Exploring Trade-offs for Online Mental Health Matching: An Agent-Based Modeling Study

Anna Fang, Yuhan Liu, Glen Moriarty, Cris Firman, Robert E. Kraut, Haiyi Zhu

Submitted to: JMIR Formative Research
on: March 11, 2024

Table of Contents

Exploring Trade-offs for Online Mental Health Matching: An Agent-Based Modeling Study

Anna Fang[1*]; Yuhan Liu[2*]; Glen Moriarty[3]; Cris Firman[3]; Robert E. Kraut[1]; Haiyi Zhu[1]

[1]Human-Computer Interaction Institute Carnegie Mellon University Pittsburgh US
[2]Princeton University Princeton US
[3]7 Cups Wilmington US
[*]these authors contributed equally

Corresponding Author:
Anna Fang
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh
US

Abstract

Background: Online mental health communities (OMHCs) are an effective and accessible channel to give and receive social support for individuals with mental and emotional issues. However, a key challenge on these platforms is finding suitable partners to interact with given that mechanisms to match users are currently underdeveloped or highly naive.

Objective: In this paper, we collaborate with one of the world's largest OMHCs to contribute the application of agent-based modeling for the design of online community matching algorithms. We develop an agent-based simulation framework and explore how it can uncover trade-offs in different matching algorithms between people seeking support and volunteer counselors.

Methods: We use a dataset spanning January 2020 to April 2022 to create a simulation framework based on agent-based modeling that replicates the current matching mechanisms of our research site. After validating the accuracy of this simulated replication, we use this simulation framework as a “sandbox” to test different matching algorithms based on the deferred-acceptance algorithm. We compare and contrast trade-offs among these different matching algorithms based on various metrics of interest such as chat ratings and matching success rates.

Results: Our study contributes the novel application of agent-based simulation to matching in online health communities, and our simulation findings suggest that various tensions and goals emerge through different algorithmic choices for these communities. For example, we found that higher chat ratings and lower blocking frequency occurs with matching people using just topic(s) of interest for discussion, compared to matching based on just demographics or first-come-first-serve methods. We also found some trade-offs in hard filter-based approaches to prioritize the protection of marginalized groups, and that other algorithms can actually improve the experience of both minority and majority groups.

Conclusions: Agent-based modeling can reveal significant design considerations in the OMHC context, including trade-offs in various outcome metrics and the potential benefits of algorithmic matching on marginalized communities.

(JMIR Preprints 11/03/2024:58241)
DOI: https://doi.org/10.2196/preprints.58241

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Exploring Trade-offs for Online Mental Health Matching: An Agent-Based Modeling Study

Abstract

Background: Online mental health communities (OMHCs) are an effective and accessible channel to give and receive social support for individuals with mental and emotional issues. However, a key challenge on these platforms is finding suitable partners to interact with given that mechanisms to match users are currently underdeveloped or highly naive.

Objective: In this paper, we collaborate with one of the world's largest OMHCs to contribute the application of agent-based modeling for the design of online community matching algorithms. We develop an agent-based simulation framework and showcase how it can uncover trade-offs in different matching algorithms between people seeking support and volunteer counselors.

Methods: Using a comprehensive dataset spanning January 2020 to April 2022 to create a simulation framework based on agent-based modeling that replicates the current matching mechanisms of our research site. After validating the accuracy of this simulated replication, we use this simulation framework as a “sandbox” to test different matching algorithms based on the deferred-acceptance algorithm. We compare and contrast trade-offs among these different matching algorithms based on various metrics of interest such as chat ratings and matching success rates.

Results: Our study suggests that various tensions emerge through different algorithmic choices for these communities. For example, our simulation uncovered the inherent consequence of increasing the waiting time for support-seekers on these sites when using intelligent matching to find more suitable matches. Our simulation also verified some intuitive effects, such as the greatest number of support-seeker/counselor matches occurred with a First-Come-First-Serve protocol while relatively fewer matches occurred with Last-Come-First-Serve. We also discuss practical findings around matching for vulnerable versus overall populations. Results by demographic group revealed significant disparities: under-aged and gender minorities experienced lower chat ratings and higher blocking rates on the site when compared to their majority counterparts, indicating the potential benefits of algorithmically matching them. We found that some protocols, such as a “filter”-based approach that matched vulnerable support-seekers only with a counselor of their same demographic, led to improvements for these groups but resulted in lower satisfaction (-12%) among the overall population. However, this trade-off between minority and majority groups was not seen when using topic as a matching criterion. Topic-based matching actually outperformed the filter based protocol among under-aged people, and led to significant improvements over the status quo among all minority and majority groups – specifically, a 6% chat rating improvement and a decrease in blocking incidents from 5.86% to 4.26%.

Conclusions: Agent-based modeling can reveal significant design considerations in the OMHC context, including trade-offs in various outcome metrics and the potential benefits of algorithmic matching on marginalized communities.

Keywords: agent-based modeling; mental health; algorithmic matching; social computing; online communities

Introduction

Background

People are increasingly turning to online mental health communities (OMHCs) for mental and emotional support [44,47]. OMHCs are a practical and accessible way for users to receive both informational and emotional support on a variety of mental and emotional concerns [35], with communities offering general support for any need or support for specified health issues. For example, communities such as 7Cups.com provide general 1-on-1 peer counseling chats while platforms like BabyCenter.com provide targeted support resources for pregnant women [18,22]. OMHCs has been found to be vital in maintaining and improving people's well-being, such as reducing depression, fostering meaningful relationships, and increasing trust in mental health treatment.

However, despite the ability of OMHCs to yield meaningful and positive relationships between users, they currently rely on naive methods (i.e. first-come-first-serve, solely based on topic of discussion) for members to find these relationships, without consideration of users' unique characteristics and preferences. Prior work suggests that current matching systems do not adequately support users' needs and capabilities; ineffective matching can also lead to fewer long-term relationships and reduced member commitment [18]. Moreover, this lack of purposeful matching methods may be particularly harmful for marginalized communities, who have both strong preferences for mental health providers with a similar background and particular reliance on online communities for support [8,11,21,28,50]**.** Given these challenges, intelligent forms of matching can provide more optimal matches with minimal efforts on a user's end [18].

However, matching is a complicated mechanism design problem [2,42]. It is challenging to meet all the possible matching goals between support seekers and providers, and prioritizing one goal might lead to worse outcomes in other goals. Given these challenges, tools such as agent-based simulation that have long been used to apply social science theories to the design of HCI systems are useful for revealing the complexities and various trade-offs in matching protocols for online community designers [40]. Importantly, running these low-cost, virtual experiments can predict community members' likely reaction to alternative design choices without disrupting existing community dynamics. Agent-based modeling thus enables researchers and community designers to pin down factors leading to desirable outcomes and understand how design choices affect behavioral outcomes by modeling the intervening processes [29].

Goal of This Study

In this paper, we answer the research question: **how can we experiment with new matching algorithms for online mental health communities?** Specifically, we seek to experiment with these new algorithms without harming or disrupting the existing community. To do this, we created a simulated "sandbox" based on agent-based modeling that allows community stakeholders to play with different matching algorithms and helps designers consider complex trade-offs in building new mechanisms for their community. In order to build this simulation based on a real OMHC, we collaborated with one of the world's largest peer support platforms. We used the platform's dataset to accurately replicate the platform in our simulation, then experiment with alternative protocols in the simulation to analyze how these new matching policies affect users' experiences. We created and tested seven new algorithmic matching policies based on prior work studying OMHC users' needs: (1) ***first-come-first-serve***, (2) ***last-come-first-serve***, (3) ***similarity-based*** that uses cosine similarity to prioritize support-seekers and support-providers of similar features/preferences, (4) ***gender-based matching***, (5) ***age-based matching***, (6) ***topic-based matching***, and lastly (7) ***filter-based***, which

focuses on protecting teenagers and gender minorities.

Our work contributes the application and example use of agent-based simulation to uncover the effects of alternative policies in the design of online community matching. Exploring different matching protocols has the potential to disrupt the particularly sensitive population of OMHC users through implementation and iteration processes; given this, our work showcases the benefits of instead using agent-based simulation to reveal the impacts of various algorithms. In addition to applying simulation to the OMHC context, we contribute practical findings for designing matching in OMHCs, such as how optimizing based on topic can improve chat experiences for vulnerable communities as well as trade-offs like how using algorithmic matching can increase the quality of conversations but also the waiting time for support-seekers.

Relevant Prior Work

Online Mental Health Support

Social support through online platforms has been shown to improve users' well-being in numerous ways, such as reducing depression, lowering suicidal ideation, spreading information about mental health, and enabling help-seeking for stigmatized populations [15,35,38]. The majority of online mental health support takes place in OMHCs where peers can speak anonymously about their experiences for free and 24/7, which is essential for groups who particularly struggle with stigma and access to resources [37] like adolescents and LGBTQ+ populations [5,10,14,20,34]. However, some groups, such as females and gender minorities, also face unique challenges of sexual and verbal harassment while chatting on OMHC platforms [18]. Thus, special consideration for protection of vulnerable groups is crucial in building better matching mechanisms, given the unique benefits and challenges they face. Our study shows how agent-based simulation can help community designers build and test ways for people, including minors and gender minorities who have specific needs, to find relevant and useful partnerships when engaging in online mental health support.

Matching for Mental Health Purposes

As people perceive those similar to themselves to be more trustworthy and likely to share their worldviews, research has thoroughly supported that a client and therapists' race, language, gender, and other variables impact therapeutic outcomes [19,24,31,45]. In traditional mental health resources, as clients have strong preferences for choosing a therapist of the same race [13] and their same gender [19]. Effects of racial matching may−be especially important for minorities such as Black clients given mitigation of general mistrust towards mental health services [33,49]. Topic and content−have been found to be important in care for client satisfaction and therapy quality [16,23,27,32]. In terms of the online context, our work builds upon prior literature by Fang and Zhu that found gender, age, and experience level are all significant factors in people's preferences for online support relationships; in particular, gender minorities being matched with those of similar gender identity and avoiding support-providers who were significantly younger resulted in support-seekers having a more positive experience [18]. In fact, users consistently share their gender and age with one another when trying to find online support relationships thus further showing the need for and lack of more efficient forms for matching. We note that, apart from Fang and Zhu, most of the prior work has studied important matching features in the traditional therapy context. Our work builds upon this prior knowledge by showcasing evaluation on community-level outcomes when matching on important features (age, gender, topic) in the *online* context specifically.

Agent-based Modeling and Online Community Design

Agent-based modeling is the simulation of the actions of agents to understand their behaviors and interactions under different conditions, and has been useful for informing the design of online communities through simulating how different design choices impact desired outcomes, and has been used in past work to explore topics such as social influence and information propagation [4,36,40]. One other primary benefit of agent-based modeling, as opposed to direct experimentation, is that effects can be observed over long periods of time as one can run the agent-based model repeatedly and for lengthy time cycles; thus, downstream and even unintended effects (rather than just first-order effects) can be identified [29]. Importantly, agent-based modeling serves as a testing ground for community designers and researchers to surface how different theories affect the community broadly but also allow them to isolate the factors leading to particular outcomes [29]. For example, Ren and Kraut have shown how agent-based modeling can be used to apply social science theories and understand trade-offs in design decisions; they applied these methods to explore motivations for online community participation and how different moderation methods affect discussion in online communities [39,40]. Given the proven power of agent-based modeling for understanding community dynamics in online communities, we apply agent-based modeling to the online mental health context to showcase its usefulness for exploring matching algorithms and understanding trade-offs in these design decisions.

Methods

Research Site

In order to study algorithmic matching and simulation in the real online mental health context, we collaborated with one of the largest existing support platforms that is currently an active and growing community. This online platform provides free 24/7 chat support and has over 54 million members and 500,000 trained volunteer counselors. Users who sign up to seek support - who we will call “support-seekers” - can chat in 1-on-1 chat rooms with trained “volunteer counselors”. Volunteer counselors complete a roughly one-hour, psychology-based training that is based on active listening and MI skills. Counselors can also receive awards on their profiles from completing additional, optional training modules such as specialized courses for specific conditions (e.g. ADHD, Depression) as well as advanced general skill courses (e.g. “Active Listening”, “Managing Emotions”). Note that although volunteer counselors have some training on the site, we refer to our study’s platform as a peer support platform given that all counselors are non-professional, only lightly trained, and members can be both support-seekers and counselors.

All users are required to provide their age to the platform, while other demographic information (e.g. gender) is optional. The primary support method on the research site is through 1-on-1 chats between one support-seeker and one volunteer counselor. The current matching process is a self-selection by volunteer counselors, where support-seekers send a request to join a live queue and wait to be picked by a volunteer counselor to begin a chat. Support-seekers also have the option to select among “topic tags” (e.g. ‘depression’, ‘relationship stress’) but are not required to do so. No other information about support-seekers besides their wait time and possibly their topic tag is displayed in the queue. We will later use these features like people’s demographic information and topic choices in our study’s matching protocols. Support-seekers can cancel their chat request at any point and may do so especially if waiting too long.

Data

The dataset consists of all chat messages between January 2020 to April 2022, which includes 8

million chats with over 1.5 million support-seekers and over 288,000 volunteer counselors. All chat data includes the anonymized message text, timestamp, and user IDs involved. The dataset also includes users' signup dates and birth years. Note that no personally identifiable information about users is available. Relevant to our study, chats can also be rated by support-seekers from 1 to 5 stars once the chat has continued for a certain length of time or after the chat ends; volunteer counselors are not able to rate a chat. Additionally, users can block one another at any point including during their conversation.

Ethics & Privacy Considerations

This paper used behavioral log data obtained through a collaboration with the online mental health community studied in this paper to conduct our analysis, and data collection followed HIPAA and confidentiality agreements. All users who register on the site are informed of and accept their anonymized data for research use. All data was anonymized before analysis and no personally identifiable information was used in this study. Note that chat messages were only analyzed to find the gender distribution of users in order to generate agents for simulation purposes, which is described in Section Building Agent-Based Simulation: Assumptions. One author of this paper worked at a 3-month internship for this study's research site. This work has been approved by the appropriate Institutional Review Board (IRB) and no users of the online mental health community were directly interacted with for this study; thus, no additional consent or compensation to users was needed for this research.

Summary of Methods

In order to showcase how agent-based modeling can be used to experiment with matching policies in the online mental health context, we followed three stages as outlined in the below diagram Figure 1.

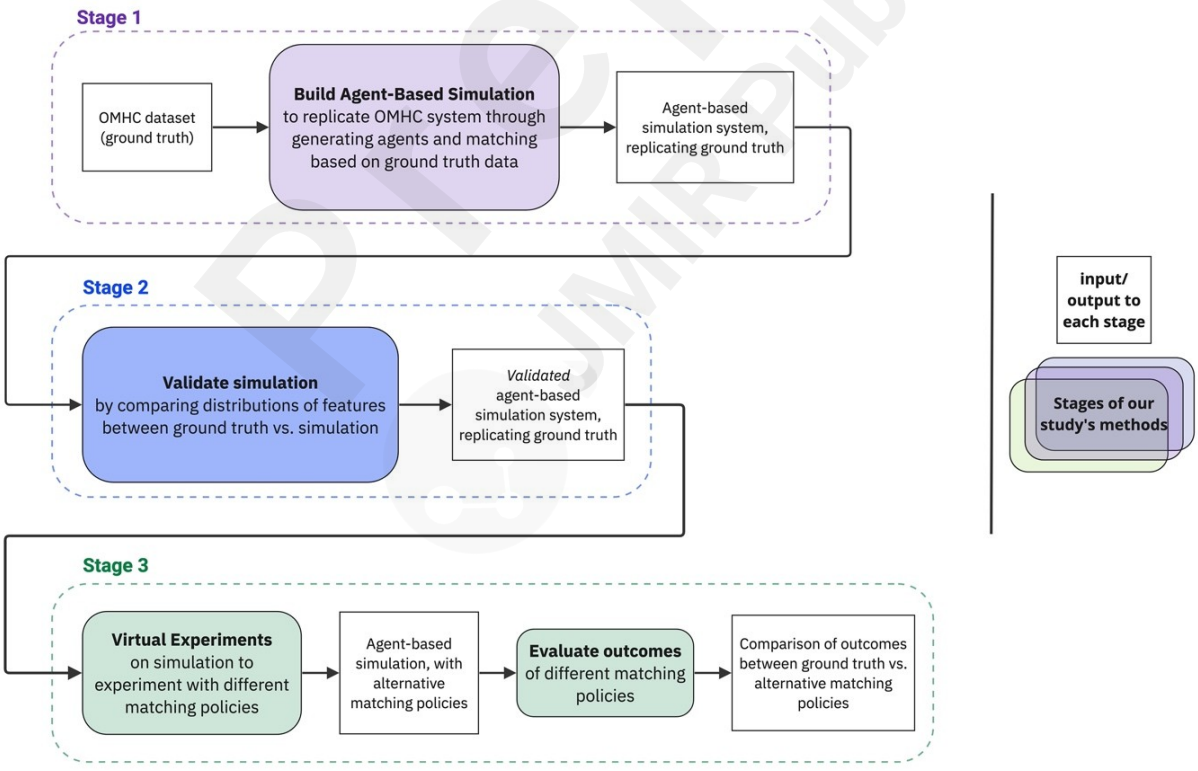


Figure 1. Workflow diagram showing inputs and outputs for the three stages of building and experimenting with agent-based simulation: (1) Building Agent-based Simulation, (2) Validation of Agent-Based Simulation, and (3) Virtual Experiments using the validated simulation.

Building Agent-Based Simulation: Assumptions

We first built a simulation based on agent-based modeling in order to replicate the current matching mechanisms of our study's research site, and utilized outcome prediction models to validate this replication. Below we outline the assumptions made in order to build our simulation, all of which were implemented based on the real OMHC dataset. Overall, we found that there was extremely high correlation (i.e. Pearson coefficient) between our simulation and the real research site. Since all had high correlation, below we show figures for only a few features as visuals of how our simulation performed compared to the research site.

Simulation Period

In our agent-based simulation, both support-seekers and volunteer counselors have the possibility of being matched during each “round”, or simulation period. We determined that a simulation period of one minute was fit for our model, as one minute is temporally granular enough to yield quick matching of users and derivable from our empirical data.

Generation of Agents

Our simulation consisted of two types of agents: support-seekers and volunteer counselors. Each simulation period (i.e. each minute) generates new support-seeker and volunteer counselor agents that are eligible for matching. Agents are considered “online” (i.e. available to chat) immediately when they are generated. In order to determine the number of agents generated, we analyzed our study’s dataset over January 2020 to April 2022 to find the average number of online support-seekers and volunteer counselors for each simulation period over a week (i.e. 10,080 minutes) (see Figure 2).

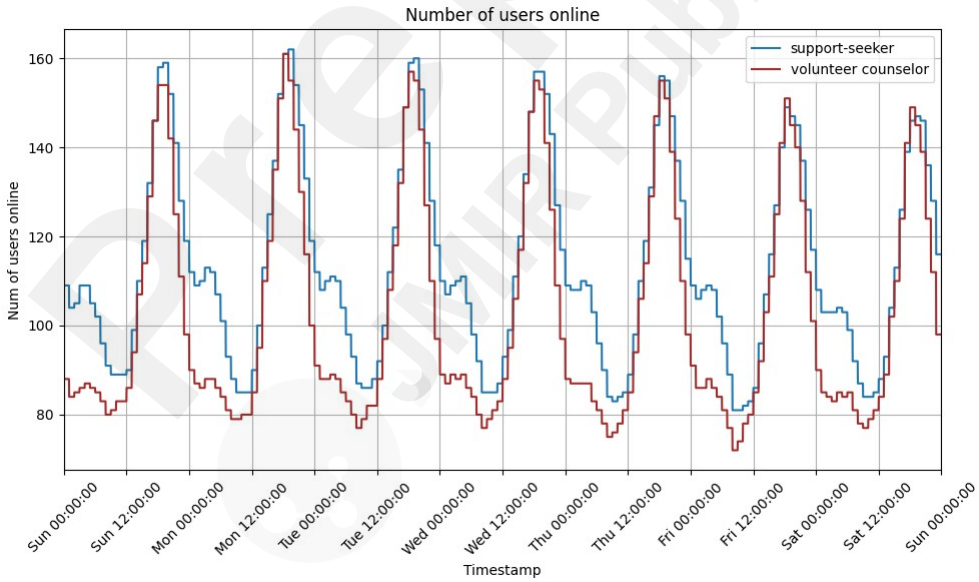


Figure 2. A line chart showing the number of online support-seekers and volunteer counselors in each simulation period. The number of support-seekers always exceeds the number of volunteer counselors, with support-seekers online at any given minute ranging between 81 and 162 (mean of 113.26, standard deviation of 22.56) and volunteer counselors online at any given minute ranging between 72 and 161 (mean of 102.49 and standard deviation of 25.07).

When generated, each agent is also given several personal characteristics that may be significant to matching (gender, birth year, topic of interest)[18]. We analyzed the OMHC dataset to find the

distribution of these characteristics at each simulation period, and assign agents' characteristics so that the simulation's distribution is identical to the distribution found from the real OMHC data. Birth year is part of the raw dataset, giving us complete and accurate real data to draw from. However, we had to conduct a labeling process for users' gender identity since a minority of users input their gender on their profile manually. Following prior work, we labeled gender according to whether a user had self-identified their own gender in chat logs (i.e. "I am a female", "I am non-binary"), which had been found to be highly accurate and only mislabel gender for 0.8% of OMHC users [18]. Using this process, we were able to label the gender of 35% of support-seekers and 50% of volunteer counselors in our dataset. We then applied the distribution of gender among those whose gender is known to our generated agents so that all agents have a gender characteristic. This distribution is shown in Figure 3. In terms of topic, we used a previously built and validated topic classifier [48] on all chats in our dataset to find the distribution of topics. The topic classifier we used from Wang et al. [48] was built using the same dataset as our study. By using Empath [46], a tool that uses neural word embeddings for generating and validating lexical categories in large-scale text data, Wang et al. tuned Empath's model by feeding in "seed words" of 18 popular topics that support-seekers discuss. The top 18 topics include romantic relationships, dating, pandemic, self-improvement, suicide, depression, parents, anxiety, family, stress, lonely, overwhelming, sexuality, LGBTQ, intimacy, home, dissociative identity, and health. Similar to assignment of the previous personal characteristics, we applied the real OMHC dataset's distribution of topics to assign support-seeker agents a topic-of-interest and counselor agent a list of up to 3 topics-of-interest that they have more experience or interest in talking about. Distributions of topics among support-seekers and volunteer counselors are shown in Figure 3.

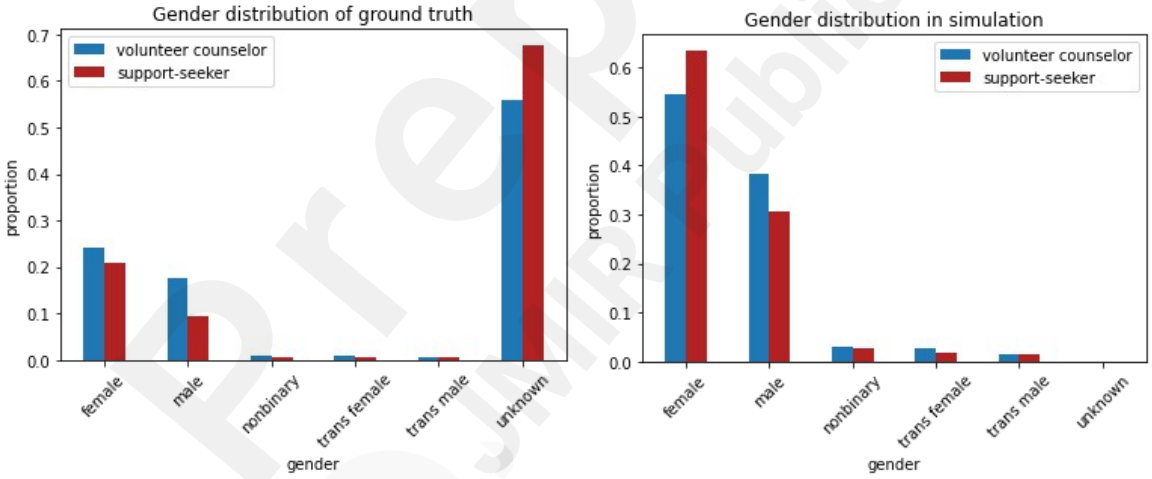


Figure 3. Gender distribution in reality (left) vs. our simulation (right). In the ground truth dataset, many support-seekers and counselors have unknown gender. In contrast, all agents are assigned gender in our simulation. We assign gender according to the gender distribution of known genders (female, male, nonbinary, trans female, trans male) in the ground truth dataset.

Patience Level

As support-seekers may leave the site while waiting for a chat, we also replicated the "patience level" of support-seekers. All support-seeker agents were given a number of minutes that they are willing to wait to be matched before they cancel their request (i.e. leave the platform). Support-seeker agents go offline when the number of simulation steps where the support-seeker remains unmatched exceeds their patience level. Similar to previous characteristics, we assigned patience levels to support-seeker agents according to the distribution we found from analyzing the dataset for how long support-seekers wait until canceling their chat requests in the queue. We assigned support-

seeker agents a patience level according to the distribution of time for support-seekers to cancel their chat requests in the dataset. The mean of patience level in reality is 4.15 with standard deviation of 3.26 whereas our simulation's patience level has a mean of 4.16 with a standard deviation of 3.27. The Pearson correlation is 0.991.

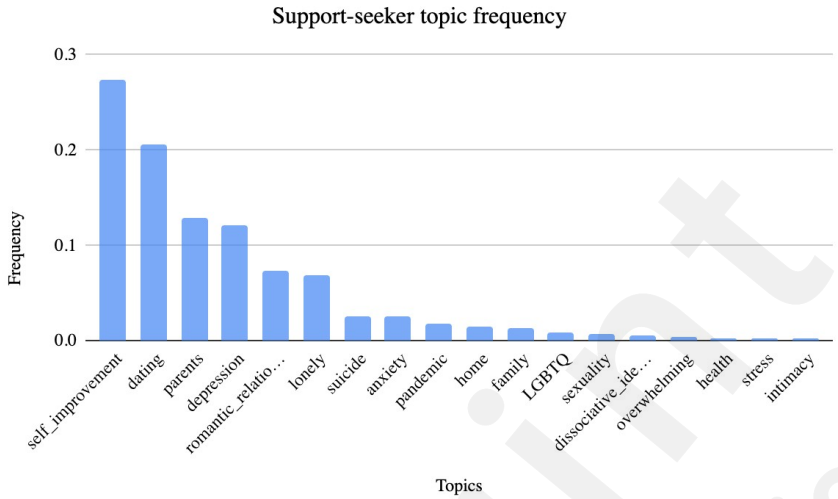


Figure 4. Distribution of topics among support-seekers in the real OMHC dataset. The most frequent topic discussed on the site is "self improvement", followed by "dating", "parents", and "depression". Our simulation's frequency distribution is the same, with a Pearson correlation of 1.

Chat Length

Once a support-seeker and volunteer counselor are matched, their chat length is set in minutes and volunteer counselor agents go offline after a chat ends. We set the chat length in our simulation to follow the distribution of conversation length found in the log data. The distribution of chat length in reality found through log data versus our simulation's distribution has a Pearson correlation of 1. The mean of chat length in reality is 17.67 min with standard deviation of 15.44 min, whereas our simulation's chat length has a mean of 17.67 min with a standard deviation of 15.42 min.

Matching Support-seekers and Volunteer Counselors

Lastly, we describe the decision of matching a support-seeker and volunteer counselor together to chat.

All agents who are online and not chatting in the current simulation period are considered available to be matched in the current round. In each simulation period, all volunteer counselor agents are presented with a list of all available support-seeker agents and pick a support-seeker to chat with depending on the matching policy. In the replication of the research site's system, volunteer counselors pick a support-seeker to chat with randomly, following an exponential distribution model for the time it takes to make their choice (Figure 5).

Figure 5. Distribution of decision time of volunteer counselors, which is modeled by an exponential distribution with lambda of 1.25.

Building Agent-Based Simulation: Prediction Models

Although there are several outcome metrics that could be used to assess a support-seeker's experience in a chat, the ability for a support-seeker to give the chat a rating of 1 to 5 stars is the most direct indicator of the match. Additionally, past work has found that users place more value on avoiding the worst chat experiences rather than aiming for the best given that extremely negative chats may include bullying and harassment and also may deter users from returning to the site [18]. Blocking another user is the main action that users take when they have a bad chat on the research site and is the most obvious reflection of a negative support-seeker and volunteer counselor relationship. Both support-seekers and volunteer counselors on our study's OMHC can block each other through the chat interface and provide a reason for blocking.–

Note that, although our dataset includes good indicators of matching quality, there is no existing data that can be used to calculate matching quality for simulated results. Therefore, we created two outcome prediction models - a chat rating prediction model and a blocking prediction model. In the validation process, we used those two models to test whether our replicated simulation is an accurate representation of the current system. In our virtual experiments, we use these prediction models to evaluate the effectiveness of our designed matching algorithms as well.

Chat Rating Prediction

We gathered all chat ratings in the dataset, with 80% of samples from the dataset used as our training set while the remaining 20% were used as our testing set. We balanced the training set using SMOTE oversampling, which generates artificial samples for minority classes based on existing samples using the K-nearest neighbor algorithm [12]. After SMOTE oversampling, we ended up with 16,730 samples for each of the five chat rating classes (1 through 5 stars).

As independent variables in our chat rating prediction, we used inputs of both volunteer counselors' and support-seekers' gender, birth year, and topic. Our experiments included random forest, logistic regression, SVM, and decision tree models in order to find the best model to predict chat rating. Note that accuracy in this context means that the output must equal the true chat rating, and is considered incorrect if it outputs any of the other four chat ratings. Given that random forest outperformed all other models in accuracy and F1-score at a 0.72 accuracy and 0.72 F1-score, we utilized the random forest classifier for our chat rating prediction model.

Blocking Prediction

In creating our training dataset for building the blocking prediction model, we labeled a support-seeker and volunteer counselor pair as 1 if at least one person blocked the other, and 0 otherwise. Similar to our chat rating prediction model, we utilized input fields of gender, birth year, and topic. For all agents, we used an 80%/20% split of samples from the dataset for training and test, respectively. Table 3 shows the number of samples in our training set, along with results from oversampling to balance the training set. We proceeded again with random forest classification as it resulted in the best performance with highest precision and recall scores of 0.91 accuracy and 0.91 F1-score.

Validation of Agent-Based Simulation

During the validation phase, we calculated and compared distributions of the features below between the real OMHC system data and our simulation's data. We then report the Pearson correlations [43] between the real vs. simulated data distributions.

Number of Users Online

To validate replication of the OMHC system, we split the dataset to a 6-month training set and 2-month test set. Figure 6 compares the number of online support-seekers and volunteer counselors, respectively, in each minute between training set and test set in a week's period of time. We found similar distributions between the training and test set, with Pearson correlations of 0.974 and 0.982 respectively.

Chat Ratings

We used our chat rating prediction model to compare the distribution of chat ratings between our simulation and ground truth. We found that we accurately simulated rating distributions on the research site, with a Pearson correlation of 0.99 between our replication and ground truth (Figure 7). Similarly, we validated our replicated system using the blocking prediction model. Using the blocking prediction model, we found similar distributions of pairs that engage in blocking as in ground truth with a Pearson Correlation of 0.949 as shown in Figure 7, leading to further confidence that our simulation is an accurate representation of the current platform's system.

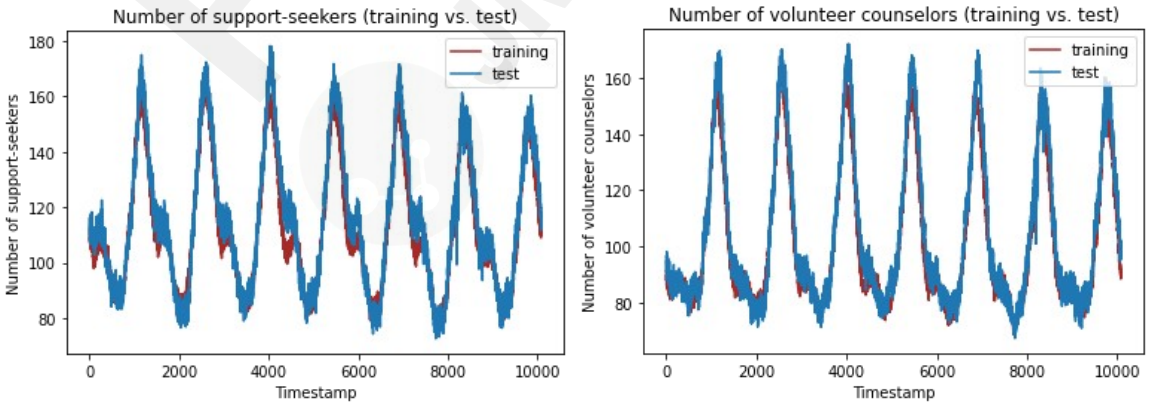


Figure 6. Distribution of number of support-seekers (left) and volunteer counselors (right) in training and test set. Pearson correlation of number of support-seekers between training and test set is 0.974 while Pearson correlation of the number of volunteer counselors between training and test set is 0.982.

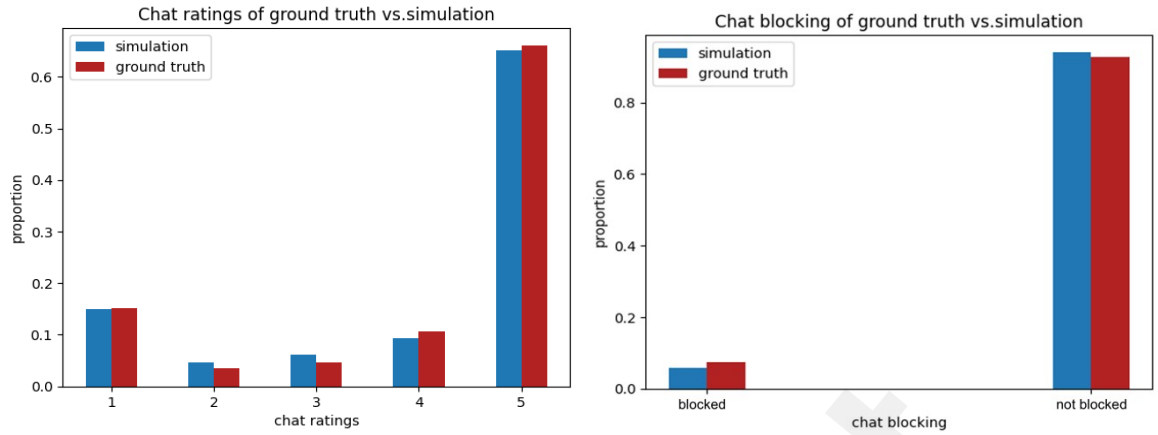


Figure 7. Comparison of chat ratings (left) and blocking (right) between reality versus our simulation. Our simulation (shown in blue) has proportions of ratings from 1-star to 5-stars were, respectively, 14.96%, 4.64%, 6.08%, 9.26%, 65.06%, while ground truth (shown in red) proportions were 15.18%, 3.51%, 4.56%, 10.63%, and 66.12%. In terms of pairs that resulted in blocking, the ground truth proportion of blocked support-seeker and volunteer counselor pairs is 5.3%, compared to our simulation's proportion of 5.86%.

Waiting Time

We also validated using the waiting time of support-seekers matched in our simulation. We found that the distribution of waiting time for support-seekers who are matched with a volunteer counselor is similar to the distribution in the real online community system, although with a lower standard deviation. The mean of waiting time in reality is 3.2 min with standard deviation of 2.9 min, whereas our simulation's chat length has a mean of 3.2 min with a standard deviation of 2.02 min. The Pearson correlation is 0.995.

Matching Rate

We also compared the proportion of support-seekers who were matched with some volunteer counselor to the total number of support-seekers available to be matched. Note that support-seekers may cancel their chat request if they are unmatched for longer than their patience level. In order to evaluate ground truth, we analyzed the proportion of chats taken by volunteer counselors on the research site to the total number of requests by support-seekers in the research site queue. We found that our simulation resulted in an overall 78.35% matching rate in a week while ground truth showed an average of 83.27% matching rate across all weeks in our dataset. Although our simulation's matching rate is slightly lower than ground truth, the likely explanation is that our simulation framework only allows support-seekers and volunteer counselors to engage in one chat at a time whereas the research site allows for volunteer counselors to take multiple chats at once if they desire. Given this, we found that our simulation's matching rate is acceptable and still closely resembled the actual state of the research site.

Virtual Experiments

Next, we applied our simulation as a "sandbox" to test different matching algorithms. Below, we review the matching algorithms in our experiment and their outcomes on metrics of chat ratings, blocking, waiting time, and matching rates.

Applicant-Proposing Deferred-Acceptance Algorithm

Our research problem consisted of two types of agents (support-seekers and volunteers) with

preferences and personal characteristics for matching. As a result, we consider it akin to the stable marriage problem, which seeks to find a stable matching between two classes of elements where both sides have an ordering of preferences. Thus, we employed the **applicant-proposing deferred-acceptance algorithm** as our matching algorithm [2,41], adapted from the established matching method for New York Public Schools [1].

In each simulation period, there are two sets of agents: support-seekers $M= \{m_1, m_2, ... ,m_n\}$ and volunteer counselors $L = \{l_1,l_2, ... ,l_n\}$. Each support-seeker has an ordered preference list $P(m)= \{l_1, l_2, ... , l_m\}$where a support-seeker's first choice is volunteer counselor l_1, second choice is volunteer counselor l_2, etc .

1. Each support-seeker "applies" to her highest ranked volunteer counselor according to their preferences, and each volunteer counselor "holds" her highest ranked application and rejects the rest.
2. At any stage at which a support-seeker has been rejected, she "applies" to her next most preferred volunteer counselor who the support-seeker agent prefers (if one remains). Each volunteer counselor holds its most preferred set of applications and "rejects" the rest.
3. The algorithm stops when no rejections are issued, and each volunteer counselor is matched to the applicants she is holding.

Any agent who is not matched in this simulation period is marked as "waiting", and continues to be available for matching in the next matching period (along with any newly generated agents). Appendix shows the pseudocode for the above algorithm.

Matching Algorithms

The key design choice in algorithmic matching is to construct the preference lists of support-seekers and volunteer counselors. Constructing the preference lists of support-seekers and volunteer counselors is a key design choice in algorithmic matching. Unlike school or physician matching with limited options, hundreds of volunteer counselors are available for any support-seeker at any given time. Since it is impractical to ask each support-seeker to rank each counselor (and vice versa), preferences must be generated using rules or prediction models. The seven algorithms are described below.

We chose methods based on our prior understanding of the research site support-seekers' needs and consultations with the research site's leadership, including the CEO and lead engineer. In general, the research site team expressed that their priorities were to optimize satisfaction on the platform (i.e. chat rating) without reducing the general matching rate. It was important to the research platform that matching protocols would allow for better experiences for support-seekers, but also that the research site could continue to serve the huge majority of its large support-seeker population. Platform leadership was also particularly interested in gender-based protocols, given that gender was a key factor in how people chose volunteer counselors. Additionally, a "hard filter"-based protocol was suggested in order to try mitigating harassment. Community leaders and our research team decided against directly optimizing for outcome metrics, such as rating-optimized or blocking-minimized protocols for validity reasons; our study's prediction models (see Section Building Agent-Based Simulation: Prediction Models) use these same metrics (e.g. rating, blocking), which would create validity issues if we were to make evaluations on protocols that directly optimize for these metrics.

Each algorithm varies in how it "recommends" a support-seeker to each volunteer counselor in each simulation period - volunteer counselors have a 90% chance of taking the recommendation and a

10% chance of random selection.

1. **First-Come-First-Serve**: support-seekers are ranked in volunteer counselors' preference lists by decreasing waiting time. The goal of first-come-first-serve is to improve the number of successful matches and to prioritize serving support-seekers who arrive to the queue first [9].

2. **Last-Come-First-Serve:** support-seekers are ranked in volunteer counselors' preference lists by increasing waiting time. The goal of last-come-first-serve is to minimize queue size when support-seekers are more likely to leave the queue the longer they wait [9].

3. **Similarity-Based**: We use three dimensions of access to us that have been shown in prior literature to be important for matching purposes: **gender**, **age**, and **topic** of discussion. Using gender, age, and topic, we define a vector for each agent with their gender identity, birth year, and topic-of-interest to calculate two agents' similarity using the cosine similarity between their vectors. Agents with higher similarity get ranked higher in other agents' preference lists.

   As reviewed above, past work has found that clients often seek therapists similar to themselves among multiple dimensions. In terms of **gender,** prior work has found that gender is one of the most important factors in choosing a support-provider in online platforms in both client preferences and outcomes, and especially so for gender minorities [3,18]. Additionally, people who are closer in **age** (especially for older populations) are better suited for one another in OMHC support [18] given their similarity in experiences and communication. Lastly, having a therapist that is knowledgeable and willing to discuss the client's needs and issues is vital to the therapeutic relationship. **Topic** relevance is a widespread idea in the space of online communities; for example, social media sites regularly infer a user's interests to recommend them relevant content (e.g. purchase suggestions, personalized advertising) using someone's previous behaviors on the site [6]. Similarly, in our case we judge a counselor's expertise on topics based on their past chats. Each support-seeker is assigned a topic-of-interest that they wish to chat about, according to distribution based on the real OMHC dataset. Based on the top 3 topics by frequency for each volunteer counselor in our OMHC dataset, we assign each simulated volunteer counselor a top 3 topic list. Volunteer counselors with a relevant topic in their list are ranked higher in the respective support-seekers' preference list, and vice versa.

   Additionally, we include matching protocols exploring each of these three dimensions (gender, age, topic) individually.

   4. **Age-based:** support-seekers and volunteer counselors who are closer in age to each other are ranked higher in each other's preference lists.
   5. **Gender-based:** support-seekers and volunteer counselors with the same gender identity are ranked higher in each other's preference lists.
   6. **Topic-based**: volunteer counselors are prioritized in a support-seeker's preference list if the counselor's topic list contains the support-seeker's topic-of-interest; similarly, support-seekers are prioritized in volunteer counselors' preference lists if the topic-of-interest is in the counselor's expertise. In other words, if a support-seeker's topic is in a counselor's topic list then we label a feature of "expertise matching" as 1 (0 otherwise) and use expertise matching as the only matching criteria.

7. **Filter-based**: Suggested by stakeholders in our study's research site, we implement filter-based methods in order to prioritize protection of two vulnerable groups – namely in this

context, teenagers and gender minorities. The filter-based method includes three different pools for agents: under-aged (18 or younger) pool, gender-minority (non cisgender women nor cisgender men) pool, and all others. Volunteer counselors can only select support-seekers who are in the same pool as themselves. Apart from this limitation on what support-seekers are available to volunteers though, volunteer counselors can pick up anyone (i.e. random selection) following the existing protocol on the site's simulation (see section "Matching Support-seekers and Volunteer Counselors").

Results

Our study's primary contribution is the application of agent-based simulation to the online mental health context. However, we also review below the outcomes of experimenting with seven specific algorithms using our simulation sandbox, shown in Tables 1-3. Full table results are in the Appendix. Although these are just seven of possible protocols for matching in this context, in the following sections we review our comparison of their outcomes to show the kinds of findings and trade-offs revealed through an agent-based simulation.

The outcome metrics used to evaluate the algorithms are defined as:

- **Average rating**: Average rating for all pairs in simulation predicted by the rating prediction model. *The higher the better.*
- **% of blocked pairs**: Percentage of pairs in the simulation that are predicted as "block". *The lower the better.*
- **Matching success rate**: the ratio of support-seekers who were successfully matched (i.e. chatted with a volunteer counselor) to all support-seekers in the simulation. *The higher the better.*
- **Average waiting time (matched)**: average waiting time of support-seekers before they match in the simulation. *The lower the better.*
- **Average waiting time (unmatched)**: average waiting time of support-seekers whose waiting time exceeds patience level and quit. *The higher the better.*

We ran t-tests to see if different algorithms' outcome metrics were statistically significant when compared to the replication protocol. We indicate both the standardized p-threshold of .05 as well as a more conservative threshold of .001 in the tables below, focusing our discussion of findings on the $P < 0.001$ findings.

Table 1. Outcome metric results for different algorithms' performance. Significant (*P<.05, **P<.001) outcomes when compared to the replication of the research site (first row) are indicated.

	Avg. rating	% of blocked pairs	Matching Success Rate	Avg waiting time for matched clients	Avg waiting time for unmatched
Replication of the research site	4.05	5.86%	78.91%	3.19	3.69
First-Come-First-Serve	4.06	5.82%	81.93%**	3.68**	2.73**
Last-Come-First-Serve	4.04	6.11%	74.81%**	2.61**	4.69**
Similarity-based	4.04	7.37%**	79.36%*	3.34**	3.53**

Age-based	4.02**	7.22%**	79.97%**	3.40**	3.41**
Gender-based	4.07*	6.04%	81.81%**	3.61**	2.86**
Topic-based	4.31**	4.26%**	80.70%**	3.45**	3.29**
Filter-based	4.03*	6.21%*	61.30%**	3.24**	3.86**

* P <.05
** P <.001

Results, overall

Average rating

Overall, the results of our virtual experiments followed intuition in that different matching policies served different goals. Chat rating was one of the metrics of the most interest to our research team in evaluating algorithm outcomes, given its interest to our community stakeholders as a good proxy for community satisfaction on the site. We found two reliable (P<.001) results, where the topic-based protocol had significantly higher ratings compared to the replication protocol and the age-based had significantly lower ratings.

The topic-based protocol performed significantly better than not only the replication of the research site, but also all other algorithms. While all other algorithms output average (mean) ratings between 4.02 to 4.07 out of 5 stars, the topic-based protocol had a 4.31 average rating out of 5 stars – giving a statistically significant increase of 6% over an already-high baseline replication. In terms of the age-based protocol, it had a statistically significant but marginal decrease in chat rating performance of 0.7% compared to the replication. Generally, we found that the age-based protocol performed well for minors (see Table 2) but did not have any improvement for adults (which makes up the majority of the population). As a result, age-similarity seems to matter for young people connecting but not for the general adult population. We discuss more about the breakdown of protocols across demographic groups in the next section (“Results, by demographics”).

Blocking

In general, blocking rates remained low for all protocols including in the replication. However, we do see one statistically significant improvement among protocols particularly when using the topic-based protocol to match pairs; the proportion of support-seeker and volunteer counselor pairs who engage in blocking (from either party) was reduced from 5.86% to 4.26%. On the other hand, similarity-based (combining age, gender, and topic) and the age-based protocol had a marginal increase in the blocking rate.

Matching success rate

In terms of overall matching rate, we found that First-Come-First-Serve, as expected, led to the highest matching rate while Last-Come-First-Serve had one of the lowest matching rates. Given that last-come-first-serve as a matching protocol is aimed at keeping queue size small but not necessarily serving the most people, it follows intuition that we see a low waiting time for matched pairs but overall relatively fewer successfully matched pairs (we explore this further below in “Waiting times”).

Worth noting is that all protocols other than similarity-based (P>.001) showed statistically significant results compared to the replication when it came to the matching success rate. First-Come-First-Serve, age-based, gender-based, and topic-based were all protocols that resulted in higher matching rates. Last-Come-First-Serve and Filter-based mechanisms resulted in statistically significant lower

matching success rates for the community. Filter-based resulted in the most striking difference – a 22.3% reduction in overall matching rate; although we find in Table 2 and discuss later in this paper that there are many benefits to a hard filter-based mechanism on certain groups, our simulation showed a substantial trade-off when it comes to the quantity of people in the community able to find support chats.

Waiting times

Wait times are generally all within 1-2 seconds of each other when comparing across protocols. We note an intriguing finding, though, in that First-Come-First-Serve initially shows counterintuitively one of the *longest* waiting times for people who are successfully matched. However, upon further reflection this result actually is to be expected. First-come-first-serve matches the greatest number of people, and thus more people with higher patience levels and thus have *higher waiting times* are also able to be matched. This pushes the average wait time of successful matches higher. The average waiting time for *unmatched* agents thus only includes the agents who were impatient (short patience levels); as a result, the average waiting time for matched clients looks higher and the average waiting time for unmatched clients looks lower. The opposite case occurs for last-come-first-serve; there are short wait times for successfully matched support-seekers, but few support-seekers are successfully matched. We thus note for community stakeholders that evaluating the average wait time requires consideration of the success rate of matching in conjunction, rather than being evaluated in isolation.

Results, by demographic

We show in Tables 2 and 3 a breakdown of different algorithms' performances for chat rating and blocking percentage, by demographic groups of adults, under-aged people, non-gender minority people. Note that we do not have racial information for this analysis, so we are limited to commenting on effects on only gender and age matching for this study. We also provide the results for matching rate and waiting times in our Appendix. In general, the largest differences are seen for under-aged people and gender minorities, showing that algorithmic matching has the greatest potential effect for these often marginalized, excluded, and vulnerable groups.

When breaking down protocols by their performances among different groups, we see several trade-offs on how algorithms perform on majority versus minority groups. It also allowed us to see that the current state of the research site has drastically different effects on various demographic groups. For blocking rate, we noted that the replication of the research site had significantly higher blocking rate of 12.31% among gender minorities – an alarming rate that is more than double that of non-gender minorities. This follows prior work finding that LGBTQ+ communities on OMHCs suffer from higher risk of harassment [22]. This finding also allows us to see the major degree of improvement allowed for by the algorithmic matching, as we found the filter-based algorithm (restricts gender minority support-seekers to only being paired with also gender minority

Table 2. Chat rating results for different algorithms' performance among different groups. Significant outcomes when compared to the replication of the research site (first row) are indicated. Our simulations show that similarity-based protocol (and its sub protocols) as well as Filter-based protocol have significant effects on the chat rating among demographic groups. For example, both Filter-based and Similarity-based protocols raised chat rating for minority groups but lowered for non-minority groups.

Chat Rating, by demographic

	Adult	**Under-aged**	**Non-gender minority**	**Gender minority**

Replication of the research site	4.00	4.28	4.06	3.80
First-Come-First-Serve	4.01	4.31	4.07	3.84
Last-Come-First-Serve	3.98	4.31	4.05	3.83
Similarity-based	3.95**	4.46**	4.04**	4.20**
Age-based	3.94**	4.39**	4.03**	3.76
Gender-based	3.94	4.32	4.06	4.14**
Topic-based	4.26**	4.53**	4.33**	4.06**
Filter-based	3.92**	4.39**	3.98**	4.69**

*P<.05
**P<.001

counselors) led to a striking improvement of 96% for blocking (from 12.31% to 0.44%) and 23% for chat rating. Our simulation results thus give quantitative evidence to previous qualitative work that has suggested LGBTQ+ support-seekers prefer a counselor of similar identity [22]. In particular, the reduction in blocking behavior is promising towards protecting minority or vulnerable groups from experiencing unwanted behaviors or relationships.

There are significant trade-offs with algorithm choice that emerged from our breakdown of results by demographic group. As reviewed above, the filter-based algorithm performed exceptionally when it comes to gender minorities. It resulted in an impressive average chat rating of 4.69 out of 5 stars for gender minorities, a 23.4% improvement over the replication protocol. It also predicted a very low proportion of pairs who block one another for both minors and gender minorities (2.35% and 0.44%, respectively, compared to the replication's results of 6.73% and 12.31%). However, it notably performs overall poorly when it comes to chat rating and blocking among adults, under-aged, and gender majority groups. In fact, the filter-based algorithm resulted in *worse* performance compared to the replication site for all other groups for both chat rating and blocking; this indicates that the hard filter approach results in worse user experiences for the majority of the site's population compared to if there was no algorithmic consideration in matching at all. This result may complement previous work that found LGBTQ+ users in OMHCs feel safer speaking with other LGBTQ+ users [18], but also reveals a likely exchange for lower overall satisfaction among majority groups on a platform such as the research site. We see a similar trade-off when it comes to the overall similarity-based model. Compared to the replication of the research site, the similarity-based protocol shows a significant decrease in chat rating for adults and non-gender minorities (-1.3% and -0.4%, respectively) but a significant increase for under-aged people and gender minorities (+4% and +10.5%, respectively). In terms of blocking, there is an almost equal rise in blocking rates for adults and non-gender minorities (+29% and +35%) as there is a decrease in blocking rates (a positive result) for under-aged people and gender minorities (-32% and -35%, respectively). Note, however, that except for gender minorities, blocking rates are relatively low among all other groups to begin with.

Table 3. Blocking results for different algorithms' performance among different groups. We indicate

statistically significant (*p<0.05, **p<0.001) findings. Our simulations show that Similarity-based protocol helps reduce blocking for under-age and gender minority individuals, but significantly raises blocking for adults and non-gender minorities. The Topic-based protocol has statistically significant reduction in blocking for Adults, Under-Aged, and Non-gender minorities. Filter-based protocol resulted in only bettering results for gender minorities, with worse performance for all other groups.

Blocking, by demographic

	Adult	Under-aged	Non-gender minority	Gender minority
Replication of the research site	6.73%	1.85%	5.45%	12.31%
First-Come-First-Serve	6.66%	1.91%	1.86%	12.50%
Last-Come-First-Serve	7.03%	1.86%	5.66%	13.13%
Similarity-based	8.69%**	1.25%**	7.34%**	8.00%**
Age-based	8.29%**	1.88%**	6.77%**	14.32%*
Gender-based	7.02%	1.52%	5.89%**	8.64%**
Topic-based	4.97%**	0.99%**	3.76%**	11.84%
Filter-based	7.38%**	2.35%**	6.66%**	0.44%**

*P<.05
**P<.001

Despite this, though, one of the most notable takeaways from our simulation results is that this trade-off between the experiences of minority and majority groups is *not* necessarily the case with all protocols. We found that the topic-based protocol performed well overall even among the marginalized communities we studied, despite no consideration of gender or age in its matching criteria, as well as improved upon experiences for the general community. When compared to the replication protocol, we found that all results for the topic-based protocol, with the exception the insignificant finding for blocking rate of gender minorities, were improvements *both* chat rating and blocking among *all* groups; other outcomes similarly showed improvement or similar performance to the replication site as shown in our full results in the Appendix. In terms of under-aged people, the topic-based protocol (chat rating of 4.53, blocking rate of 0.99%) actually outperformed the hard-filter based protocol (chat rating of 4.39, blocking rate of 2.35%) that is the most restrictive protocol for prioritizing the matching of demographic groups. As a result, a key takeaway from these results is that directly using demographics of gender and age as matching criteria is not necessary to improving the experiences of the targeted groups. Although topic-based matching does not include any demographic consideration, it still led to significant improvement for groups that are marginalized or vulnerable from their demographics (age, gender) and shows that improving the experiences of minority groups is not necessarily exclusive to improvement for the overall population as well. However, we note that our findings of demographics' effects in matching are limited to gender and age; racial matching being an important factor in matching for traditional services, but our analysis lacks enough racial data to analyze for this study and thus we cannot draw

conclusions the role of any racially- or culturally-based matching.

Discussion

Agent-based modeling for designing OMHCs

In this work, we showcased our important contribution of applying agent-based simulation to the OMHC context and provided a framework for how it can help the creators and designers of online communities weigh the various trade-offs when building mechanisms to best help their support-seekers find meaningful relationships online. Our study suggests that different goals can be achieved through different algorithmic choices for these communities, from optimizing the quality of conversations to the protection of gender and age minorities on these platforms.

Our research was guided by previous literature that revealed how algorithmic matching is particularly beneficial in the online mental health context, and the numerous considerations that are necessary for effective partnerships to aid users' well-being [7,18]. Through simulating the algorithms and mechanisms of the research site, we found that there are numerous trade-offs to be made in deciding how to match users together in OMHCs. For example, communities aiming to optimize the number of users that engage in chats may wish to experiment with first-come-first-serve methods given simulation results that it has the highest overall matching rate. Alternatively, algorithmic matching may lead to better conversations between users but also increase the time that users must wait for a match. Our simulation revealed that there is indeed substantial improvement that can be made in bettering the quality of conversations using algorithmic matching, such as more than 6% increase in the average chat rating just using topic. However, users may become impatient while waiting for a "best fit" chat partner, log off the site, and thus fewer people overall are helped by the platform. We see this reflected in our experiments - for example, first-come-first-serve is able to match the most people, with a matching rate of nearly 82% despite being the simplest algorithm without optimization for users' characteristics.

We also considered and offered the benefits of agent-based modeling in conversations with the stakeholders of the research site. Although these conversations introduced other important considerations in algorithmic matching implementation, such as the computational resources and refactoring of the codebase to implement matching, the stakeholders of our study's research site found our results insightful and beneficial to the community. In particular, the site's leadership expressed that simulating algorithmic matching helped them weigh the effects and trade-offs of different design choices for their site without disrupting the existing community through continuous algorithmic testing. Other conversations were also had among our research team and the site's leadership, such as which outcome measures were most crucial in evaluating the platform's success for support-seekers. The research site of our work has begun developing an algorithmic matching system for their platform based on this work.

It is also worth noting that the research site' stakeholders wanted to know more about the reasoning behind the numerical results for each matching algorithm. As a result, we believe it is important that presentation of these experiments is accompanied by model explanation and algorithmic transparency as well in order to best help community stakeholders understand results and best make decisions based on these types of simulation. Indeed, this suggestion is supported by a plethora of past research that has found that model transparency is necessary for trust and understanding when deploying new algorithms in online communities, for both community decision-makers and community members [17,25,26]. We have released our simulation open-source; see below "Data Availability" section.

Matching Criteria for Online Support Chats

It is important to note that there does not exist a universally best design for all communities; instead, the choice of algorithms and mechanisms for mental health communities is specific to the community's context, its goals, and its support-seekers [30,48]. However, although our experiments are not necessarily generalizable to every OMHC, they do provide some initial insight into making algorithmic choices for these communities.

Past work in traditional mental health services (e.g. therapy) has found that using demographics for matching clients and providers is important for outcomes, particularly so for minority and vulnerable groups [3,19,24]. Our findings in the online community context support this notion in that matching based on gender and age indeed improved outcomes for under-age and gender minority groups. However, a more surprising result from these experiments in our agent-based simulation model is that matching approaches that do *not* use demographics as matching criteria can also serve to protect vulnerable groups. As discussed, we found matching solely based on topic resulted in high rating and low blocking percentages among all groups and was an improvement on the baseline. Indeed, topic-based matching outperformed for average rating, proportion of blocking, and the matching success rate when compared to the similarity-based protocol that also included age and gender as matching criteria. However, it is possible that using race or cultural background as a matching criterion could lead to improvements given their importance in research on matching for traditional mental health contexts. In terms of our findings on gender and age, though, topic as an effective metric for matching people in vulnerable groups makes some intuitive sense; people who are teenagers, for example, are likely to share similar issues compared to older adults (e.g. parenting, divorce) and people who are gender minorities may also use an OMHC to seek support for gender issues, transitioning, or other general LGBTQ+ questions [18]. As a result, topic may inherently connect people who are of not only similar demographics but also similar *experiences* with one another – this includes those who aren't necessarily in what our simulation considered a vulnerable group (minors, gender minorities), and thus we see the topic-based protocol performing well on the overall population (Table 1) largely due to its improvement for majority groups in addition to minority groups (Table 2 and Table 3). It is important to note the privacy risks when it comes to people's personal information given to online platforms, and there is some evidence that people have reservations when it comes to revealing this information [18]; thus, our work importantly reveals that bettering people's experiences is not necessarily reliant on more demographic or other personal information about them.

We note another surprising result that using similarity between people amongst all features – age, gender, topic– in our similarity-based protocol generally did not results in significant improvements compared to other protocols. Supporting prior work on the importance of demographic features [3,18,19,24], our evaluation of outcomes by demographic features (see Tables 2 and 3) show improved experiences for the group that a protocol uses as a matching feature (i.e. the gender-based protocol for gender minorities, the age-based protocol for minors). However, the *general* similarity-based protocol that uses all features (age, gender, topic) only showed marginal improvement among under-aged and gender minorities when it came to blocking proportions (Table 3) albeit slightly more substantial improvements for the same groups when it came to chat rating (Table 2). When evaluating the findings in Tables 2 and 3, we hypothesize that simply matching on all characteristics is not necessarily conducive to an overall improvement for vulnerable groups; instead, intentional choice must be made for selecting certain features (like gender for gender minorities) to aid their experience. Overall, we find that just as much or even more improvement is seen for vulnerable groups when solely matching on one characteristic, such as gender for gender minorities or age for under-aged people, and even more so that topic is as good or even better as a sole matching criterion

compared to cosine similarity among all features. However, we also acknowledge a potential issue in our simulation's use of cosine-similarity as it may not be the best way to measure similarity between people in this context; further work may be necessary to test other similarity measurements. Additionally, as mentioned, the lack of racial and cultural background data for our study may have limited our findings.

Limitations

Our work has several limitations. First, although our paper's primary contribution is using agent-based modeling to show how we can simulate algorithmic outcomes for online mental health communities, we note that we cannot replicate the complex systems of these communities completely. Secondly, chat ratings and blocking are just two possible outcome metrics to evaluate the performance of a support-seeker and volunteer counselor pair. These measurements have limitations; for example, volunteer counselors are unable to rate chats and support-seekers may use the 1 to 5 stars scale differently. There are other possible metrics that may have been good measurements of the performance of a support-seeker and volunteer counselor pair. For example, the research site periodically sends out emotional wellness tests to support-seekers to evaluate how their mental health has improved over time; unfortunately, during our data collection period, support-seekers rarely completed these tests and thus we did not have adequate data to analyze these results for our simulation. Retention is another metric that could be utilized to evaluate how a support-seeker and volunteer counselor pair impacted the users; however, retention is an unclear metric in the OMHC contexts as it can indicate either failure or success of the community. Third, our matching analysis lacks one key demographic source: race. As we mentioned in our Related Work, race, ethnicity, and cultural matching are important factors that clients consider when finding a mental health service provider. However, our dataset lacks racial data for a huge majority of users and we did not conduct automatic detection of racial disclosure within chats. As a result, our findings can only comment on matching from a demographic-standpoint on gender and age. Lastly, we note our previous mention that optimizing directly for rating, blocking, or other outcome metrics may be an effective and intuitive protocol method for online platforms but we were not able to experiment with these protocols given our prediction model and evaluation methodology.

Conclusions

In this paper, we used agent-based modeling in the online mental health context to reveal trade-offs of algorithmically matching peers. Evaluating data from the online community the research site.com, we provided a simulation model to compare current matching mechanisms and various algorithmic matching policies, and observed their differing effects on outcome metrics including wait time and chat experiences of support-seekers. Our results indicated that algorithmic matching policies based on the applicant-proposing deferred-acceptance algorithm can lead to better chat experiences for OMHC support-seekers while still matching them for chats quickly. Our simulation can aid designers of OMHC and other online communities with a need for matching through enlightening the tensions between goals of matching as well as its impact on different communities.

Acknowledgements

First author Yuhan Liu provided conceptualization, building and evaluation of the simulation system, statistical analyses, and contributed to the writing of the original draft of this paper. Co-first author Anna Fang contributed to the conceptualization, data curation, advised methodology and evaluation, and all writing and editing of this paper. Authors Dr. Glen Moriarty and Cris Firman provided feedback on initial conceptualization of this project as well as platform data for this project. Dr. Robert E. Kraut supervised initial conceptualization, methodology, and supervised the writing and editing of this paper. Finally, last author Dr. Haiyi Zhu led supervision of this work, including

conceptualization, feedback on methodology, reviewing, and editing this paper. We also thank and acknowledge the work of our colleagues Haard Shah and Tony Wang for their topic classifier work done at Georgia Tech that contributed to this study.

Data Availability

The data sets used during our study are not available to the public due to data privacy and collaboration agreements; our dataset was obtained through collaboration with the anonymous online mental health community. However, we have released our simulation code open-source at *https://github.com/liuyuhan1997/Agent-based-Simulation-for-OMHC-Matching* so research site stakeholders, other OMHC designers, and even students can utilize our simulation sandbox to learn and understand the trade-offs in designing real-world matching algorithms.

Conflicts of Interest

None declared.

References

1. Atila Abdulkadiroğlu, Parag A. Pathak, and Alvin E. Roth. 2005. The New York City High School Match. *The American economic review* 95, 2: 364–367.
2. Atila Abdulkadiroğlu, Parag A. Pathak, Alvin E. Roth, and Tayfun Sönmez. 2005. The Boston Public School Match. *The American economic review* 95, 2: 368–371.
3. Steven J. Ackerman and Mark J. Hilsenroth. 2003. A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical psychology review* 23, 1: 1–33.
4. Javier Alvarez-Galvez. 2016. Network Models of Minority Opinion Spreading: Using Agent-Based Modeling to Study Possible Scenarios of Social Contagion. *Social science computer review* 34, 5: 567–581.
5. M. Álvarez-Jiménez, J. F. Gleeson, and S. Rice. 2016. Online peer-to-peer support in youth mental health: seizing the opportunity. *Epidemiology* . Retrieved from https://www.cambridge.org/core/journals/epidemiology-and-psychiatric-sciences/article/online-peertopeer-support-in-youth-mental-health-seizing-the-opportunity/D0790F6E3EE39E6C70DDF74F80E8177A
6. Anitha Anandhan, Liyana Shuib, Maizatul Akmar Ismail, and Ghulam Mujtaba. 2018. Social Media Recommender Systems: Review and Open Research Issues. *IEEE Access* 6: 15608–15628.
7. 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: e21819.
8. Laima Augustaitis, Leland A. Merrill, Kristi E. Gamarel, and Oliver L. Haimson. 2021. Online Transgender Health Information Seeking: Facilitators, Barriers, and Future Directions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–14.
9. Achal Bassamboo and Ramandeep Singh Randhawa. 2016. Scheduling Homogeneous Impatient Customers. *Management science* 62, 7: 2129–2147.
10. Max Birchwood and Swaran P. Singh. 2013. Mental health services for young people: matching the service to the need. *The British journal of psychiatry. Supplement* 54: s1-2.
11. Walter O. Bockting, Michael H. Miner, Rebecca E. Swinburne Romine, Autumn Hamilton, and Eli Coleman. 2013. Stigma, mental health, and resilience in an online sample of the US transgender population. *American journal of public health* 103, 5: 943–951.
12. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *The journal of artificial intelligence research* 16: 321–357.
13. Hardin L. K. Coleman, Bruce E. Wampold, and Sherry L. Casali. 1995. Ethnic minorities' ratings of ethnically similar and European American counselors: A meta-analysis. *Journal of counseling psychology* 42, 1: 55–64.
14. Shelley L. Craig and Gina Keane. 2014. The Mental Health of Multiethnic Lesbian and Bisexual Adolescent Females: The Role of Self-Efficacy, Stress and Behavioral Risks. *Journal of gay & lesbian*

mental health 18, 3: 266–283.
15. Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1: 71–80.
16. Robert Elliott. 2008. Research on client experiences of therapy: introduction to the special section. *Psychotherapy research: journal of the Society for Psychotherapy Research* 18, 3: 239–242.
17. Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300724
18. Anna Fang and Haiyi Zhu. 2022. Matching for Peer Support: Exploring Algorithmic Matching for Online Mental Health Communities. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2: 1–37.
19. Judith R. Felton. 1986. Sex makes a difference—How gender affects the therapeutic relationship. *Clinical social work journal* 14, 2: 127–138.
20. Jessica N. Fish. 2020. Future Directions in Understanding and Addressing Mental Health among LGBTQ Youth. *Journal of clinical child and adolescent psychology: the official journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53* 49, 6: 943–956.
21. U. S. S. General. Mental health: Culture, race, and ethnicity. *A supplement to mental health: A report of the Surgeon*.
22. Xinning Gui, Yu Chen, Yubo Kou, Katie Pine, and Yunan Chen. 2017. Investigating Support Seeking from Peers for Pregnancy in Online Health Communities. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW: 1–19.
23. Christine Howes, Matthew Purver, and Rose Mc Cabe. Investigating topic modelling for therapy dialogue analysis. Retrieved December 7, 2023 from https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/5300/PURVERInvestigatingTopic2013POST.pdf
24. Anna M. Jackson. 1973. Psychotherapy: Factors associated with the race of the therapist. *Psychotherapy* 10, 3: 273–277.
25. Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass. *Proceedings of the ACM on human-computer interaction* 4, GROUP: 1–35.
26. Jakko Kemper and Daan Kolkman. 2019. Transparent to whom? No algorithmic accountability without a critical audience. *Information, communication and society* 22, 14: 2081–2096.
27. Karin Kichler and Johanna Lalouschek. 1987. Therapy talk: Start, beginning, and outset. *Neurotic and psychotic language behavior*: 125–138.
28. Heejung S. Kim, David K. Sherman, and Shelley E. Taylor. 2008. Culture and social support. *The American psychologist* 63, 6: 518–526.
29. R. Kraut and Y. Ren. Bepress guest access. Retrieved November 15, 2023 from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1304&context=icis2007
30. Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press.
31. Frederick T. Leong. 1986. Counseling and psychotherapy with Asian-Americans: Review of the literature. *Journal of counseling psychology* 33, 2: 196–206.
32. Lesley Lowes and Gill Paul. 2006. Participants' experiences of being interviewed about an emotive topic. *Journal of advanced nursing* 55, 5: 587–595.
33. Maxie C. Maultsby. 1982. A Historical View of Blacks' Distrust of Psychiatry. In *Behavior Modification in Black Populations: Psychosocial Issues and Empirical Findings*, Samuel M. Turner and Russell T. Jones (eds.). Springer US, Boston, MA, 39–55.
34. Elizabeth A. McConnell, Antonia Clifford, Aaron K. Korpak, Gregory Phillips 2nd, and Michelle Birkett. 2017. Identity, Victimization, and Support: Facebook Experiences and Mental Health Among LGBTQ Youth. *Computers in human behavior* 76: 237–244.
35. J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2: 113–122.
36. D. Plikynas, A. Raudys, and S. Raudys. 2015. Agent-based modelling of excitation propagation in social

media groups. *Journal of experimental & theoretical artificial intelligence: JETAI* 27, 4: 373–388.
37. Julie Prescott, Rathbone Amy Leigh, and Terry Hanley. 2020. Online mental health communities, self-efficacy and transition to further support. *Mental Health Review Journal* 25, 4: 329–344.
38. Julie Prescott, Amy Leigh Rathbone, and Gill Brown. 2020. Online peer to peer support: Qualitative analysis of UK and US open mental health Facebook groups. *Digital health* 6: 2055207620979209.
39. Yuqing Ren and Robert E. Kraut. 2010. Agent-based modeling to inform online community theory and design: Impact of discussion moderation on member commitment and contribution. *Second round revise and resubmit at Information Systems Research* 21, 3. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2cbeb64680bdede3fe24d989809e9175c2bac820
40. Yuqing Ren and Robert E. Kraut. 2014. Agent-Based Modeling to Inform Online Community Design: Impact of Topical Breadth, Message Volume, and Discussion Moderation on Member Commitment and Contribution. *Human–Computer Interaction* 29, 4: 351–389.
41. Alvin E. Roth and Elliott Peranson. 1999. The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design. *The American economic review* 89, 4: 748–780.
42. Alvin E. Roth and Marilda Sotomayor. 1992. Chapter 16 Two-sided matching. In *Handbook of Game Theory with Economic Applications*. Elsevier, 485–541.
43. P. Sedgwick. 2012. Pearson's correlation coefficient. *BMJ* . Retrieved from https://www.bmj.com/content/345/bmj.e4483.pdf+html
44. Anthony C. Stratford, Matt Halpin, Keely Phillips, Frances Skerritt, Anne Beales, Vincent Cheng, Magdel Hammond, Mary O'Hagan, Catherine Loreto, Kim Tiengtom, Benon Kobe, Steve Harrington, Dan Fisher, and Larry Davidson. 2019. The growth of peer support: an international charter. *Journal of mental health* 28, 6: 627–632.
45. S. Sue. 1988. Psychotherapeutic services for ethnic minorities. Two decades of research findings. *The American psychologist* 43, 4: 301–308.
46. Binbin Than Fast and Michael S. Chen. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, San Jose, California, USA; New York, NY, USA, 4647–4657.
47. G. Turner. 1999. Peer support and young people's health. *Journal of adolescence* 22, 4: 567–572.
48. Tony Wang, Haard K. Shah, Raj Sanjay Shah, Yi-Chia Wang, Robert E. Kraut, and Diyi Yang. 2023. Metrics for Peer Counseling: Triangulating Success Outcomes for Online Therapy Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23), 1–17.
49. Arthur L. Whaley. 2001. Cultural mistrust: An important psychological construct for diagnosis and treatment of African Americans. *Professional psychology, research and practice* 32, 6: 555–562.
50. James E. Yadavaia and Steven C. Hayes. 2012. Acceptance and Commitment Therapy for Self-Stigma Around Sexual Orientation: A Multiple Baseline Evaluation. *Cognitive and behavioral practice* 19, 4: 545–559.

Appendix

Pseudocode for Applicant-Proposing Deferred Acceptance Algorithm

```
Parameter S {support-seekers available to be matched in the current simulation period}

Parameter V {volunteer counselors available to be matched in the current simulation period}

  function stableMatching {
      Initialize all s in S and v in V to be available
      while there exists available s who still has a volunteer v to apply to {
          v = first volunteer counselor on support seeker s' list to whom v has not yet applied
         if v is available
           (s, v) become matched
         else some pair (s', v) already exists
           if v prefers s to s'
              s' becomes available
             (s, v) become matched
           else
             (s', v) remain matched
      }
  }
```

Additional outcome measures for protocols by demographic

	Matching success rate	Avg. waiting time (matched)	Avg. waiting time (not matched)
Replication of research site			
adult	78.97%	3.19	3.69
under-age	78.65%	3.18	3.68
non-gender-minority	78.92%	3.18	3.69
gender-minority	78.76%	3.26	3.74
first come first serve			
adult	81.92%**	3.68**	2.73**
under-age	82.02%**	3.70**	2.72**
non-gender-minority	81.96%**	3.69**	2.73**
gender-minority	81.60%**	3.68**	2.73**
last come first serve			
adult	74.85%**	2.61**	4.71**
under-age	74.64%**	2.62**	4.60**
non-gender-minority	74.81%**	2.62**	4.68**
gender-minority	74.75%**	2.58**	4.87**
similarity			

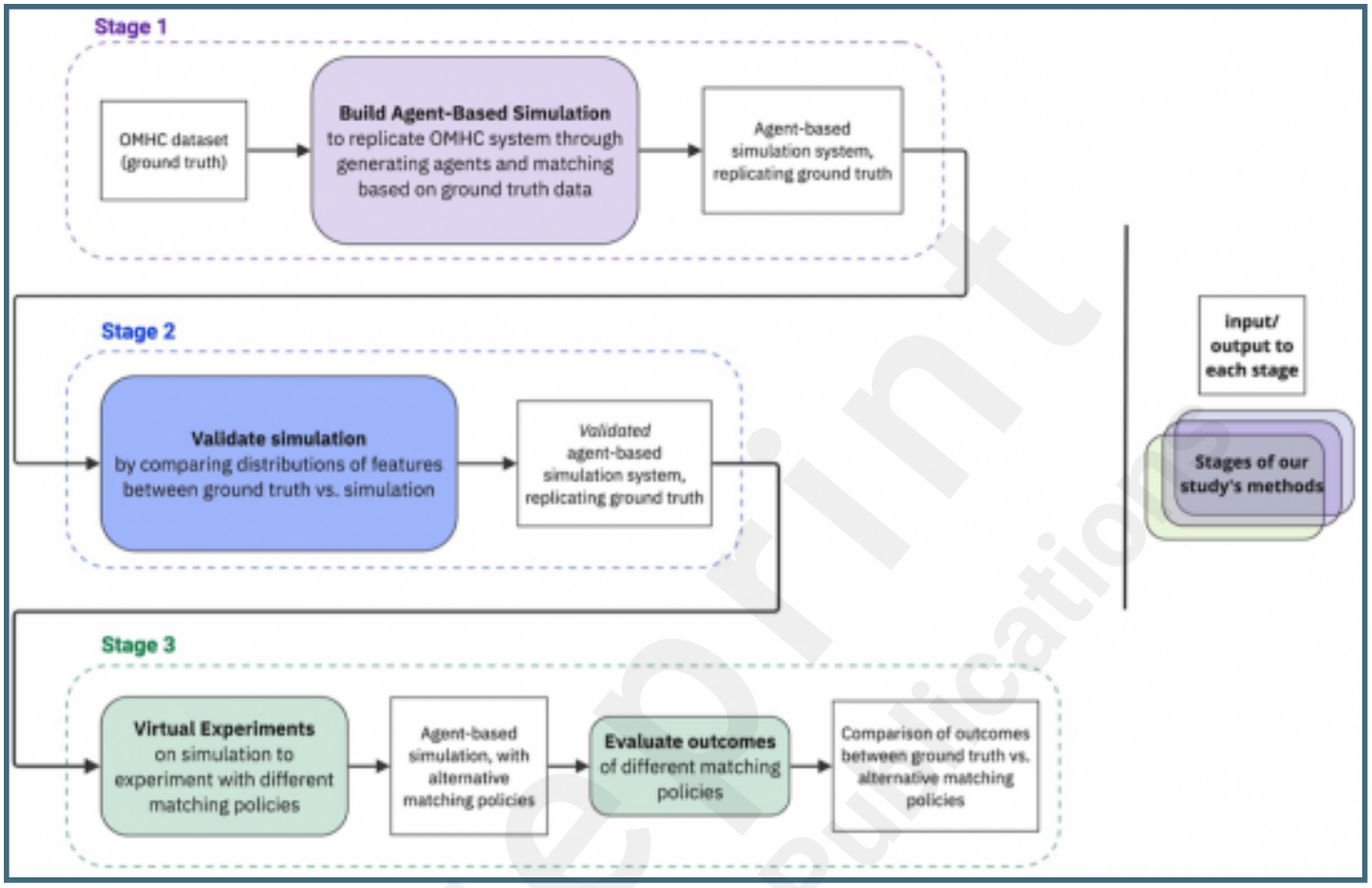
adult	79.27%	3.35**	3.52**
under-age	79.78%*	3.28**	3.55*
non-gender-minority	80.09%**	3.31**	3.49**
gender-minority	67.81%**	3.86**	3.83
- age-based			
adult	81.26%**	3.35**	3.33**
under-age	74.08%**	3.65**	3.65**
non-gender-minority	79.98%	3.40**	3.40**
gender-minority	79.86%	3.43**	3.51*
- gender-based			
adult	81.80%**	3.61**	2.87**
under-age	81.88%**	3.61**	2.86**
non-gender-minority	82.31%	3.61	2.81
gender-minority	74.07%	3.64	3.41
- topic			
adult	80.66%	3.45	3.30
under-age	80.90%**	3.49	3.24
non-gender-minority	80.67%	3.45	3.29
gender-minority	81.24%*	3.49	3.29
filter			
adult	56.07%	3.54	3.90
under-age	85.53%**	2.23	3.28
non-gender-minority	59.85%	3.32	3.85
gender-minority	84.48%	2.20	4.02

*P < .05
**P<.001

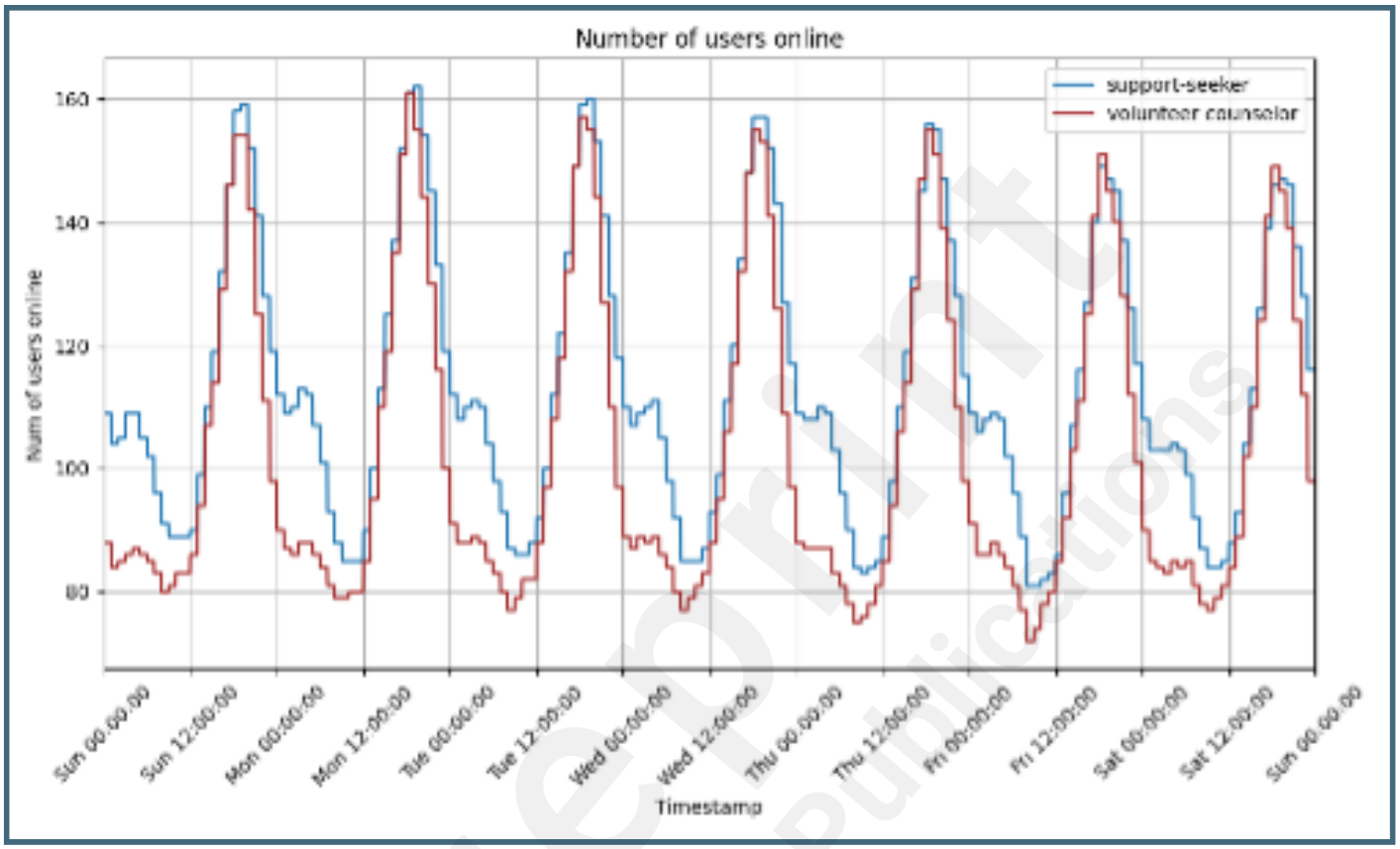
Supplementary Files

Figures

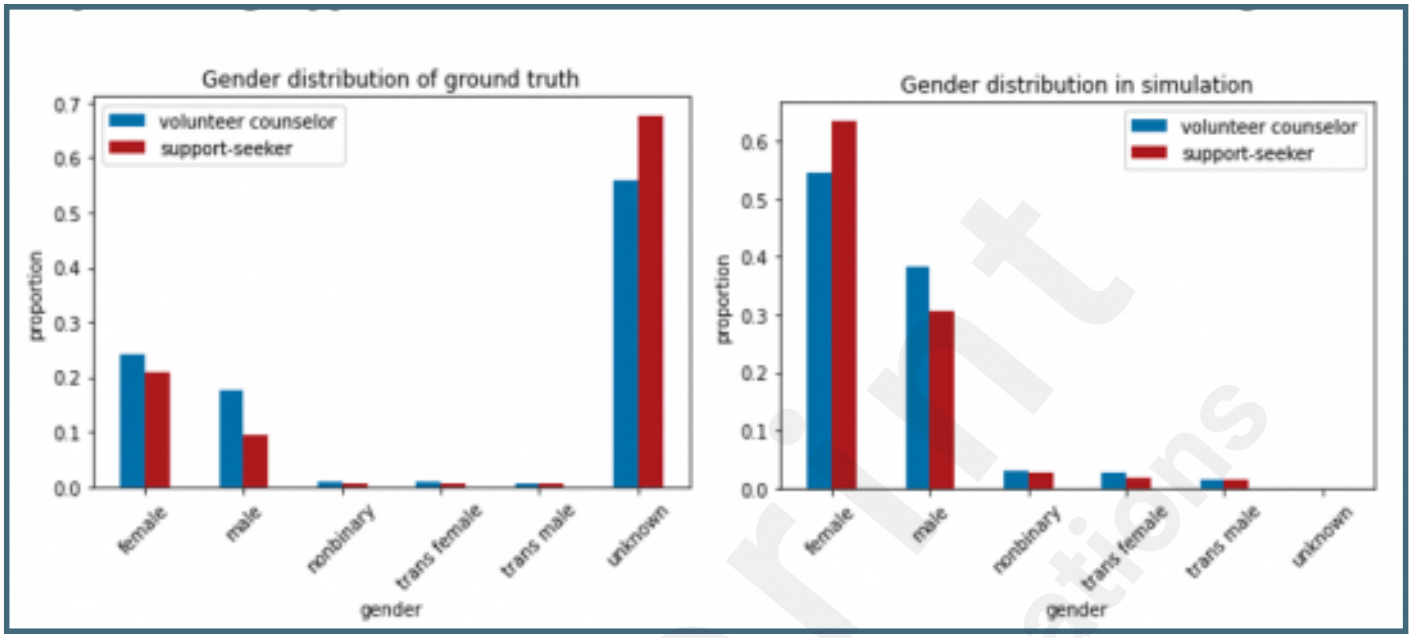
Workflow diagram showing inputs and outputs for the three stages of building and experimenting with agent-based simulation: (1) Building Agent-based Simulation, (2) Validation of Agent-Based Simulation, and (3) Virtual Experiments using the validated simulation.



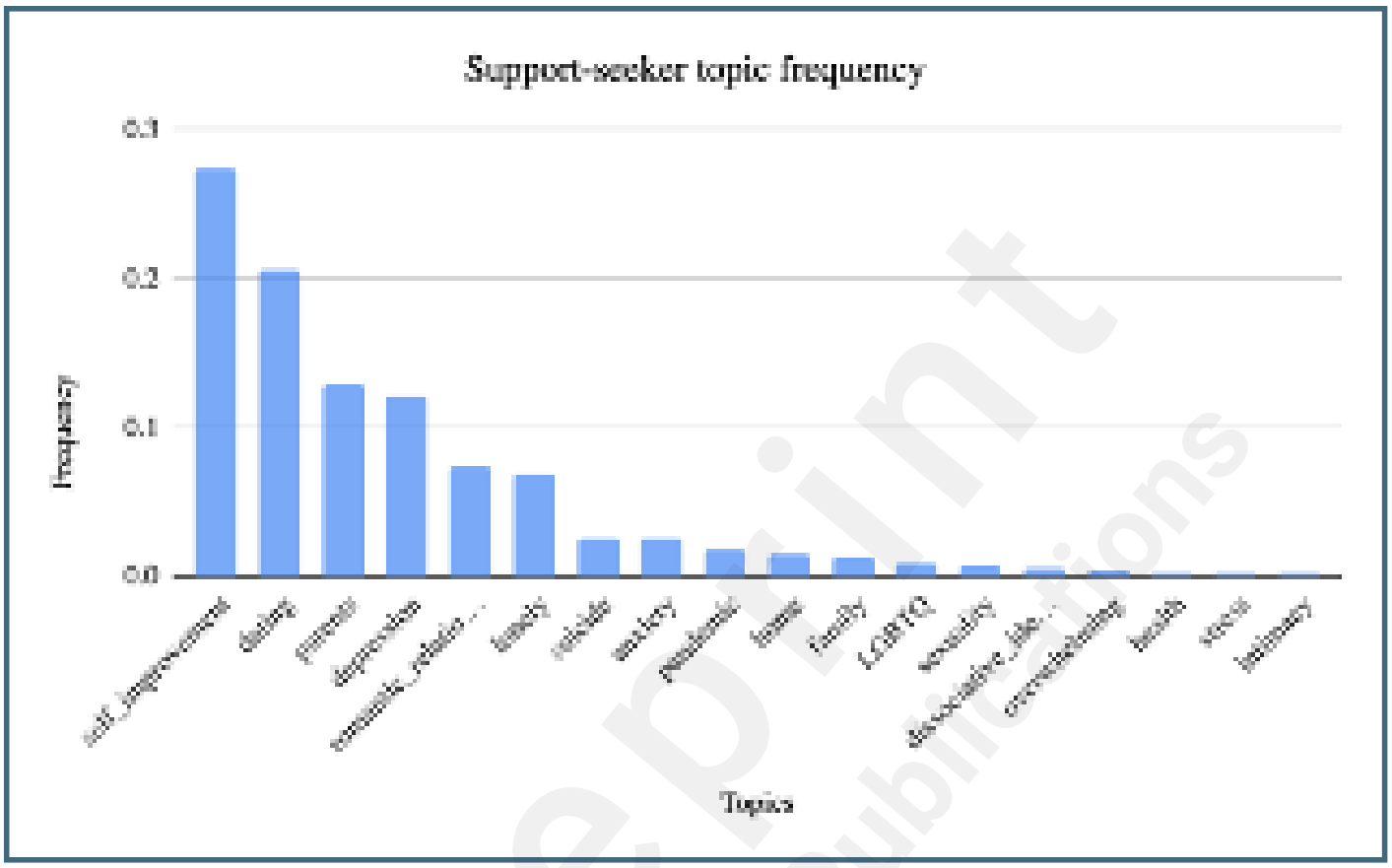
A line chart showing the number of online support-seekers and volunteer counselors in each simulation period. The number of support-seekers always exceeds the number of volunteer counselors, with support-seekers online at any given minute ranging between 81 and 162 (mean of 113.26, standard deviation of 22.56) and volunteer counselors online at any given minute ranging between 72 and 161 (mean of 102.49 and standard deviation of 25.07).



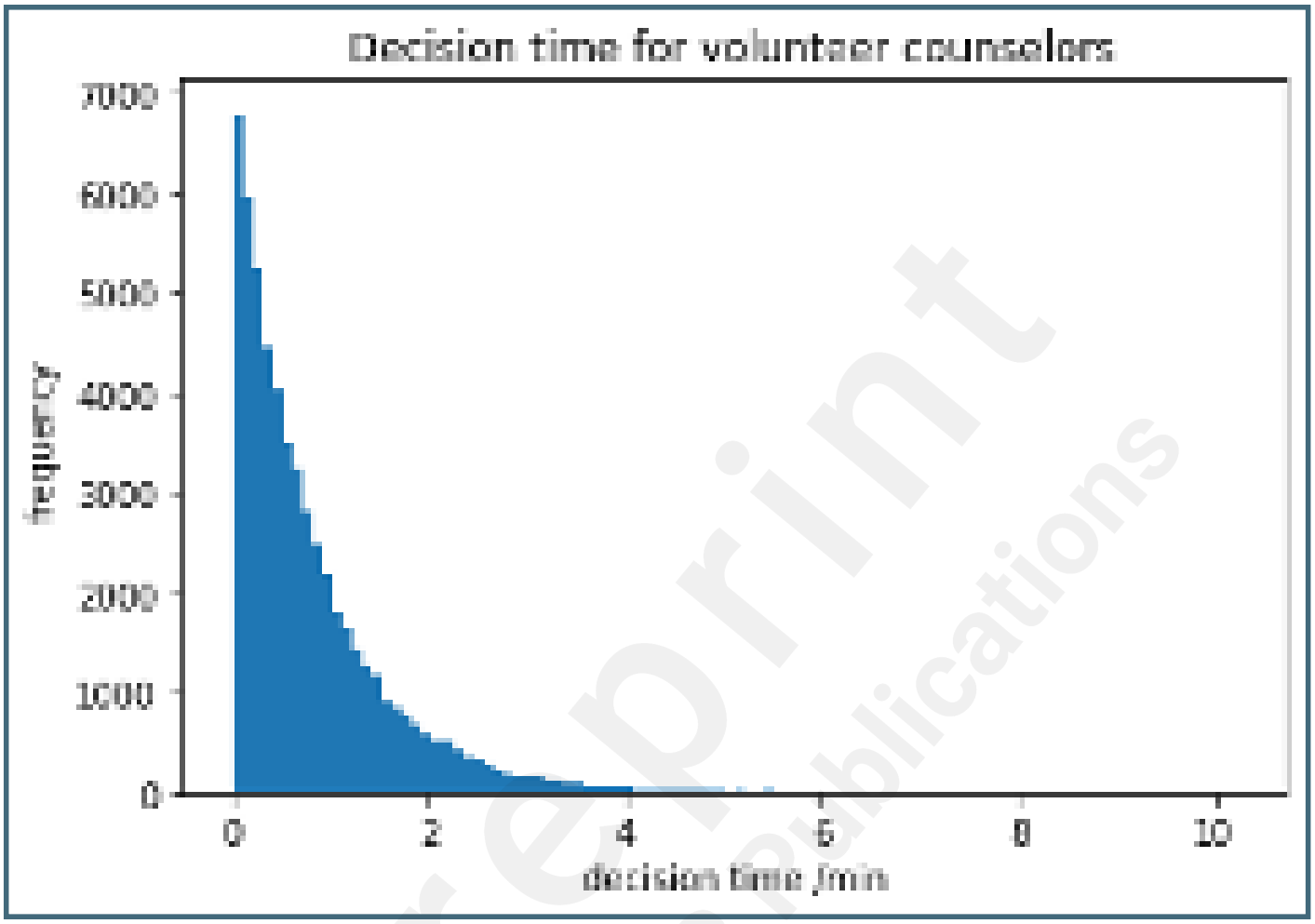
Gender distribution in reality (left) vs. our simulation (right). In the ground truth dataset, many support-seekers and counselors have unknown gender. In contrast, all agents are assigned gender in our simulation. We assign gender according to the gender distribution of known genders (female, male, nonbinary, trans female, trans male) in the ground truth dataset.



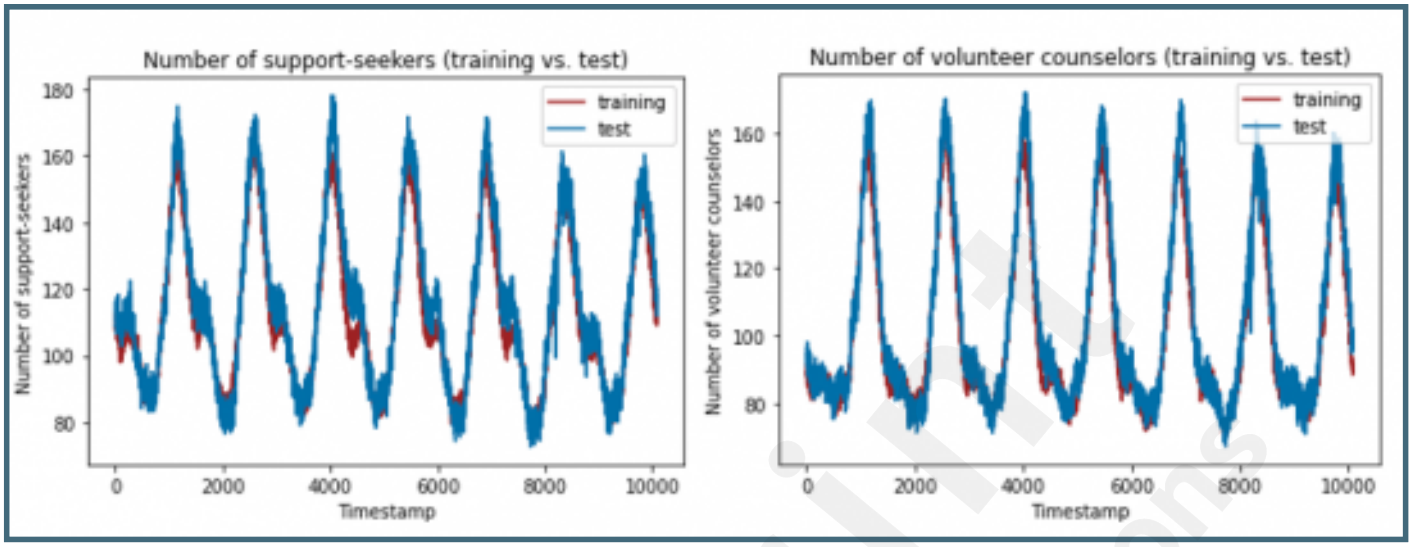
Distribution of topics among support-seekers in the real OMHC dataset. The most frequent topic discussed on the site is “self improvement”, followed by “dating”, “parents”, and “depression”. Our simulation’s frequency distribution is the same, with a Pearson correlation of 1.



Distribution of decision time of volunteer counselors, which is modeled by an exponential distribution with lambda of 1.25.



Distribution of number of support-seekers (left) and volunteer counselors (right) in training and test set. Pearson correlation of number of support-seekers between training and test set is 0.974 while Pearson correlation of the number of volunteer counselors between training and test set is 0.982.



Comparison of chat ratings (left) and blocking (right) between reality versus our simulation. Our simulation (shown in blue) has proportions of ratings from 1-star to 5-stars were, respectively, 14.96%, 4.64%, 6.08%, 9.26%, 65.06%, while ground truth (shown in red) proportions were 15.18%, 3.51%, 4.56%, 10.63%, and 66.12%. In terms of pairs that resulted in blocking, the ground truth proportion of blocked support-seeker and volunteer counselor pairs is 5.3%, compared to our simulation's proportion of 5.86%.