Understanding Patient Journeys with Telehealth: A Poisson-Factor-Marked Hawkes Process

Quan Zhang
Michigan State University, quan.zhang@broad.msu.edu

Yixuan Liu
Purdue University, liu3179@purdue.edu

Understanding Patient Journeys with Telehealth: A Poisson-Factor-Marked Hawkes Process

Completed Research Paper

Quan Zhang
Michigan State University
632 Bogue St, East Lansing
Michigan, U.S.A.
quan.zhang@broad.msu.edu

Yixuan Liu
China Europe International Business School
699 Hongfeng Rd, Pudong
Shanghai, China
yixuanliu@ceibs.edu

Abstract

The emerging telehealth platforms connect patients with physicians using telecommunication technologies and are transforming the traditional healthcare delivery process. Meanwhile, patient care journeys spreading across online and offline health service channels call for new research methodologies. Using a dataset from a telehealth platform, we develop a novel Poisson-factor-marked Hawkes process to model such a journey and quantify the mutual-modulating effects of various patient activities. Our estimation results demonstrate the disparate impacts of the patient's health conditions and physician characteristics on choosing care channels. Taking advantage of the self-generation property of our model, we simulate policy and strategic interventions, which highlights the practical value of the proposed model and offers implications for better patient routing and service design for telehealth platforms.

Keywords: Multichannel healthcare, correlated Poisson factor analysis, generative model

Introduction

Facilitated by the emerging telehealth platforms, patients are provided with an online channel to connect with healthcare practitioners and to quickly access consultation, primary care, and specialty care services. As defined by HRSA (2020), telehealth (or telemedicine) "takes advantage of electronic information and telecommunication technologies to support long-distance clinical health care, patient and professional health-related education, public health, and health administration." The COVID-19 pandemic has forced the healthcare industry to leap into the digital future, significantly influencing the healthcare delivery process. An emerging "Digital-First" health care approach flips the traditional in-person-centric healthcare delivery model into one that prioritizes digital engagement and underscores the need for telehealth (Gartner 2020). The global telehealth market is projected to grow from 90 billion in 2021 to 636 billion in 2028 with a dramatic increase in the adoption rate (Fortune Business Insights 2021).

Telehealth transforms traditional healthcare delivery by introducing an additional online channel for quick and low-cost healthcare delivery (Bestsennyy et al. 2021). The accessibility and convenience of telemedicine have demonstrated the potential for more cost-effective medical services (Zeltzer et al. 2021). To advance the knowledge of how patients choose between the traditional offline channel and the emerging online telehealth channel, and more importantly, its implication on the patient health outcome, we propose a modeling framework with considerable latitude in temporal dimensions and details of each medical care interaction.

Firstly, a patient's health condition evolves over time and is affected by all the previous health-related interventions. Ignoring the patient journey may result in biased estimates or important factors being overlooked.

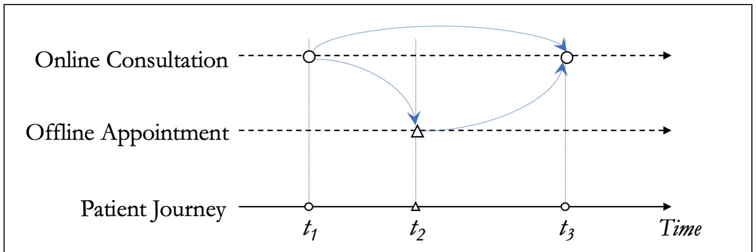

Figure 1. An Example Patient Journey

This gap motivates us to model a patient journey as a point process. Specifically, we see the journey of a patient as a collection of events (points) falling in the time space. Figure 1 shows an example patient journey in our study where two types of events are accounted for: online consultations (through a telehealth platform) and offline office visits. The patient first consults a doctor online at time t_1, and then makes an office visit appointment at time t_2 afterward. As the patient monitors her condition, she decides to go with the online channel for the follow-up visit, which occurs at t_3. Note that, the online visit at time t_3 can be driven by both the online visit at t_1 and the offline visit at t_2. Solely focusing on the office visit and estimating its impact on the online visit at t_3 may give rise to bias from overlooking the impact of the first online communication.

To properly model the patient journey, the effects of events within and across both channels are captured. Specifically, *self-* (or *mutual-*) *exciting effect* represents the positive impact of past events on the same (or a different) type of future events whose arrivals are thus accelerated. Moreover, previous events can also inhibit future ones. For instance, a telehealth visit about a chronic disease can substitute for future traditional clinical visits (Ayabakan et al. 2020). Similarly, an office visit with proper diagnosis and treatment can cure the patient, making future medical visits unnecessary. Thus, it is also critical to capture the *self-* and *mutual-inhibiting effects* across both channels. In particular, the self- (or mutual-) inhibiting effect is the negative impact of past events on the same (or a different) type of future events whose arrivals are thus decelerated. We collectively refer to the self-exciting and inhibiting effect as a *self-modulating effect*, and the mutual-exciting and inhibiting effect as a *mutual-modulating effect*.

The patient journey with telehealth is also characterized by the content of consulting conversations and physician characteristics. During consultations, patients describe their symptoms and concerns. Subsequently, physicians provide medical advice based on their clinical experience. Moreover, it has been reported that physician seniority can be correlated with inappropriate diagnosis (Young et al. 2020). This effect of physician seniority is further modulated by cultural factors, which also affect patients' trust in physicians' advice (Ju, Zhang, et al. 2020). Since medical service is a typical example of credence goods (Dulleck and Kerschbamer 2006), as suggested by the signaling theory (Spence 1978), the content of the dialogues and the physician experience can potentially signal the quality of the service. Thus, the rich details of patient-physician interactions are incorporated into the proposed model to understand its influence on patients' decisions on further medical care and the dynamics of patient conditions.

To this end, we propose Poisson-factor-marked Hawkes process to accommodate all the aforementioned modeling features. In particular, both self- and mutual- modulating effects between patients' online consultation and offline appointment behaviors are considered. In addition, we highlight the importance of capturing physician characteristics and conversation content for online encounters. The proposed model is applied to a dataset shared by an online telehealth service platform operating in China, which includes information of medical care processes related to gynecology and obstetrics. The dataset has several advantages. Firstly, the complete journey of patients who consult physicians online via the platform and make office-visit appointments to see a physician offline is recorded. Moreover, the consultation service takes the format of secure messaging through the platform, and the content of the conversation is included in the data. To model the patient journey, we use a correlated Poisson factor model to extract information from the consultation content, and a logistic model for the choice of the physician class in consultations. Utilizing both information, we model patient activity streams by a mutually modulating marked Hawkes process. We refer to the whole procedure as a Poisson-factor-marked Hawkes Process.

The proposed model offers a better characterization of the patient journey across both channels, as its prediction power outperforms several benchmark counting processes. The estimation results show that two factors can best describe the content of conversations: one represents non-urgent medical needs, and the other represents more urgent medical conditions that may require further attention. Based on this finding, the severity of patients is approximated by the weights on each factor which is input into the physician choice model and the Hawkes process. The Hawkes process estimations demonstrate the modulation impact of patient's condition severity and physician seniority. After a consultation about urgent concerns, patients tend to see a doctor offline instead of online. In addition, consulting a junior physician leads to a higher probability of future consultation and substitutes for a future appointment. However, consulting a senior physician reverses the effect. These results serve as a starting point for further quantifying the marginal effects and simulating various policy and strategic interventions.

Apart from improving the understanding of patient journeys, the proposed model has the important advantage of fully depicting the data-generating process. This property allows for broader model applications based on predictions and simulations. We showcase one model application to understand the impact of different initial encounters. In particular, we impose various initial events (consulting different doctors with various severity levels of health conditions or making an appointment) and simulate patient journeys after that. The simulation result further confirms our previous findings and points out the importance of proper patient routing in the utilization of different channels.

In summary, this paper makes three contributions. First, we bring in a Poisson-factor-marked Hawkes process to model patient-physician interactions through telehealth and offline channels and offer new insights about the modulating roles of physician characteristics and consultation content. This extends the literature on telemedicine. Second, the proposed model has several strengths. We introduce a correlation Poisson factor model to capture the association between patients' questions and physicians' responses. Besides, we allow for both exciting and inhibiting effects under the framework of the mutual-modulating Hawkes process. This contributes to the methodologies in the Bayesian probabilistic topic models and the applications of the Hawkes process. Third, taking advantage of the generative property of the proposed model, we utilize the model to simulate policy interventions. These applications provide practitioners with implications for better patient routing and service design for telehealth platforms.

Literature Review and Contribution

Three strands of literature relate to our work. First, burgeoning empirical research has studied the impact of telemedicine on channel utilization, costs, and healthcare outcomes. Second, extensive work develops Bayesian topic models or applies them to different domains. Third, a growing number of studies in the business-related fields use a Hawkes process to model disparate mutual effects. We discuss each in turn.

First, our paper contributes to the growing literature on telemedicine. An important consideration when investigating the impact of telemedicine is its impact on the utilization of other channels, including office visits and telephone contacts. Depending on the research contexts and research methodologies, previous studies draw various mixed conclusions. From the practitioner side, the survey has indicated a significant variation regarding telehealth adoption, with most physicians either with an adoption rate below 10% or above 90% (AMA 2022). While on the patient side, such a disparity also exists. Examining the telehealth procedures through videoconferencing, Ayabakan et al. (2020) document a complementary effect in inpatient settings for non-chronic patients; however, they also show a significant reduction in outpatient visits after a telehealth visit. Looking into videoconferencing as well, Shah et al. (2018) show that this form of virtual visit reduces in-person visits in accountable care organizations. Another widely adopted form of telehealth is secure messaging that is similar to the online consultation service provided by the telehealth platform we study. This service is also known as e-visits and is defined as "non-face-to-face patient-initiated communications through an online patient portal" (CMS 2020). Green et al. (2013) argue that e-visits have the potential to resolve the surging demand for physician services. Supporting this argument, Zhou et al. (2007) find a negative correlation between access to secure messaging and primary care office visits. However, North et al. (2014) document null effects. Conversely, Bavafa et al. (2018) discover that e-visits increase office visits, as the telehealth channel serves as a gateway to future offline visits. Most existing papers along this line study the context of primary care where patients have a long-term established relationship with their primary care physicians. However, the telehealth platform we study is a marketplace for patients to choose any physicians, which introduces heterogeneity among physician choices. Secondly, on top of examining the effect of the "last event" on an upcoming one, we study the dynamics of patient journeys and account for individual health conditions by the innovative modeling framework we construct.

Second, our work contributes to methodology in Bayesian probabilistic topic models. Probabilistic topic models, especially latent Dirichlet allocation (LDA) (Blei et al. 2003), have enjoyed tremendous success in various applications in management science (Bao and Datta 2014; Bellstam et al. 2021; Brandt et al. 2021; Lee et al. 2020; Pu et al. 2020). These models unveil common latent topics among a collection of documents and interpret semantic concepts of the topics by representative words. In particular, LDA finds the proportion of a document that originates from a topic. With the sum of the proportions equal to one, LDA does not capture the absolute relevance of documents to topics and is problematic if applied to our context. Specifically, a dialogue between a patient and a physician with high degrees of relevance to every topic cannot be distinguished by LDA from those with low degrees, and consequently, the information extracted cannot reflect how detailed the patient's questions are or how well the questions are addressed by the physician. To overcome this obstacle, substantive extensions have been made, such as the Gamma-Poisson model (Canny 2004) and the more general Poisson factor analysis (Zhou et al. 2012). However, all these approaches fail to model topic correlations among documents that are indispensable in our application since questions from a patient and answers from a physician are highly correlated. Therefore, we propose a correlated Poisson factor model to jointly analyze the consultation questions and answers. Moreover, we propose a variational inference algorithm that admits big data analysis and topic extraction from unseen consultations.

Third, our study extends applications of Hawkes processes to understanding both excitation and inhibition between the two healthcare channels, and our proposed model allows prediction and simulation of future events. The Hawkes process is originally used to study the arrival patterns of earthquakes (Hawkes 1971) and has been applied to many fields in natural, biomedical, and social sciences (Jankowiak and Gomez-Rodriguez 2017; Okawa et al. 2021; Reynaud-Bouret and Schbath 2010; Yang and Zha 2013) where the event arrival has a clustering pattern. Among research in management science, Xu et al. (2014) introduce the Hawkes process to studying exciting effects of multiple types of online advertisements on purchase conversion. Using Hawkes processes for self- and mutual-excitation, Daw et al. (2021) model service times in contact centers, and Mukherjee et al. (2022) investigate auto recall clustering. These studies assume homogeneity of events and do not account for the effects of specific event characteristics on the size of excitation. To relax this homogeneity assumption, one can involve event features in a Hawkes process as covariates. For example, Yoo et al. (2019) study content diffusion on Twitter and use counts of followers as covariates. Aggarwal et al. (2021) use a Hawkes process to understand idea generation with problems and ideas as covariates. However, these approaches do not account for the generating process of covariates and thus cannot be used to study the marginal effects of a preceding event or to simulate event streams with potential interventions. In stark contrast, our proposed Poisson-factor-marked Hawkes process is a generative model and admits both prediction and simulation of future events such that one can easily study the impact of an intervention on the event distributions. Specifically, we model the time and type (consultation or appointment) of a patient activity by a Hawkes process, the content of consultation by Poisson factors, and a patient's choice of the physician class in consultation by logistic regression. Moreover, our model allows both excitation and inhibition among events whereas all the aforementioned studies only model exciting effects.

Research Context and Data

The research context is a large online healthcare service platform operating through a mobile app in China. Founded in 2015, the platform operates in more than 70 cities in 17 provinces by 2019, with the majority in East and South China. The app is free for patients to register for seeking services. All licensed physicians are free to register and provide services through the platform. In China, almost all physicians are affiliated to hospitals or clinics. Thus, to further induce more providers onboard, the platform has partnered with hospitals in various provinces and cities. By 2018, the platform has collaborated with around 500 hospitals. More than 50,000 healthcare service providers have registered on the platform, serving nearly 10 million registered users. The healthcare system in China does not currently have a gatekeeping system. Unlike in the

	5% quantile	25% quantile	median	75% quantile	95% quantile
Age	21	26	30	34	47
# Consultations	1	1	1	2	6
# Appointments	1	2	4	8	20
Time interval (days)	0.19	2.98	10.28	29.30	124.83

Table 1. Summary of Patient Age, Events per Patient, and Time Intervals

U.S., patients in China can choose any types of physicians, including primary care physicians and specialists, to make appointments without referral. This also applies to the services provided on the platform.

The platform provides several online health-related services. The main services are online consultation and offline appointment. For the consultation service, a user can chat with a doctor of his or her choice through the platform. Doctors set their service prices and duration, which can last for one or more days. During the service period, a user and a doctor can exchange text messages, photos, and voice messages with each other. The content of their conversations can only be viewed by them but not by others. The platform charges a commission from doctors' consultation service. Note that, physicians will not make a formal clinical diagnosis and are not allowed to prescribe medications. Instead, during the conversations, physicians generally discuss potential causes of users' conditions, provide certain guidance on seeking further medical care, and answer other health-related questions. For the appointment service, users can make offline office visit appointments with doctors through the platform. The appointment schedules and fees are set by hospitals with which doctors are affiliated. The platform charges no fee for such a service.

The dataset shared by the platform include the activity streams from users in six cities of one province in South China, between March 2016 (when the platform started to operate in the focal province) and April 2019. For each user, we observe their gender, age, and registration time. The activity stream of a user include purchases of consultation and appointment services. For the consultation service, the dataset includes the time of purchase, the text content of the conversation, the provider information, and the fee. For appointments, we know the time a user makes an appointment and information of the provider. For all the providers, we observe their affiliations (hospitals), specialties (represented by the medical departments in the hospitals), and ranks. In Chinese healthcare system, there are four ranks for physicians, which depend on physicians' educational background, level of scholarship, and years of experience. Physicians of the top-two ranks are often called "experts" in China, with whom the appointment and other medical fees are higher. We refer to these physicians as senior physicians and the others as junior physicians.

In this study, we focus on 8,260 users from one city, who have interacted both online and offline with physicians from the department of gynecology and obstetrics. The app we collaborate with was the most popular telehealth platform in the focal city during our study period. It partners with most public primary hospitals and Obstetrics & Gynecology-specialized hospitals. The firm also gives the highest priority to developing services related to Obstetrics & Gynecology (OB/GYN). As a result, OB/GYN represents the most popular specialty on the platform. The dataset we use contains 52,125 appointments and 17,238 consultations where 13,221 consultations are provided by senior physicians and 4,017 by junior physicians. In Table 1, we summarize the patients' age at registration, the numbers of consultations and appointments for each patient, and the time interval between two events. The time intervals have a median value of 10.28 days and can be as large as over four months, likely suggesting a clustering pattern of the patient activities on the platform. In other words, a patient's consultations and appointments tend to concentrate in a relatively short period of time followed by a long period of inactivity. The potential event clustering is statistically testable by checking an optimal number of clusters. We conduct K-means clustering (other clustering methods also apply) for each of the 6,890 patients who have at least three activities (consultations and/or appointments) and use the gap statistic (Tibshirani et al. 2001) to determine the number of clusters. As a result, 5,085 patients have more than one cluster, which demonstrates the clustering pattern of patients' activities.

Poisson-Factor-Marked Hawkes Process

We propose a Poisson-factor-marked Hawkes Process to investigate patient journey dynamics. Specifically, we use a correlated Poisson factor analysis to extract information from the consultation content and accommodate correlation between questions and answers in a consultation. Furthermore, we use a logistic model for the choice of the physician class in consultation service. Lastly, we utilize the extracted information from consultations and the choices of the physician class as marks in a mutually modulating Hawkes process to model patient activity streams with a noticeable clustering pattern. The three components complete the data generating process of patient journeys.

Patient activities (consulting a physician and making an appointment) on the platform are patient-initiated events. We let (t_i, κ_i) denote the i-th event in temporal order with the event type κ_i occurring to a patient at time t_i. With $i = 1, 2, \ldots$, the patient's actions over time compose an event stream $\{(t_i, \kappa_i)\}_i$ where $0 \leq t_1 < \ldots < t_i < t_{i+1} < \ldots$ and $\kappa_i \in \{1, 2, \ldots, K\}$. In our empirical setting, each patient has $K = 2$ types of events: consulting a physician ($\kappa_i = 1$) and making an appointment ($\kappa_i = 2$). A consultation is characterized by the dialogue between the patient and the physician, i.e., the patient's questions and the physician's answers. Specifically, the consultation dialogue of event i is represented by the word count vector $\boldsymbol{y}_i^{(Q)} = (y_{i1}^{(Q)}, \ldots, y_{iV_Q}^{(Q)})$ of the questions and $\boldsymbol{y}_i^{(A)} = (y_{i1}^{(A)}, \ldots, y_{iV_A}^{(A)})$ of the answers where $V_Q = 97$ and $V_A = 141$ are the numbers of all keywords in questions and answers, respectively. It is not necessary that all the keywords appear in each consultation; we have $y_{iv}^{(Q)} = 0$ or $y_{iv}^{(A)} = 0$ or if keyword v of questions or answers does not appear in event i.

Correlated Poisson Factor Analysis of Consultations

We propose a correlated Poisson factor model to extract information from consultation and use variational inference for an easy factor score estimation of unseen consultations. Suppose an event i of consultation has the count vectors $\boldsymbol{y}_i^{(Q)}$ of the patient questions and $\boldsymbol{y}_i^{(A)}$ of the physicians' answers. With S being a prespecified number of factors (topics) in the dialogue, we assume that the word counts $\boldsymbol{y}_i^{(Q)}$ and $\boldsymbol{y}_i^{(A)}$ are generated by the following Poisson factor model:

$$
\begin{aligned}
&y_{iv}^{(Q)} \sim \text{Poisson}(\textstyle\sum_{s=1}^{S} z_{is}^{(Q)} \phi_{sv}^{(Q)}), v = 1, \ldots, V_Q, \quad y_{iv}^{(A)} \sim \text{Poisson}(\textstyle\sum_{s=1}^{S} z_{is}^{(A)} \phi_{sv}^{(A)}), v = 1, \ldots, V_A, \\
&\phi_{sv}^{(Q)} > 0,\ \textstyle\sum_{v=1}^{V_Q} \phi_{sv}^{(Q)} = 1,\ v = 1, \ldots, V_Q, \quad \phi_{sv}^{(A)} > 0,\ \textstyle\sum_{v=1}^{V_A} \phi_{sv}^{(A)} = 1,\ v = 1, \ldots, V_A, \text{ for } s = 1, 2, \ldots, S, \\
&z_{is}^{(Q)} > 0, z_{is}^{(A)} > 0 \text{ for } s = 1, 2, \ldots, S, \quad (z_{i1}^{(Q)}, z_{i2}^{(Q)}, \ldots, z_{iS}^{(Q)}, z_{i1}^{(A)}, z_{i2}^{(A)}, \ldots, z_{iS}^{(A)}) \sim \text{log-Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \qquad (1)
\end{aligned}
$$

Defining $\boldsymbol{\phi}_s^{(Q)} := (\phi_{s1}^{(Q)}, \phi_{s1}^{(Q)}, \ldots, \phi_{sV_Q}^{(Q)})$ and $\boldsymbol{\phi}_s^{(A)} := (\phi_{s1}^{(A)}, \phi_{s1}^{(A)}, \ldots, \phi_{sV_A}^{(A)})$, $\boldsymbol{\phi}_s^{(Q)}$ and $\boldsymbol{\phi}_s^{(A)}$ are the factor loading vector of $\boldsymbol{y}_i^{(Q)}$ and $\boldsymbol{y}_i^{(A)}$; with the sum-to-one constraint on the factor loading vectors, each element $\phi_{sv}^{(Q)}$ (or $\phi_{sv}^{(A)}$) is between 0 and 1 and represents the relative importance of word v in factor s of the questions (or answers). Consequently, each factor s can be interpreted by looking at the loading vectors $\boldsymbol{\phi}_s^{(Q)}$ and $\boldsymbol{\phi}_s^{(A)}$, where representative words v of factor s in questions (or answers) have large weights $\phi_{sv}^{(Q)}$ (or $\phi_{sv}^{(A)}$). Note that $\boldsymbol{\phi}_s^{(Q)}$ and $\boldsymbol{\phi}_s^{(A)}$ are global parameters that remain unchanged across consultations.

Writing $\boldsymbol{z}_i^{(Q)} := (z_{i1}^{(Q)}, z_{i2}^{(Q)}, \ldots, z_{iS}^{(Q)})$ and $\boldsymbol{z}_i^{(A)} := (z_{i1}^{(A)}, z_{i2}^{(A)}, \ldots, z_{iS}^{(A)})$, $\boldsymbol{z}_i^{(Q)}$ and $\boldsymbol{z}_i^{(A)}$ are the factor score vector of the questions and the answers, respectively, in the consultation of event i, and each element $z_{is}^{(Q)}$ (or $z_{is}^{(A)}$) characterizes the relevance of the consultation to factor s of the questions (or answers). The factor scores $\boldsymbol{z}_i^{(Q)}$ and $\boldsymbol{z}_i^{(A)}$ depend on specific information of event i of consultation. We further assume the factor scores of questions and answers follow a multivariate log-normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (or equivalently $(\log z_{i1}^{(Q)}, \ldots, \log z_{iS}^{(Q)}, \log z_{i1}^{(A)}, \ldots, \log z_{iS}^{(A)}) \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) to accommodate the correlation of factors within and between questions and answers in a consultation.

We use the correlated Poisson factor analysis for its easy interpretation and generation of the factor scores that can be used to model and simulate the occurrences of patients activities. Other generative language models with an easy interpretation may also apply. We estimate the Poisson factor model by variational inference with different numbers of factors S, evaluate the log-likelihood, and use the Bayesian information criterion (BIC) to determine S (as shown in Table 2, shortly). Finally, we set the number of factors $S = 2$ and interpret the two factors as different types of need for medical service. In fact, as shown in the estimation section, we conclude that factor $s = 1$ represents the routine/wellness health-related questions from patients, while factor $s = 2$ characterizes relatively urgent inquiries. In the remainder of this section, we present the rest of the model by letting $S = 2$.

Choice of Physician Class for Consultation

A patient chooses between junior and senior physicians if she has decided to request an online consultation. Let $w_i = 0$ or 1 denote that the patient consults a junior or a senior physician in event i of consultation. The choice w_i may depend on the patient's health condition at that moment, which is self-evaluated and hardly known by healthcare providers before the service. To circumvent the difficulty in assessing the patient's condition, we find an approximation using the correlated Poisson factor analysis. Specifically, a positive valued factor score $z_{is}^{(Q)}$ quantifies to what degree a patient's question in consultation i is related to factor s; a larger value of $z_{is}^{(Q)}$ implies that the questions are more specifically asked with respect to factor s. With factor $s = 1$ representing routine healthcare demands and factor $s = 2$ urgent inquiries, $\frac{z_{i2}^{(Q)}}{z_{i1}^{(Q)}+z_{i2}^{(Q)}}$ is between 0 and 1 and can be used as a proxy of the patient's health condition. In consequence, we model the choice of the physician class w_i by a logistic regression

$$\log \frac{p(w_i = 1)}{1 - p(w_i = 1)} = \gamma_0^{(n)} + \gamma_1 \frac{z_{i2}^{(Q)}}{z_{i1}^{(Q)} + z_{i2}^{(Q)}} \tag{2}$$

where $\gamma_0^{(n)} \sim \text{Normal}(\mu_{\gamma_0}, \sigma_{\gamma_0}^2)$ is a random intercept quantifying patient n's baseline preference when choosing the physician class and γ_1 is a fixed slope that does not differ among patients.

The formulation of (2) evaluates the impact of urgency/health condition on the choice of the physician class. The prediction of w_i by $\frac{z_{i2}^{(Q)}}{z_{i1}^{(Q)}+z_{i2}^{(Q)}}$ seems to contradict the temporal order of patient actions, as the factor scores $z_{i1}^{(Q)}$ and $z_{i2}^{(Q)}$ are inferred from the questions after w_i is determined. In fact, $\frac{z_{i2}^{(Q)}}{z_{i1}^{(Q)}+z_{i2}^{(Q)}}$ as an approximation of the health condition is used as a "pivot quantity" that is unaffected by consultation. In other words, the questions have been already in the mind of the patient and her degree of urgency/health condition evokes the demand of consulting a junior or senior physician.

We particularly focus on the choice between junior and senior physicians for the online consultation service for the following reasons. First, we take the platform's perspective and aim to understand patient journeys to facilitate better service design and patient routing. Specifically, taking advantage of the self-generation property of the proposed model, we study by simulation how an arbitrary event affects a patient's future behaviors. Second, the rich text data from the consultation service allows a higher flexibility in modeling patient's online activity stream in detail. Lacking such granular data for offline interactions between patients and doctors or offline medical records, we do not model choices of the physician classes in offline visits. Hereby, we have defined the generating processes of marks, including the Poisson factors of dialogues and the choices of the physician class in consultations. Next we utilize the marks of past events to model arrivals of future ones by defining intensity functions of a mutually modulating marked Hawkes process.

Marked Hawkes Process for Mutual Modulation

To account for the effects of a patient's past actions on future ones, we define the patient's event history $\mathcal{H}(t-) = \{(t_i, \kappa_i) \,|\, i : t_i < t\}$ that includes the time and type of an event up to but not including time t. Analogously, $\mathcal{M}(t-) := \{(w_i, \boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)}) \,|\, i : t_i < t, \kappa_i = 1\}$ denotes the set of marks before time t, including the patient's choice of the physician class and factor scores of dialogues in consultations. Given $\mathcal{H}(t-)$ and $\mathcal{M}(t-)$, we formulate a marked Hawkes process by intensity functions λ_k at time t with $k = 1$ and 2 for consultation and making an appointment, respectively. Concretely, the probability of an event of type k occurring in an infinitesimal time interval $[t, t+dt)$ is equal to $\lambda_k(t \,|\, \mathcal{H}(t-), \mathcal{M}(t-))dt$. Note that λ_k depends on the entire history of past events, including the inter-events time intervals, time elapsed since the previous event, and the marks related to previous events.

We use a softplus function $\mathcal{F}(x) = \log(1 + e^{cx})/c$ with $c = 10$ as a smooth approximation of $\max(0, x)$. Suppose there are N patients. Given patient n's event history $\mathcal{H}(t-)$ and marks $\mathcal{M}(t-)$ before time t, her intensity of consultation ($k=1$) or making an appointment ($k=2$) at time t, $\lambda_k(t \mid \mathcal{H}(t-), \mathcal{M}(t-))$, is

$$\begin{aligned}
\lambda_k(t \mid \mathcal{H}(t-), \mathcal{M}(t-)) =& \mathcal{F}\Big(\alpha_k^{(n)} + \sum\nolimits_{i:t_i<t, \kappa_i=1} \psi_{k1}(t, t_i, w_i, \boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)}) + \sum\nolimits_{i:t_i<t,\kappa_i=2} \psi_{k2}(t, t_i)\Big), \\
\psi_{k1}(t, t_i, w_i, \boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)}) =& \Big(\beta_{k10} + \beta_{k11} w_i + \beta_{k12}(z_{i1}^{(Q)} + z_{i2}^{(Q)}) + \beta_{k13} \frac{z_{i2}^{(Q)}}{z_{i1}^{(Q)} + z_{i2}^{(Q)}} + \\
& (\beta_{k14} + \beta_{k15} w_i)(z_{i1}^{(A)} + z_{i2}^{(A)}) + (\beta_{k16} + \beta_{k17} w_i) \frac{z_{i2}^{(A)}}{z_{i1}^{(A)} + z_{i2}^{(A)}}\Big) e^{-\eta_{k1}(t-t_i)}, \\
\psi_{k2}(t, t_i) =& \beta_{k20} e^{-\eta_{k2}(t-t_i)},
\end{aligned} \tag{3}$$

where $\alpha_k^{(n)}$ is patient n's random baseline intensity of an upcoming event of type $k=1$ or 2 and independently follows log-Normal(m_k, σ_k^2) for each patient $n = 1, 2, \ldots, N$. $\psi_{k1}(t, t_i, w_i, \boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)})$ quantifies the decaying impact of a past consultation at time t_i on a potentially upcoming event of type k at time t and it depends on the content of the consultation that is reflected by $z_{i1}^{(Q)} + z_{i2}^{(Q)}$, $\frac{z_{i2}^{(Q)}}{z_{i1}^{(Q)} + z_{i2}^{(Q)}}$, $z_{i1}^{(A)} + z_{i2}^{(A)}$, and $\frac{z_{i2}^{(A)}}{z_{i1}^{(A)} + z_{i2}^{(A)}}$. Specifically, $z_{i1}^{(Q)} + z_{i2}^{(Q)} > 0$ is the total relevance of the patient's questions to routine and urgent healthcare demands and thus interpreted as how detailed the questions are in the consultation. Analogously, $z_{i1}^{(A)} + z_{i2}^{(A)} > 0$ is interpreted as how specifically the physician has addressed the patient's concern. Simply, $z_{i1}^{(Q)} + z_{i2}^{(Q)}$ and $z_{i1}^{(A)} + z_{i2}^{(A)}$ characterize the length of the questions and answers. As explained, $\frac{z_{i2}^{(Q)}}{z_{i1}^{(Q)} + z_{i2}^{(Q)}}$ between 0 and 1 approximates the patient's health condition and a larger value indicates a worse condition. $\frac{z_{i2}^{(A)}}{z_{i1}^{(A)} + z_{i2}^{(A)}}$ is the proportion of the physician's answers/advice on the urgent inquiries of the patient and thus can reflect how the patient's health condition is evaluated by the physician; a larger value may indicate a more severe condition from the physician's perspective. The coefficients $\beta_{k1h} \in \mathbb{R}$, $h = 0, 1, \ldots, 7$, quantify the effects of the information in the past event i of consultation on the instantaneous probability of an upcoming event of type k, and these effects exponentially decay over time with a decay rate $\eta_{k1} > 0$. In addition, $\psi_{k2}(t, t_i)$ is the decaying impact of a past appointment at time t_i on an upcoming event of type k at time t with an instantaneous effect $\beta_{k20} \in \mathbb{R}$ and a decay rate $\eta_{k2} > 0$. The softplus function $\mathcal{F}$ guarantees a positive value of the intensity functions (Mei and Eisner 2017). To be estimated are global parameters β's, η's, m_k and σ_k^2, $k = 1, 2$ that are unchanged among patients. The random baseline intensities $\{\alpha_k^{(n)}\}_{n,k}$ are also estimable when necessary, such as in individual-level studies.

Model Estimation

The logistic model for the choice of the physician class as in equation (2) is estimated by the maximizing the likelihood and is omitted in our paper. In this section, we briefly discuss the inference of the correlated Poisson factor model and the marked Hawkes process with technical details omitted due to the page limit. Concretely, we use variational inference to estimate the Poisson factor model by maximizing the evidence lower bound

$$\max \mathbb{E}\left[\log \frac{p_1(\boldsymbol{\mu}) p_2(\boldsymbol{\Sigma}) \prod_{i:\kappa_i=1} p_{\boldsymbol{\phi}}(\boldsymbol{y}_i^{(Q)}, \boldsymbol{y}_i^{(A)} \mid \boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)}) p(\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}{q_1(\boldsymbol{\mu}) q_2(\boldsymbol{\Sigma}) \prod_{i:\kappa_i=1} q(\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)} \mid \boldsymbol{y}_i^{(Q)}, \boldsymbol{y}_i^{(A)})}\right] \tag{4}$$

with each component formulated by

$$\begin{aligned}
p_{\boldsymbol{\phi}}(\boldsymbol{y}_i^{(Q)}, \boldsymbol{y}_i^{(A)} \mid \boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)}) &= \Big(\prod\nolimits_{v=1}^{V_Q} \text{Poisson}(y_{iv}^{(Q)}; \sum_{s=1}^{S} z_{is}^{(Q)} \phi_{sv}^{(Q)})\Big)\Big(\prod\nolimits_{v=1}^{V_A} \text{Poisson}(y_{iv}^{(A)}; \sum_{s=1}^{S} z_{is}^{(A)} \phi_{sv}^{(A)})\Big), \\
p(\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \text{log-Normal}([\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)}]; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \\
q(\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)} \mid \boldsymbol{y}_i^{(Q)}, \boldsymbol{y}_i^{(A)}) &= \text{log-Normal}\big(\boldsymbol{z}_i^{(Q)}; \boldsymbol{\mu}_Q(\boldsymbol{y}_i^{(Q)}), \text{diag}(\boldsymbol{\sigma}_Q(\boldsymbol{y}_i^{(Q)})^2)\big) \times \\
& \quad \text{log-Normal}\big(\boldsymbol{z}_i^{(A)}; \boldsymbol{\mu}_A(\boldsymbol{y}_i^{(A)}), \text{diag}(\boldsymbol{\sigma}_A(\boldsymbol{y}_i^{(A)})^2)\big)
\end{aligned}$$

where $\boldsymbol{\mu}_Q$, $\boldsymbol{\sigma}_Q$, $\boldsymbol{\mu}_A$, and $\boldsymbol{\sigma}_A$ are feed-forward neural networks that take inputs of length V_Q or V_A and output an S-dimensional real vector. Moreover, $p_1(\boldsymbol{\mu})$ and $p_2(\boldsymbol{\Sigma})$ are standard multivariate normal priors on $\boldsymbol{\mu}$ and

	$S = 1$	$S = 2$	$S = 3$	$S = 4$
log-likelihood	-8.12×10^4	-7.45×10^4	-7.28×10^4	-7.15×10^4
BIC	3.63×10^7	3.59×10^7	3.67×10^7	3.74×10^7

Table 2. Log-likelihood and BIC for the Correlated Poisson Factor Model with Different S

the Cholesky factor of $\mathbf{\Sigma}$, and $q_1(\boldsymbol{\mu})$ and $q_2(\mathbf{\Sigma})$ are their Gaussian variational distributions whose means and covariances are to be estimated. The expectation in (4) is taken with respect to $\boldsymbol{\mu} \sim q_1(\boldsymbol{\mu})$, $\mathbf{\Sigma} \sim q_2(\mathbf{\Sigma})$, and $(\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)}) \sim q(\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)} \,|\, \boldsymbol{y}_i^{(Q)}, \boldsymbol{y}_i^{(A)})$ and evaluated by reparameterization (Blei et al. 2017; Kingma and Welling 2014). The evidence lower bound is maximized by (stochastic) gradient descent with respect to $\boldsymbol{\phi}_s^{(Q)}$, $\boldsymbol{\phi}_s^{(A)}$, the parameters of neural networks $\boldsymbol{\mu}_Q$, $\boldsymbol{\mu}_A$, $\boldsymbol{\sigma}_Q$, and $\boldsymbol{\sigma}_A$, and the means and covariances of Gaussian distributions $q_1(\boldsymbol{\mu})$ and $\boldsymbol{q}_2(\mathbf{\Sigma})$. Note that $q(\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)} \,|\, \boldsymbol{y}_i^{(Q)}, \boldsymbol{y}_i^{(A)})$ approximates the joint posterior of $\boldsymbol{z}_i^{(Q)}$ and $\boldsymbol{z}_i^{(A)}$. In stark contrast to Bayesian inference based on MCMC that is used for inference of LDA (Blei et al. 2003) and the Poisson factor model (Zhou et al. 2012) that does not account for document correlation, the closed-form expression of $q(\boldsymbol{z}_i^{(Q)}, \boldsymbol{z}_i^{(A)} \,|\, \boldsymbol{y}_i^{(Q)}, \boldsymbol{y}_i^{(A)})$ allows predictions of $\boldsymbol{z}_i^{(Q)}$ and $\boldsymbol{z}_i^{(A)}$ of an unseen consultation i with no need to re-run the inference algorithm. This is imperative in big data applications where training a model is computationally expensive.

We find the parameters β's and η's in the marked Hawkes process (3) by maximizing log-likelihoods. Suppose we observe a patient's activities between time 0 and T. The log-likelihood of the patient's event times and types given the marks $\mathcal{M}(T)$ is

$$\ell(\mathcal{H}(T) \,|\, \mathcal{M}(T)) = \sum_{i: t_i \le T} \log \lambda_{\kappa_i}(t_i \,|\, \mathcal{H}(t_i-), \mathcal{M}(t_i-)) - \int_0^T \lambda(t \,|\, \mathcal{H}(t-), \mathcal{M}(t-)) dt \tag{5}$$

where $\lambda(t \,|\, \mathcal{H}(t-), \mathcal{M}(t-)) = \lambda_1(t \,|\, \mathcal{H}(t-), \mathcal{M}(t-)) + \lambda_2(t \,|\, \mathcal{H}(t-), \mathcal{M}(t-))$ represents the intensity of either a consultation or an appointment made at time t. We estimate parameters of the Hawkes process by maximizing the summation of log-likelihoods of all patients after marginalizing out the random baselines $\{\alpha_1^{(n)}, \alpha_2^{(n)}\}_n$ by Monte Carlo integration with reparameterization of log-normal distributions. All the optimization is implemented in PyTorch (Paszke et al. 2017) and starts from different random values of the parameters; we report results based on the parameter estimates that deliver the largest likelihood among candidate solutions. Nonparametric bootstrapping is used for uncertainty quantification of β's and η's in the intensity functions.

Results

We are interested in quantifying the effect of a patient seeking online/offline medical service on such decisions in the future. On top of that, we want to understand how physician seniority and the content of conversations affect a user's choice between the two channels of healthcare. Moreover, we compare different benchmark models with the proposed Poisson-factor-marked Hawkes process to demonstrate the critical roles of each model component.

Estimation of the Correlated Poisson Factor Analysis

We first present the estimation of the correlated Poisson factor model for consultation conversations. We evaluate the log-likelihood and use the Bayesian information criterion (BIC) to determine the number of factors S. As shown in Table 2, the log-likelihood remarkably decreases as the number of factors increases from one to two, and slightly goes down when S gets bigger. When $S = 2$, BIC achieves a minimum value among the models compared. Eventually, We set the number of factors $S = 2$ and report the keywords in each factor of the questions and answers and their weights $\{\phi_{sv}^{(Q)}\}_{s,v}$ and $\{\phi_{sv}^{(A)}\}_{s,v}$ in Table 3. For brevity, we only show the keywords whose weights are greater than 1%, and these keywords account for 69% or more of the information conveyed by the corresponding factors.

The top five keywords are *baby*, *breastfeeding*, *medication*, *now*, and *have a look/wait* for factor $s = 1$

Question factor $s = 1$		Question factor $s = 2$		Answer factor $s = 1$		Answer factor $s = 2$	
Keyword v	Weight $\phi_{sv}^{(Q)}$	Keyword v	Weight $\phi_{sv}^{(Q)}$	Keyword v	Weight $\phi_{sv}^{(A)}$	Keyword v	Weight $\phi_{sv}^{(A)}$
baby	0.1472	examination	0.1406	medication	0.0840	examination	0.1420
breastfeeding	0.0692	hospital	0.0622	baby	0.0802	menstruation	0.0457
medication	0.0570	appointment	0.0514	have a look/wait	0.0460	hospital	0.0445
now	0.0532	menstruation	0.0504	breastfeeding	0.0391	pregnant	0.0374
have a look/wait	0.0505	right now	0.0445	infect	0.0309	uterus	0.0345
cough	0.0291	medication	0.0345	comfortable	0.0274	abdomen	0.0325
poop	0.0287	pregnant	0.0339	perhaps	0.0260	medication	0.0324
cold	0.0250	result	0.0336	examination	0.0226	appointment	0.0287
severe	0.0243	surgery	0.0280	poop	0.0219	surgery	0.0271
yesterday	0.0236	normal	0.0240	cold	0.0216	bleeding	0.0256
comfortable	0.0213	problem	0.0231	hospital	0.0213	vaginal	0.0253
hospital	0.0204	vaginal	0.0213	oral	0.0199	perhaps	0.0237
examination	0.0204	bleeding	0.0209	problem	0.0196	normal	0.0236
particles	0.0199	treat	0.0182	allergy	0.0194	now	0.0199
before	0.0197	influence	0.0181	virus	0.0190	treat	0.0195
abdomen	0.0196	perhaps	0.0175	treat	0.0188	have a look/wait	0.0188
allergy	0.0192	uterus	0.0173	normal	0.0165	problem	0.0167
normal	0.0186	tomorrow	0.0172	particles	0.0160	result	0.0162
fever	0.0178	severe	0.0166	now	0.0157	comfortable	0.0161
diarrhea	0.0170	before	0.0155	replenish	0.0156	ovary	0.0145
influence	0.0169	baby	0.0153	appointment	0.0155	abortion	0.0131
sleep	0.0168	comfortable	0.0151	fever	0.0150	influence	0.0121
probiotics	0.0154	have a look/wait	0.0135	mental state	0.0144	tomorrow	0.0109
snot	0.0143	abortion	0.0127	probiotics	0.0141	baby	0.0104
problem	0.0128	yesterday	0.0120	severe	0.0137		
tomorrow	0.0126	ovary	0.0111	symptom	0.0136		
perhaps	0.0123	progesterone	0.0107	influence	0.0134		
natural delivery	0.0110	abdomen	0.0103	cough	0.0128		
oral	0.0108			skin	0.0128		
infect	0.0104			pregnant	0.0106		
Total weights	83.5%		78.9%		71.7%		69.1%

Table 3. Top Keywords and Weights for Questions and Answers in Consultation

of questions and *examination*, *hospital*, *appointment*, *menstruation*, and *right now* for factor $s = 2$ of questions. It is reasonable to conclude that factor $s = 1$ corresponds to patients' daily/routine needs of healthcare while factor $s = 2$ represents urgent inquiries or demands of urgent care[1]. Furthermore, the top five keywords are *medication*, *baby*, *have a look/wait*, *breastfeeding*, and *infect*, and for factor $s = 1$ of answers and *examination*, *menstruation*, *hospital*, *pregnant*, , and *uterus* for factor $s = 2$ of answers. Therefore, factors $s = 1$ and 2 represent physicians' answers/advice addressing the corresponding factors of patients' questions.

With factor scores $z_{is}^{(Q)}$ and $z_{is}^{(A)}$, $s = 1, 2$, quantifying the degrees of how relevant the consultation i is to factor s of questions and answers, the previous claims can be justified: $z_{i1}^{(Q)} + z_{i2}^{(Q)}$ and $z_{i1}^{(A)} + z_{i2}^{(A)}$, indicating the length of the questions and answers, are interpreted as how specifically the questions are asked and addressed in the consultation, $\frac{z_{i2}^{(Q)}}{z_{i1}^{(Q)}+z_{i2}^{(Q)}}$ can serve as a good approximation of the patient's health condition, and $\frac{z_{i2}^{(A)}}{z_{i1}^{(A)}+z_{i2}^{(A)}}$, the proportion of the physician's answers/advice on the urgent inquiries can represent the patient's health condition that is evaluated by the physician.

Estimation Results of the Choice Model for Physician Class

We report the estimation of the choice model of the physician class in Table 4. The significantly negative γ_1 implies that patients are more likely to consult a senior physician if their condition is less urgent. A number of factors may help explain this finding. As documented by Bavafa et al. (2018), the online channel can be a gateway for patients to seek further offline medical care. Moreover, price-sensitive users on the platform are more likely to consult a junior physician since they do not expect a significant difference in the responses

[1]Urgent care is referred to as the medical care provided for illnesses or injuries which require prompt attention, but are not life-threatening. It should not be confused with emergent care, as urgent care are typically not of such seriousness as to require the services of an emergency room.

Random intercept	mean (μ_{γ_0})	variance ($\sigma^2_{\gamma_0}$)	Slope	Estimate	Standard error
	2.069	3.561	Urgency (γ_1)	-0.178*	0.086

Table 4. Model Estimation for the Choice of Physician Class in Consultation[2]

Parameter	Estimate	Standard error	Parameter	Estimate	Standard error
Intercept (β_{110})	0.489*	0.093	Answer length (β_{114})	-0.003	0.004
Senior (β_{111})	-0.147*	0.054	Answer length $\times$ Senior (β_{115})	0.003	0.004
Question length (β_{112})	0.007*	0.002	Answer urgency (β_{116})	-0.089*	0.042
Question urgency (β_{113})	-0.116*	0.042	Answer urgency $\times$ Senior (β_{117})	-0.199*	0.079

Table 5. Effects of a Consultation on a Future Consultation

Parameter	Estimate	Standard error	Parameter	Estimate	Standard error
Intercept (β_{210})	-0.106*	0.040	Answer length (β_{214})	-0.002	0.002
Senior (β_{211})	0.120*	0.039	Answer length $\times$ Senior (β_{215})	0.000	0.003
Question length (β_{212})	0.005*	0.002	Answer urgency (β_{216})	0.060*	0.027
Question urgency (β_{213})	0.084*	0.040	Answer urgency $\times$ Senior (β_{217})	0.004	0.048

Table 6. Effects of a Consultation on a Future Appointment

from two classes of physicians. Meanwhile, patients who expect to resolve their minor and routine medical concerns online are more quality-sensitive; they may be willing to pay a higher price for consulting a senior physician online. Since fully explaining this phenomenon requires more granular data and it is not the main focus of the current study, we leave it for future empirical research. We also estimate $\gamma_0^{(n)}$ for each patient n use these values to predict their probabilities of choosing a junior or a senior physician.

Estimation Results of the Marked Hawkes Process

We report in Tables 5 to 7 the coefficients β's in the intensity functions of the marked Hawkes process. The self-modulating coefficient, β_{110}, describes the instantaneous effect of a consultation with a junior physician on the intensity of a future consultation, which is modulated by the patient health condition and the information conveyed during the conversation. The incremental effect of consulting a senior physician is captured by the parameter β_{111}. The significantly positive estimate of β_{110} indicates that the probability of a patient choosing the consultation service will be increased due to his/her past experience with this online service. However, this effect is less prominent if a previous consultation was with a senior physician in the presence of the negative β_{111}. In contrast, Table 6 demonstrates an opposite effect on future appointments. The mutual-modulating coefficient, β_{210}, describes the instantaneous effect of a consultation with a junior physician on the intensity of a future appointment. The negative β_{210} implies that consulting a junior physician inhibits future appointment. But, β_{211} (and $\beta_{210} + \beta_{211}$) is estimated to be positive, showing that a consultation with a senior physician is more likely to induce future appointments. These results suggest that consulting a junior physician leads to a higher probability of a follow-up consultation instead of an appointment, while consulting a senior physician reverse the effect. This finding also implies the potential of routing patients to different ranks of physicians to regulate the patient traffic. We further explore this aspect through a series of simulation studies, shortly.

We further discuss the impact of the content of consulting conversations on future behaviors. Based on the estimates in Table 5 and 6, longer and more specific questions lead to a higher probability of seeking both online and offline medical cares ($\beta_{112} > 0$ and $\beta_{212} > 0$). However, the urgency of a patient's condition (inferred from the questions) has differing impact on which channel to seek further cares. After consulting on an urgent condition, a patient prefers to see a doctor offline ($\beta_{213} > 0$), instead of online ($\beta_{113} < 0$). The effects of physicians' answers are similar: if physicians indicate the urgency of conditions, a patient tend to see a doctor offline ($\beta_{216} > 0$), instead of online ($\beta_{116} < 0$). In addition, the effect on the online channel will be more pronounced if the response is from a senior physician ($\beta_{117} < 0$).

[2]The star symbol * indicates that the 95% confidence interval excludes 0. Standard errors are reported in parentheses. The same notation is also used in Tables 4 to 7.

Parameter	Estimate	Standard error	Parameter	Estimate	Standard error
Effect on consultation (β_{120})	0.0685*	0.002	Effect on appointment (β_{220})	0.1318*	0.002

Table 7. Effects of an Appointment on a Future Consultation or Appointment

k	$\eta_{kk'}$		m_k	σ_k^2
	$k'=1$	$k'=2$		
1	3.484	3.088	-1.306	4.948
2	4.164	2.301	-0.033	1.529

Table 8. Decay Rate $\eta_{kk'}$ and Log-normal Mean m_k and Variance σ_k^2 of Baseline Intensities

We show in Table 7 the estimated effects of an appointment on future consultations and appointments. The impacts on both channels are significantly positive ($\beta_{120} > 0$ and $\beta_{220} > 0$), suggesting a user's continuous needs for medical service. This finding is consistent with our empirical context. Since we focus on gynecology and obstetrics related health services, there is a significant portion of pregnant women and women who are in perinatal period among users. These users indeed require frequent medical visits. Additionally, Table 8 reports the decay rates $\eta_{kk'}$. The effect of a past consultation on future consultations exponentially decays at rate η_{11}, which is smaller than η_{21}, the decay rate of the effect of a past consultation on appointments. This finding demonstrates that the self-modulating effect of consultation is more lasting. We can reach a similar conclusion for the effect of an appointment by comparing the decay rates η_{k2}'s, where $\eta_{12} > \eta_{22}$. It is also reported in Table 8 the log-normal mean and variance for the random baseline intensities among all patients. With $m_1^2 < m_2^2$ and $\sigma_1^2 > \sigma_2^2$, the average baseline intensities of consulting online is smaller than that of in-person visit, but is of higher heterogeneity among patients.

Prediction and Model Comparison

We showcase the prediction power of the proposed Poisson-factor-marked Hawkes process compared to some benchmark counting processes and highlight the importance of the marks (content of consultation and choices of the physician class) and mutual modulation. To measure the performance of a model for multiple types of events in temporal order, a common way is to use Brier score (BS) that quantifies the accuracy of cumulative incidence function (CIF) estimations (Zhang and Zhou 2018). Specifically,

$$\text{CIF}_k(i,t) = P(t_i \le t, \kappa_i = k)$$

is the probability that event i is of type k and happens by time t (Crowder 2001; Fine and Gray 1999; Kalbfleisch and Prentice 2011). Brier score (Gerds et al. 2008; Steyerberg et al. 2010) represents the average squared errors between the observed event status and its estimated CIF and is defined as

$$\text{BS}_k(t) = \frac{1}{I}\sum\nolimits_{i=1}^{I}\left[\mathbf{1}(t_i \le t, \kappa_i = k) - \text{CIF}_k(i,t)\right]^2,$$

where the indicator function $\mathbf{1}(.)$ is equal to 1 if the condition $(t_i \le t, \kappa_i = k)$ holds and equal to 0 otherwise. Smaller Brier scores indicate better model fit or prediction.

To illustrate the prediction accuracy and the importance of each component of our model, we compare with four benchmark models: 1) a Poisson process, 2) an unmarked Hawkes process (HP), 3) a marked Hawkes process (MHP) without using content of consultations (the factor scores of the questions and anwers), and 4) a marked Hawkes process that does include the phisician class and its interaction with the factor scores. To be specific, the Poisson process 1) assumes a constant intensity of making a consultation and an appointment so that a patient's future activities are independent of her history. The unmarked Hawkes process 2) includes fixed mutually modulating effects between a consultation and making an appointment, but does not utilize the information of the consultation. Furthermore, the marked Hawkes processes 3) and 4) intentionally overlook some of the information of consultations, respectively.

We use the events of all patients in their first 150 days to train the model and those afterwards for testing.

Model $\Delta t =$	Consultation ($k = 1$) 3 days	6 days	9 days	Appointment ($k = 2$) 3 days	6 days	9 days
1) Poisson process	0.071	0.100	0.118	0.224	0.281	0.317
2) Unmarked HP	0.062	0.085	0.101	0.148	0.232	0.299
3) MHP: no Q&A	0.060	0.083	0.099	0.145	0.230	0.300
4) MHP: no physician class	0.058	0.078	0.092	0.141	0.221	0.279
5) Proposed model	**0.056**	**0.077**	**0.091**	**0.137**	**0.211**	**0.268**

Table 9. Model Comparison in Brier Score

The training data contain 57.8% of all the events. For each event i in the testing data, we find the Brier score

$$[\mathbf{1}(t_i \le t_{i-1} + \Delta t, \kappa_i = k) - \mathrm{CIF}_k(i, t_{i-1} + \Delta t \,|\, \mathcal{H}(t_{i-1}))]^2, \tag{6}$$

and report in Table 9 the scores averaging over all the testing events for $k = 1, 2$ and Δt equal to 3 days, 6 days, and 9 days, respectively. We find that the proposed model outperforms all the benchmarks in predicting probabilities of upcoming events in different time spans. The superior prediction accuracy demonstrates the value of accounting for the exciting or inhibiting interactions among events and the information of consultation. Note that Brier scores of the marked Hawkes process measures not only the accuracy of event prediction, but also accounting for how far we look forward into the future, namely Δt in (6) which is arbitrary in the testing stage. In comparison, though classification models, like logistic and probit regressions, are also able to predict future event probabilities, without modeling event times as dependent variables they have to aggregate information in a prespecified time span in which training and prediction are restricted.

Model Application

We have shown the effects of past consultation dialogues and choices of the physician class on the intensities (instantaneous probabilities) of a patient's future behaviors. However, the marginal effects in the longer term cannot be fully quantified by the coefficients β's and η's. In this section, we utilize the self-generative property of the proposed model to explain how patients' prior activities change their future behaviors.

We have estimated β's that represent the instantaneous effects of consulting physicians of different classes or visiting a physician offline. To better quantify such effects in the longer term and route patients accordingly, we impose an initial event, namely event 0, and simulate patient activity streams afterwards. We aim to investigate how the physician class and the patient's health condition in the consultation of event 0 affect the distribution of future events in terms of the next-fifteen-day event probabilities and event counts. Specifically, for each patient, we impose an event 0 of type $\kappa_0 = 1$ (consultation) at $t_0 = 0$ and set w_0 equal to 0 and 1, respectively, representing consulting a junior and senior physician. We fix the question factor scores $z_{01}^{(Q)}$, $z_{02}^{(Q)}$ of event 0 such that $z_{01}^{(Q)} + z_{02}^{(Q)} = 3.93$, which is the median value of $z_{01}^{(Q)} + z_{02}^{(Q)}$ among all consultations estimated by the correlation Poisson factor model. Furthermore, we set $z_{02}^{(Q)}/(z_{01}^{(Q)} + z_{02}^{(Q)})$ equal to 0.1 and 0.9, respectively, to represent a routine and emergent inquiry in the initial consultation.

Leveraging the correlated Poisson factor analysis, we simulate $\boldsymbol{z}_0^{(A)} = (z_{01}^{(A)}, z_{02}^{(A)})$, the factor scores of the physician's answer given on $\boldsymbol{z}_0^{(Q)} = (z_{01}^{(Q)}, z_{02}^{(Q)})$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ by

$$\boldsymbol{z}_i^{(A)} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{z}_i^{(Q)} \sim \text{log-Normal}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\log \boldsymbol{z}_i^{(Q)} - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}), \tag{7}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of dimension S and $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{22}$, and $\boldsymbol{\Sigma}_{21}$ of dimension $S \times S$ are block vectors or matrices of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ such that

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

By now, we have characterized event 0, including the time, type, and mark values. The following events $i = 1, 2, \ldots$ are simulated by the marked Hawkes process characterized by (3) where marks are generated by (1) and (2).

We simulate such event streams with an arbitrary event 0 for 10,000 patients and report in Table 10 the empirical probability that at least one consultation or appointment occurs in the first fifteen days, denoted

Event 0	$\frac{z_{02}^{(Q)}}{z_{01}^{(Q)}+z_{02}^{(Q)}}$	Consultation		Appointment		Appointment rate
		$p(N_1(15) \geq 1)$	$\mathbb{E}N_1(15)$	$p(N_2(15) \geq 1)$	$\mathbb{E}N_2(15)$	
None	NA	0.265	0.977	0.426	1.275	0.566
Consultation (junior)	0.1	0.353	1.159 (+)	0.395	1.152 (−)	0.499 (−)
Consultation (junior)	0.9	0.315	1.039 (+)	0.429	1.270 (−)	0.550 (−)
Consultation (senior)	0.1	0.302	0.983 (+)	0.435	1.254 (−)	0.561 (−)
Consultation (senior)	0.9	0.255	0.857 (−)	0.453	1.359 (+)	0.613 (+)
Real data	NA	NA	1.045	NA	1.342	0.562

Table 10. Probabilities and Average Numbers of Consultations and Appointments in 15 Days

by $p(N_k(15) \geq 1)$, $k = 1, 2$, as well as the expected number of the two types of events, denoted by $\mathbb{E}N_k(15)$. Also reported is the appointment rate that is equal to $\mathbb{E}N_2(15)/(\mathbb{E}N_1(15) + \mathbb{E}N_2(15))$. We do not count the arbitrary initial event when calculating such measures. In addition, a cold-starting stream with no event at time 0 is simulated as a benchmark, and the corresponding statistics are given in row 1 of the table. For simulated event streams reported in row 2 to 5, we indicate in parenthesis the change in direction compared to the cold start. We also calculate the fifteen-day average number of consultations and appointments per patient in the original data and report them in the last row of the table.

Compared to a cold start, if consulting a junior physician during the initial encounter, the direction of its effect does not vary across different patient condition severities. Specifically, consulting a junior physician raises the number and probability of an upcoming consultation, but drops those of an upcoming appointment in the next fifteen days, regardless of the severity of the conditions. However, consulting a senior physician demonstrates a separating effect. That is, senior physicians are more likely to direct patients with severe conditions to offline appointments instead of online services, while doing the opposite to patients with minor conditions. Taken all together, compared to a cold start, the proportion of appointments in patients' event streams decreases by 0.6-6.7% when the initial encounter happens online with a junior physician. In contrast, the appointment rate increases by 4.7% if initially consulting with a senior physician for a severe health issue, while it decreases by 0.5% if it is for a minor issue. Additionally, in the real data, the average number of consultations and appointments in fifteen days is 1.045 and 1.342, respectively, and the appointment proportion is 0.562. These recovered data statistics by our simulation demonstrate that our Poisson-factor-marked Hawkes process has well approximated the ground true generating process of the patients' activities.

Since we have controlled the severity of the initial health encounter in the simulation, the difference across the two types of physicians stated above can be mainly attributed to behavioral factors. The heterogeneity can possibly be explained by the disparate incentives that two types of physicians have – junior physicians have a stronger incentive to induce patients to use the online telehealth platform. The offline healthcare market in China has been experiencing the long-standing issue of "the inverted-pyramid of demand" where the majority of patients who experience minor health issues flood into top-level hospitals to see experienced and renowned physicians, creating a significant congestion and a severe resource misallocation (Deng et al. 2021). Thus, junior physicians with fewer demands offline are more willing to be active on the telehealth platform to make extra income. However, senior physicians do not have such an incentive, and thus, are more likely to provide medical advice based on the severity of conditions.

Telemedicine platforms have been introduced as a substitute channel for unnecessary offline visits. With the emerging "Digital-First" health care approach, National Health Service (NHS) England harnesses it to cut health inequalities (NHS England 2020), which has been realized by telehealth platforms like Babylon (UK). Our simulation studies have pointed out the importance of guidelines and the corresponding training for physicians who practice online. Although there may be certain differences in diagnosis between two classes of physicians, such a systematic disparity should be less likely to happen if clear telehealth practice guides are provided and the incentive among the platform, different types of physicians and other parties is properly aligned.

Conclusion

Patients often resolve health-related concerns through a journey of multiple contacts with healthcare providers. With telehealth technologies' wide-spreading, patients are offered an additional online channel to interact with physicians. Quantifying such journeys is a challenging but critical question to understand patient behavior better and improve the design of patient routing strategies with telehealth platforms. We address this challenge by proposing a novel model of Poisson-factor-marked Hawkes process, which captures the modulating effects across services of both channels and the impacts of physician characteristics and the content of patient-physician online conversations. We demonstrate several interesting findings. Physician seniority and patient urgency (inferred from the consultation content) have differing effects on future encounters. Patients are more likely to follow up through the telehealth platform when interacting online previously with a junior physician. On the contrary, senior physicians more often direct patients to online channels for future care. On top of that, to be expected, patients whose conversations cover more urgent concerns prefer office visits.

We highlight the prediction power of the proposed model and the contributions to the Bayesian probabilistic topic models and the applications of Hawkes process. Our model introduces topic correlations between patients' questions and the corresponding responses from physicians. Based on this correlated Poisson factor analysis, we identify two factors, representing different types of demands for the online health service. We further incorporate the weights on these factors in the Hawkes process, which allows for both exciting and inhibiting effects across online and offline medical encounters. These components jointly capture the data generation process of patients' activity streams, including the activity times and types, and dialogues during consultations.

Our model also has important applications and offers several practical implications. Employing the self-generating property of the proposed model, we simulate the care journeys by imposing different initial encounters. Further underlining the estimation results, simulation studies show a 0.6-6.7% drop in the portion of office appointments in the next-fifteen-day window after patient consulting a junior physician, compared to a cold start; however, the portion will be increased mildly if this initial online contact is with a senior doctor for a severe health issue while be decreased if it is for a minor issue. Contributing to the current discussions on the "Digital-First" health care approach, we showcase the importance of routing patients to the correct type of providers and guiding physicians with precise standards. We also propose various other applications and simulation scenarios. For instance, we illustrate an application scenario for AI-assisted telehealth decision-making, which can help a platform prioritize demands based on the predicted probability of patients' activity in the near future and better allocation health resources.

The proposed model can be extended and augmented in different ways. First, the focal study focuses on online interactions from the telehealth platform's perspective. Future research could apply the same framework and incorporate various factors related to offline health encounters. Second, instead of focusing on one medical specialty for the outpatient setting, research can also be done by modeling the activity streams of patients across different medical departments for inpatient and outpatient care. Third, expanding the options of two classes of doctors, a potential avenue for future research is to model the choice of each individual physician. This model can be applied to recommendation systems to optimize the patient experience.

Acknowledgements

Quan Zhang is supported in part by the Broad College Summer Research Grant, 2021.

References

Aggarwal, V., Hwang, E. H., and Tan, Y. (2021). "Learning to be creative: A mutually exciting spatiotemporal point process model for idea generation in open innovation," *Information Systems Research* (32:4), pp. 1214–1235.

AMA (2022). *2021 Telehealth Survey Report*. www.ama-assn.org.

Ayabakan, S., Bardhan, I., and Zheng, E. (2020). "Impact of telehealth use on healthcare utilization: A quasi-experimental study of maryland patients," Available at SSRN 3707829.

Bao, Y. and Datta, A. (2014). "Simultaneously discovering and quantifying risk types from textual risk disclosures," *Management Science* (60:6), pp. 1371–1391.
Bavafa, H., Hitt, L. M., and Terwiesch, C. (2018). "The impact of e-visits on visit frequencies and patient health: Evidence from primary care," *Management Science* (64:12), pp. 5461–5480.
Bellstam, G., Bhagat, S., and Cookson, J. A. (2021). "A text-based analysis of corporate innovation," *Management Science* (67:7), pp. 4004–4031.
Bestsennyy, O., Gilbert, G., Harris, A., and Rost, J. (2021). *Telehealh: A quarter-trillion-dollar post COVID-19 reality?* https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality.
Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational inference: A review for statisticians," *Journal of the American statistical Association* (112:518), pp. 859–877.
Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent dirichlet allocation," *Journal of machine Learning research* (3:Jan), pp. 993–1022.
Brandt, T., Dlugosch, O., Abdelwahed, A., van den Berg, P. L., and Neumann, D. (2021). "Prescriptive analytics in urban policing operations," Manufacturing & Service Operations Management.
Canny, J. (2004). "GaP: A factor model for discrete data," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 122–129.
CMS (2020). *Medicare telemedicine health care provider fact sheet*. CMS.gov. 2020.
Crowder, M. J. (2001). *Classical competing risks,* CRC Press.
Daw, A., Castellanos, A., Yom-Tov, G., Pender, J., and Gruendlinger, L. (2021). "The co-production of service: Modeling service times in contact centers using hawkes processes," Available at SSRN 3817130.
Deng, Z., Jiang, N., Song, S., and Pang, R. (2021). "Misallocation and price distortions: A revenue decomposition of medical service providers in China," *China Economic Review* (65), p. 101574.
Dulleck, U. and Kerschbamer, R. (2006). "On doctors, mechanics, and computer specialists: The economics of credence goods," *Journal of Economic Literature* (44:1), pp. 5–42.
Fine, J. P. and Gray, R. J. (1999). "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American Statistical Association* (94:446), pp. 496–509.
Fortune Business Insights (2021). *Telehealth Market Size*. https://www.fortunebusinessinsights.com/industry-reports/telehealth-market-101065.
Gartner (2020). *The digital first engagement framework for healthcare delivery organizations*. https://www.gartner.com/en/doc/732384-the-digital-first-engagement-framework-for-healthcare-delivery-organizations. 2020.
Gerds, T. A., Cai, T., and Schumacher, M. (2008). "The performance of risk prediction models," *Biometrical Journal: Journal of Mathematical Methods in Biosciences* (50:4), pp. 457–479.
Green, L. V., Savin, S., and Lu, Y. (2013). "Primary care physician shortages could be eliminated through use of teams, nonphysicians, and electronic communication," *Health Affairs* (32:1), pp. 11–19.
Hawkes, A. G. (1971). "Spectra of some self-exciting and mutually exciting point processes," *Biometrika* (58:1), pp. 83–90.
HRSA (2020). *Telehealth Programs*. https://www.hrsa.gov/rural-health/telehealth.
Jankowiak, M. and Gomez-Rodriguez, M. (2017). "Uncovering the spatiotemporal patterns of collective social activity," in *Proceedings of the 2017 SIAM International Conference on Data Mining,* SIAM, pp. 822–830.
Ju, C., Zhang, S., et al. (2020). "Influencing factors of continuous use of web-based diagnosis and treatment by patients with diabetes: model development and data analysis," *Journal of Medical Internet Research* (22:9), e18737.
Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data,* vol. 360. John Wiley & Sons.
Kingma, D. P. and Welling, M. (2014). "Stochastic gradient VB and the variational auto-encoder," in *Second International Conference on Learning Representations, ICLR,* vol. 19, p. 121.
Lee, G. M., He, S., Lee, J., and Whinston, A. B. (2020). "Matching mobile applications for cross-promotion," *Information Systems Research* (31:3), pp. 865–891.
Mei, H. and Eisner, J. M. (2017). "The neural hawkes process: A neurally self-modulating multivariate point process," *Advances in Neural Information Processing Systems* (30).

Mukherjee, U. K., Ball, G. P., Wowak, K. D., Natarajan, K. V., and Miller, J. W. (2022). "Hiding in the herd: The product recall clustering phenomenon," *Manufacturing & Service Operations Management* (24:1), pp. 392–410.

NHS England (2020). *Digital first primary care*. https://www.england.nhs.uk/gp/digital-first-primary-care/.

North, F., Crane, S. J., Chaudhry, R., Ebbert, J. O., Ytterberg, K., Tulledge-Scheitel, S. M., and Stroebel, R. J. (2014). "Impact of patient portal secure messages and electronic visits on adult primary care office visits," *Telemedicine and e-Health* (20:3), pp. 192–198.

Okawa, M., Iwata, T., Tanaka, Y., Toda, H., Kurashima, T., and Kashima, H. (2021). "Dynamic Hawkes processes for discovering time-evolving communities' states behind diffusion processes," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining,* pp. 1276–1286.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). "Automatic differentiation in PyTorch," in *Neural Information Processing Systems 2017 Workshop on Autodiff,*

Pu, J., Chen, Y., Qiu, L., and Cheng, H. K. (2020). "Does identity disclosure help or hurt user content generation? Social presence, inhibition, and displacement effects," *Information Systems Research* (31:2), pp. 297–322.

Reynaud-Bouret, P. and Schbath, S. (2010). "Adaptive estimation for Hawkes processes; application to genome analysis," *The Annals of Statistics* (38:5), pp. 2781–2822.

Shah, S. J., Schwamm, L. H., Cohen, A. B., Simoni, M. R., Estrada, J., Matiello, M., Venkataramani, A., and Rao, S. K. (2018). "Virtual visits partially replaced in-person visits in an ACO-based medical specialty practice," *Health Affairs* (37:12), pp. 2045–2051.

Spence, M. (1978). "Job market signaling," in *Uncertainty in economics,* Elsevier, pp. 281–306.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). "Assessing the performance of prediction models: A framework for some traditional and novel measures," *Epidemiology (Cambridge, Mass.)* (21:1), p. 128.

Tibshirani, R., Walther, G., and Hastie, T. (2001). "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (63:2), pp. 411–423.

Xu, L., Duan, J. A., and Whinston, A. (2014). "Path to purchase: A mutually exciting point process model for online advertising and conversion," *Management Science* (60:6), pp. 1392–1412.

Yang, S.-H. and Zha, H. (2013). "Mixture of mutually exciting processes for viral diffusion," in *International Conference on Machine Learning,* PMLR, pp. 1–9.

Yoo, E., Gu, B., and Rabinovich, E. (2019). "Diffusion on social media platforms: A point process model for interaction among similar content," *Journal of Management Information Systems* (36:4), pp. 1105–1141.

Young, G. J., Flaherty, S., Zepeda, E. D., Mortele, K. J., and Griffith, J. L. (2020). "Effects of physician experience, specialty training, and self-referral on inappropriate diagnostic imaging," *Journal of General Internal Medicine* (35:6), pp. 1661–1667.

Zeltzer, D., Einav, L., Rashba, J., and Balicer, R. D. (2021). *The impact of increased access to telemedicine*. Tech. rep. National Bureau of Economic Research.

Zhang, Q. and Zhou, M. (2018). "Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks," in *Advances in Neural Information Processing Systems,* pp. 5002–5013.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). "Beta-negative binomial process and Poisson factor analysis," in *Artificial Intelligence and Statistics,* PMLR, pp. 1462–1471.

Zhou, Y. Y., Garrido, T., Chin, H. L., Wiesenthal, A. M., and Liang, L. L. (2007). "Patient access to an electronic health record with secure messaging: impact on primary care utilization," *Am J Manag Care* (13:7), pp. 418–424.