                 |
| # first visits to other authors’ sites                       | # unique author visits (recent)                    |
| # repeat visits to other authors’ sites                      | # unique authors (all time)                        |
| # unique days visiting other authors’ sites                  | # first visits to site from peers                  |
| # interactions on other authors’ sites                       | # repeat author visitors to site                   |
| # interactions on their own sites                            | # unique days other authors visited                |
| # self-authored sites interacted with                        | # interactions from other authors                  |
| Author tenure (log)                                          | # initiations to site                              |
|                                                             | # peer interactions by site authors                |
|                                                             | # self-site interactions by site authors           |
|                                                             | # initiations by site authors                      |
|                                                             | Site tenure (log)                                  |

sites (see section 4.1.2 and Appendix I.2 respectively). We provide additional method details and sensitivity analysis in this section.

In Figures 8 and 9, we showed post-study outcomes by comparing 5 weeks (35 days) of pre-study behavior to 12 weeks (91 days) of post-study behavior. We conducted a sensitivity analysis to determine the impact of two decisions: examining behavior during the post-study period (rather than the “during study” period) and choosing a time window of 12 weeks. Figure 16 is a high-detail summary of those decisions, demonstrating that differences are small depending on the selected time window: smaller post-study time windows are generally associated with higher-variance effect size estimates. In pre/post panel data analyses, it is common to use the same time window pre- and post-intervention, which helps to control for longer-term behavioral trends. We tried both approaches, and found no difference except increased variance, so we use data from the full pre-study time window (5 weeks) even if the post-study time window is less than 5 weeks.

Using a similar approach, we extend this analysis to investigate post-click outcomes on recommended site behavior. In Figure 9, we showed post-study outcomes by comparing 5 weeks (35 days) of pre-click behavior to 12 weeks (91 days) of post-click behavior. Similar to Appendix I, values were calculated at the time a site was first clicked and in the case of non-clicked and pseudo-control sites, a click time was randomly chosen from the set of actual click times in the first batch that site was/could of be recommended. We conducted a sensitivity analysis to examine the impact a participant visiting a recommended site had on the sites behavior during the post-click period. Figure 17 is a high-detail summary demonstrating we were unable to capture statistically significant effects from participant clicks on recommended site behavior. Moreover, we find that effect estimates based on comparison to the non-clicked and pseudo-control groups are similar.

As noted by Hernán and Robins, successful causal inference predominantly rests on untestable assumptions [51]. The doubly robust estimates we produce depend on five assumptions:

- **Exchangeability** Exchangeability refers to the probability of being in the treatment group is independent of the causal outcome (depending only on $A$). Exchangeability is likely false: it is reasonable to assume that authors that opted to participate in a recommendation intervention are more likely to engage with a recommendation intervention than non-participating
Fig. 17. Estimated impacts of Site Suggestion emails on recommended site behavior post click, both using non-click sites (left column) and pseudo-control sites (right column). The estimators shown are as described in sec. 5.3.3: raw (solid, blue), OLS (dashed, orange), and DR (dotted, green). Figure 9 captures the vertical slice at week 13 in the right column. 95% confidence intervals are computed via bootstrapping (1000 iterations).
authors, independent of e.g. their tenure as an author or their prior peer interaction behavior. If participants are not very different from non-participating authors in terms of their responsiveness to treatment, then it may still be reasonable to assume exchangeability. The same reasoning holds for clicked and non-clicked sites.

- **Positivity** Positivity requires that the distributions of the covariates for the treated and pseudo-control groups to fully overlap. Positivity is approximately true in our data, i.e. all values of the covariates observed in the treated group are also observed in the pseudo-control group. The positivity assumption is why we omit causal estimates of the impact of deploying Site Suggestion emails to the whole author population in terms of e.g. click rates; we have no ability to estimate the behavior of the pseudo-control population when exposed to Site Suggestion emails.

- **Consistency** Consistency refers approximately to the assumption that the treatment varies only in ways unrelated to the outcomes. Recommendation varies, so participants will be exposed to different versions of the treatment; different participants are shown different recommendations, so clicked sites will be exposed to visits from different participants. Intuitively, these differences can be causally related to the outcomes: a participant with no prior interaction behavior, for example, may receive less relevant recommendations than a participant with a long history of peer interaction. A site clicked by a participant that visits and leaves a comment is different than a site clicked by a participant that only visits. Joachims et al. argue that recommender systems ought to be viewed as policies that select interventions in order to optimize a desired outcome [65]; it’s a policy designed to select variable interventions. By assuming consistency, we assume that the recommendations are of similar quality and that participant visits have similar effects on the visited sites. The degree to which this assumption is violated will bias the resulting estimates.

- **No measurement error** Measurement error in the observed behaviors is likely non-existent, given the use of a complete database snapshot.

- **No model misspecification** To adjust for confounding, the models we use need to include all relevant covariates and assume the correct functional form. By fitting linear models, we make a parsimonious but possibly false assumption of functional relationship between the covariates and the outcome of interest. More serious is the assumption that all relevant covariates are available; as discussed, we believe that unmeasured confounders such as an interest in peer connection affect response to treatment in a way that is independent of the measured covariates.

Given this discussion, it seems reasonable to object that three of the assumptions are very likely false. In practice, we are assuming that these assumptions are close enough to true, such that we can still attain some insight from computing causal estimates based on these assumptions in order to compare to the observational estimates.

### K Sample Size Calculations

In sec. 5.3.4, we described a power analysis for an uncontrolled peer recommendation study. In this section, we provide additional details and add estimates for an RCT. In Fig. 10, we showed sample size estimates for two variations of a replication of our feasibility study by computing the effect of two author behaviors associated with receiving recommendations. Here, we extend this analysis to include 5 behaviors: unique repeat visits to recommended sites, interactions with recommended sites, unique repeat visits to all sites, interactions with all sites, and recommended site updates. We consider the 35 days before and after the first exposure to recommendation or first stranger visit
for the one-time email and 35 days before and the study period plus 35 days after for the 12 week recurring email intervention.

Table 19. Mean and standard deviation estimates for a one-time and recurring recommendation email intervention. Site descriptions in parenthesis for participant groups denote the target of the intended action.

| Behavior          | Recurring 12-Week | One-Time |
|-------------------|-------------------|----------|
|                   | $M_P$ ($\sigma^2$) | $M_C$ ($\sigma^2$) | $M_P$ ($\sigma^2$) | $M_C$ ($\sigma^2$) |
| Participants      |                   |          |                   |                      |
| (Recommended Sites) | Study ($n_p=79$) | Study ($n_p=73$) |          |                      |
| Second visits     | 0.06 (0.06)       | 0.04 (0.02) | -0.05 (0.03)      | 0.02 (0.01)          |
| Interactions      | 0.47 (16.36)      | 0.07 (0.22) | 0.04 (0.72)       | 0.09 (0.49)          |
| Participants      |                   |          |                   |                      |
| (All Sites)       | Participants ($n_p=79$) | Pseudo-control ($n_C=1759$) | Participants ($n_p=73$) | Pseudo-control ($n_C=1759$) |
| Second visits     | -0.07 (0.07)      | 0.02 (0.01) | -0.05 (0.03)      | 0.02 (0.01)          |
| Interactions      | -0.34 (14.81)     | 0.11 (0.49) | -0.04 (0.72)      | 0.09 (0.49)          |
| Recommended Sites |                   |          |                   |                      |
| Updates           | 0.024 (2.11)      | 0.023 (1.80) | 0.02 (1.33)       | 0.04 (3.45)          |

Table 20. Observed effect sizes and required sample size needed for an appropriately powered RCT for a one-time recommendation email and a recurring, weekly recommendation email. Site descriptions in parenthesis for participant groups denote the target of the intended action.

| Behavior          | Recurring 12 Week | One-Time |
|-------------------|-------------------|----------|
|                   | Effect Size       | Sample Size | Effect Size | Sample Size |
| Participants (Rec) |                   |          |                   |                      |
| Second visits     | 0.28              | 75       | 0.26             | 81                |
| Interactions      | 0.12              | 444      | 0.14             | 314               |
| Participants (All) |                   |          |                   |                      |
| Second visits     | 0.85              | 5        | 0.67             | 10                |
| Interactions      | 0.43              | 29       | 0.18             | 189               |
| Recommended Sites |                   |          |                   |                      |
| Updates           | -0.14             | 299      | 0.23             | 110               |

Table 19 outlines the mean, variance, and sample size for each group in terms of weekly actions used in the calculations. The structure of our data let us consider two potential future interventions: a one-time recommendation email and a recurring, weekly recommendation email. We estimate the effects of a one-time recommendation email by including only participant exposures and visits from the first batch of recommendation emails (B1). We use the pseudo-control group (see sec. 4.1.2 and Table 2) and non-clicked recommendations (see Appendix I) to analyze the effect of receiving...
recommendations and stranger visits. For participants (recommended sites) behaviors, we report weekly actions from a 5 week pre-inflection period to a post-inflection period: 12 weeks + 5 weeks and 5 weeks. For participants (all sites) and recommended sites behaviors, we report the difference in weekly actions from similar pre- and post-inflection periods.

Using these statistics, Table 20 estimates the sample size required for a replication (same as Fig. 10 and for an RCT, with 50% not receiving recommendation emails. For author behaviors targeted at recommended sites, we compute the effect as simply as $d = M_P / SD$. For all other behaviors, we control for prior behavior by computing the effect as $d = (M_C^{before} - M_C^{after}) - (M_P^{before} - M_P^{after}) / SD_{pooled}$, where P is the group of participants or clicked recommended sites and C is the control group of pseudo-control participants or non-clicked recommended sites. We present $M^{before} - M^{after}$ because the expected churn of users dictates over time they will become less active. We find this is true in our sample; all users that published at least one journal update in July 2021 subsequently published, on average, 1.32 (5.44) less updates in the following month. Here, a positive effect size means that being a participant results in less of a difference from before and after when compared to the control group. Using G*Power 3.1.9.7, the estimated required sample sizes for an appropriately powered RCT were calculated based on observed effect sizes using a one tailed point biserial model at 80% power with $\alpha = 0.5$ [32].

Unsurprisingly, we see large differences in participant behavior towards recommended sites when compared to the pre-study period. More interesting is the fact that participants, on average, sustained an increase in weekly actions including non-recommended sites after receiving recommendations. While this translates to a seemingly large effect size during calculation, we attribute these differences primarily to the significant differences pre-study between both groups outlined in Table 4.1.2. Here, we believe any differences that exist between groups can largely be attributed to confounding variables (i.e. author/site tenure).

Received January 2024; revised July 2024; accepted October 2024