A Prescription Trend Analysis using Medical Insurance Claim Big Data

Kazutoshi Umemoto † Kazuo Goda † Naohiro Mitsutake ‡ Masaru Kitsuregawa †
† The University of Tokyo ‡ Institute for Health Economics and Policy
† {umemoto,kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp ‡mitsutake@ihep.jp

***Abstract*—Understanding the spread of diseases and the use of medicines is of practical importance for various organizations, such as medical providers, medical payers, and national governments. This study aims to detect the change in the prescription trends and to identify its cause through an analysis of *Medical Insurance Claims* (MICs), which comprise the specifications of medical fees charged to health insurers. Our approach is two-fold. (1) We propose a latent variable model that simulates the medication behavior of physicians to accurately reproduce monthly prescription time series from the MIC data, where prescription links between the diseases and medicines are missing. (2) We apply a state space model with intervention variables to decompose the monthly prescription time series into different components including seasonality and structural changes. Using a large dataset consisting of 3.5-year MIC records, we conduct experiments to evaluate our approach in terms of accuracy, usefulness, and efficiency. We also demonstrate three applications for our medical analysis.**

I. Introduction

The extent of the spread of diseases and the frequency of the use of medicines vary over time. Diseases have time-varying factors such as seasonality and epidemics [1]. A number of new medicines have been developed by pharmaceutical companies. In Japan, for example, more than 100 applications of new medicines are approved by the Minister of Health, Labour and Welfare every year.

Understanding the trends in medicine prescriptions for diseases is a key issue, and, in particular, detecting the change in prescription trends is of practical importance for various organizations. For pharmaceutical companies, knowing the trend of new medicines plays an important role in planning an appropriate marketing strategy to provide their products and in spreading cutting-edge prescription to all medical institutions. The accurate tracking of trending prescriptions also enables national governments to confirm the proper use of medicines and to make medical charges more reasonable. Furthermore, if new indications can be detected early from the actual use of medicines in clinical practice, the feasibility of clinically-based drug repositioning[1] will be worth exploring as an alternative to conventional bioinformatics approaches [3], [4].

Traditionally, electronic health records (EHRs) [5] and X-ray images [6]–[8] have been frequently used for the resources of data mining in the medical domain. However, because EHRs contain highly confidential data, it is not easy to obtain data from many medical institutions. In addition, when using X-ray images, one cannot analyze the diseases that do not require an X-ray inspection. These make it difficult to conduct *universal analysis for population-scale healthcare* using EHR data.

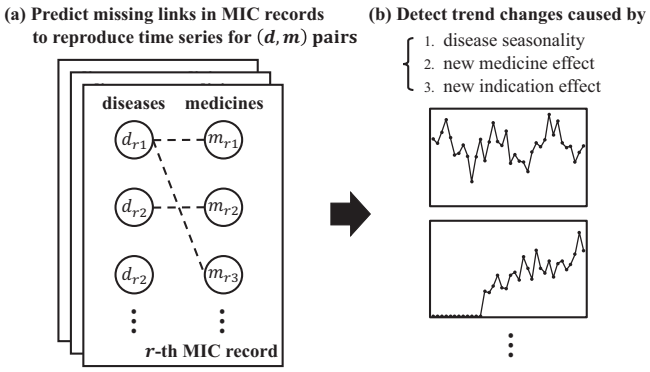


Fig. 1: Overview of our two-fold approach.

As another resource, the present study focuses on *Medical Insurance Claims* (MICs), which comprise the specifications of medical fees that medical institutions charged to health insurers (see Section III-A for details). An advantage of using MIC data is the *full coverage* of patients: as the Japanese government has achieved a universal healthcare service system since 1961, every citizen is obligated, by the law, to take out any health insurance, where most MIC records are computerized.[2] While we use Japanese MIC data in this work, similar systems have been adopted in other nations including Korea and Taiwan.[3]

Given the MIC big data, this paper addresses the problem of detecting the change in prescription trends and identifying its cause. Our approach is two-fold as shown in Figure 1. The first step is to accurately reproduce monthly prescription time series from the MIC data, where prescription links between diseases and medicines are missing. To this end, we propose a latent variable model that simulates the medication behavior of physicians (Figure 1a). The second step is to detect the trend change from the reproduced prescription time series. We achieve this by applying a state space model with intervention variables that can decompose the prescription time series into different components including seasonality and structural

[1] Drug repositioning is the application of known drugs to new indications. This approach has received great attention owing to its advantages over traditional drug development in terms of drug safety and development cost [2].

[2] As of September, 2017, the penetration rate of electronic MIC records was 93.2% on a medical institution basis and 98.2% on a record basis.

[3] http://www.esri.go.jp/jp/prj/int_prj/prj-2004_2005/macro/macro16/09-1-R.pdf

changes. We then categorize the trend change into disease-, medicine-, and prescription-caused changes by assessing the structural change component (Figure 1b). We evaluate the effectiveness of our approach with real big data consisting of 3.5-year MIC records.

We make the following contributions with this work:

- We introduce the change detection of prescription trends as a promising application for mining electronic MIC big data, which has a large impact on not only academic but also practical domains such as medical, administrative, and economic fields. To the best of our knowledge, this is the first attempt to use MIC data for this purpose. We also present two more applications for universal medical analysis.
- We develop a probabilistic medication model that uses latent variables to predict prescription links missing from MIC records. Our experiments show that the model performs significantly better than a cooccurrence-based approach in terms of both predictive capability of unseen medicines and prescription relevance assessed by a medical professional.
- We empirically show that our model based on a state space model with intervention variables can find from hundreds of thousands of disease-medicine pairs the ones with a change in prescription trends due to, for example, new medicine and new indication effects. We also propose a method for efficiently finding approximate change points and compare its performance with a method finding the exact solution in terms of both computational cost and approximation accuracy.

II. Related Work

Medical Data Mining. Much effort has been devoted to medical data mining, aiming at improving the quality of medical services either directly or indirectly [5]–[8]. In addition to sensitive medical data such as EHRs [5] and X-ray images [6]–[8], other data resources are beginning to be analyzed from the different perspective of user understanding and knowledge discovery on the medical domain. Paparrizos *et al*. [9] used search logs to predict those who will issue first-person diagnostic queries about devastating diseases. Mishra *et al*. [10] also leveraged search logs to find searchers with time-critical health information needs (*e.g*., seeking an urgent care facility). Aramaki *et al*. [11] proposed using Twitter as a social sensor to detect influenza epidemics. As alternative data, we use MIC records in this work to detect the change in prescription trends. Matsubara *et al*. [1] developed FUNNEL, an analytical model for long-term epidemiological data across a wide area. FUNNEL is so flexible as to generalize existing epidemiological models like SIRS [12] and can find important patterns of epidemiological time series. We borrow from Matsubara *et al*.'s work [1] several factors (*e.g*., seasonality) affecting prescription trends while also considering the own ones unique to our problem setting (Section III).

Link Prediction. The problem of predicting links between nodes has been studied extensively as it has a broad range of applications such as product purchase prediction and human relationship understanding [13]–[15]. To understand the association between words and tags in documents, Blei *et al*. [16] proposed Correspondence Topic Model, which selects a topic for each tag from the ones assigned to words in the same document. Iwata *et al*. [17] extended that model to deal with the situation in which documents contain noisy tags unrelated to the content. The MIC data we use in this work also lacks links indicating prescriptions between diseases and medicines. The absence of these links causes adverse effects on the accurate reproduction of prescription time series, from which we detect trend changes. Thus, we propose a similar model with latent variables that simulates physicians' medication behavior.

Time Series Analysis. The autoregressive (AR), autoregressive integrated moving average (ARIMA), and state space models have been well-known, representative methods throughout the history of time series analysis. Many researchers have addressed the problem of time series analysis by building on these models [18]–[20]. The state space model, in particular, is a technique encompassing the AR and ARIMA models, and the parameter estimation and forecasting can be efficiently achieved thanks to the Kalman filter [21]. It allows one to incorporate domain knowledge and/or his/her assumption to the model and to interpret the change in time series through the component decomposition [22]. Taking these advantages into account, we apply a state space model with intervention variables for our purpose. The change and burst detection has also been studied by the data mining community for years [23]–[26]. In this work, we detect the change in prescription trends in a fully automatic manner by considering the fitting quality of our model.

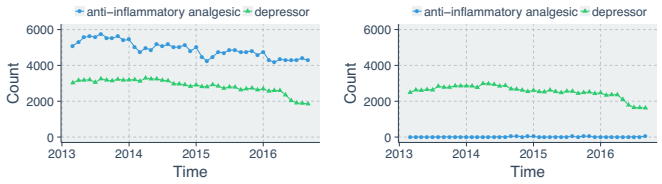
III. Challenges

This section describes two main challenges to be addressed in this work. One is attributed to the structure of MIC data, while the other is related to factors affecting the change in prescription trends.

A. Medical Insurance Claims (MICs)

MIC records that we use in this work consist of (1) medical institutions, (2) patients covered by health insurances, (3) the patients' diseases diagnosed by physicians, (4) medical services (*e.g*., actions taken, medicines prescribed, and devices used) offered for the disease treatments, and (5) medical fees for the services. Each medical institution creates a single MIC record for every patient on a monthly basis to recover the medical fees incurred from the health insurers. Receiving the MIC records, the health insurers assess the validity of the medical treatments and, if the assessment is favorable, pay the medical institutions the medical fees (except for out-of-pocket expenses borne by the patients).

As the purpose of generating MIC records is to charge the incurred medication fees to health insurers, detailed clinical information (*e.g*., the results of medical tests and X-ray images) is omitted from these records, unlike in EHRs. This



(a) Cooccurrence-based approach (b) Probabilistic medication model

Fig. 2: Adverse effect of missing prescription links on prescription count prediction for hypertention.

gives rise to the following challenge when using MIC data for prescription trend analysis.

Missing prescription links. As both a list of diseases diagnosed and a list of medicines prescribed are included in MIC data, we can easily understand how often each disease is diagnosed and how often each medicine is prescribed per month. However, there is no direct method to count the prescriptions for each disease-medicine pair. This is because links indicating the prescription relationship between diseases and medicines are missing from the MIC data. As noted above, each MIC record contains monthly medical treatments offered to every patient. For instance, if a patient visits the same hospital several times a month to have treatments for different diseases, individual treatments are aggregated into the same MIC record. In fact, the average frequencies of diseases and medicines per MIC record are as large as 7.435 and 4.788, respectively, in our dataset (Section VIII). To detect the change in prescription trends in a reliable manner, it is necessary to reproduce the prescription time series for each disease-medicine pair as accurately as possible by predicting the missing prescription links.

One straightforward approach to the above issue is to assume the number of cooccurrences between each disease and medicine in MIC data as the prescription count for that disease-medicine pair. Figure 2a shows the prescription time series of two medicines (*i.e.*, a depressor and an anti-inflammatory analgesic) for hypertension, predicted by this approach. Note that only the former medicine has efficacy for hypertension. Nevertheless, the cooccurrence-based approach shows a higher prediction of the prescription count for the latter medicine than that for the former one. In this way, this approach has a mis-prediction problem especially for medicines appearing frequently in the MIC data.

B. Prescription Trend

The number of prescriptions of medicines for diseases varies over time due to various factors. We carried out a close observation of real MIC data to organize components that are required to model the dynamics of the prescription time series. In what follows, we show several examples of prescription time series that have been estimated by our model (Section IV) for addressing the aforementioned missing link problem in MIC data.

Seasonality. Some diseases cause epidemics in particular seasons [1] while others behave stably during any season. Figure 3a shows the prescription time series of medicines for hay fever, heatstroke, and influenza. We can observe from this figure that hay fever, heatstroke, and influenza are prevalent during spring, summer, and winter, respectively. In this way, the seasonality of diseases brings periodic change to the number of prescriptions. A model for analyzing prescription trends must be able to distinguish such periodic change from structural changes described below.

TABLE I: Notation ((t) is omitted when it is clear from the context).

Symbol	Description
$\mathcal{R}^{(t)}$	monthly MIC dataset at time t
$R^{(t)}$	number of MIC records in $\mathcal{R}^{(t)}$
$D^{(t)}$	number of unique diseases in $\mathcal{R}^{(t)}$
$M^{(t)}$	number of unique medicines in $\mathcal{R}^{(t)}$
$\boldsymbol{d}_r^{(t)}$	bag of diseases in MIC record r ($\boldsymbol{d}_r^{(t)} = \{d_{rn}^{(t)}\}_{n=1}^{N_r^{(t)}}$)
$\boldsymbol{m}_r^{(t)}$	bag of medicines in MIC record r ($\boldsymbol{m}_r^{(t)} = \{m_{rl}^{(t)}\}_{l=1}^{L_r^{(t)}}$)
$\boldsymbol{\eta}^{(t)}$	$D^{(t)}$-dimensional parameter of disease distribution at time t
$z_{rl}^{(t)}$	(latent) disease for which medicine $m_{rl}^{(t)}$ is prescribed
$\boldsymbol{\theta}_r^{(t)}$	$D^{(t)}$-dimensional parameter of latent disease distribution for MIC record r
$\boldsymbol{\phi}_d^{(t)}$	$M^{(t)}$-dimensional parameter of medicine distribution for disease d
$q_{rld}^{(t)}$	probability of selecting disease d for l-th medical treatment in MIC record r

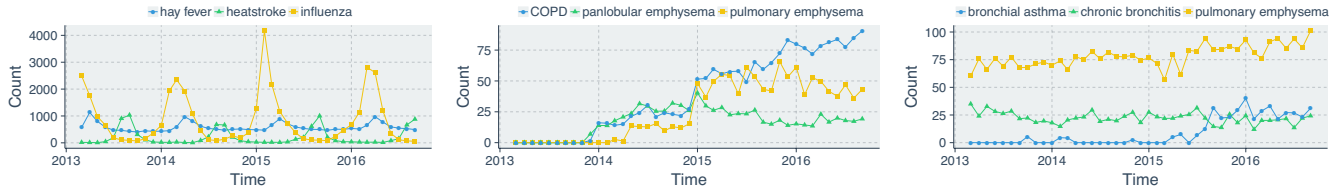
Medicine-derived structural changes. While seasonality is a disease-specific property affecting the number of prescriptions, some medicines also have their unique effects on the prescription trend changes. The release of new medicines is a prime example of such medicine-specific effects. Figure 3b shows the prescription time series of a bronchodilator for three of its target diseases. In the figure, the number of prescriptions for these diseases suddenly increased from zero around November in 2011, from when this medicine has been on sale. The revision of medicine price is another example. It is possible that a medicine whose price is discounted at some point in time will be prescribed more frequently from then on. These medicine-derived structural changes also need to be taken into consideration in the model.

Prescription-derived structural changes. Structural changes can also be caused by interaction effects between diseases and medicines. Examples of this effects include indication expansion (*i.e.*, adding new indications to known medicines). Figure 3c shows the prescription time series of another bronchodilator that is known to be efficacious for chronic obstructive pulmonary disease (COPD; the generic term for chronic bronchitis and pulmonary emphysema). Because bronchial asthma was announced as the new indication for this medicine around the end of 2014, the prescription for this indication has started to increase gradually since then.

Outliers. The prescription time series can have extreme spikes at some time points due to other external factors such as pandemics. In Figure 3a, for example, we can see that the prescription count for influenza during the winter season in 2014 is much larger than that in other years. Due to these outliers in real MIC data, the model needs to be robust against random fluctuations.

IV. Prescription Link Prediction

To address the missing link problem in the MIC data, we propose a latent variable model that simulates how physicians prescribe medicines for the diseases that they diagnose. Table I summarizes the notation used throughout the paper.



(a) Seasonality and outliers (b) Medicine-derived structural changes (c) Prescription-derived structural changes

Fig. 3: Factors affecting monthly prescription counts (a) for diseases related to climates and/or environments, (b) of newly released medicines, and (c) of existing medicines with new indications.

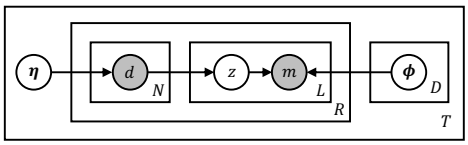


Fig. 4: Generative process of diseases and medicines in our model.

Suppose that we have T monthly MIC datasets, *i.e.*, $\mathcal{R}^{(t)}$ $(t \in \{1, \ldots, T\})$. Let $R^{(t)}$ be the total number of MIC records in $\mathcal{R}^{(t)}$, where $D^{(t)}$ and $M^{(t)}$ kinds of diseases and medicines appear, respectively. Each MIC record $r \in \{1, \ldots, R^{(t)}\}$ is represented by $(\boldsymbol{d}_r^{(t)}, \boldsymbol{m}_r^{(t)})$, where $\boldsymbol{d}_r^{(t)} = \{d_{rn}^{(t)}\}_{n=1}^{N_r^{(t)}}$ is a bag of diseases diagnosed in r, and $\boldsymbol{m}_r^{(t)} = \{m_{rl}^{(t)}\}_{l=1}^{L_r^{(t)}}$ is a bag of medicines prescribed in r. In what follows, we omit the superscript (t) when it is clear from the context.

A. Generative Process

Figure 4 illustrates how our model generates diseases and medicines in MIC data. As shown in the figure, diseases, realizations of latent variables, and medicines are generated in this order. In what follows, we explain each generation step in detail.

Disease Diagnosis. A physician makes diagnoses through patient examinations. Our model simulates this behavior by generating diseases $\boldsymbol{d}_r$ for each MIC record r. More specifically, $d_{rn} \in \boldsymbol{d}_r$, the n-th disease diagnosed in the MIC record r, is chosen from the multinomial distribution with the D-dimensional parameter $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_D)$, where $\eta_d \geq 0$ and $\sum_{d=1}^{D} \eta_d = 1$ (*i.e.*, $d_{rn} \sim \mathrm{Multinomial}(\boldsymbol{\eta})$). Note that diseases frequently diagnosed by physicians could vary over time due to, for example, seasonality. To take this temporal effect into account, we assume in our model that disease diagnoses (and medicine prescriptions described below) follow different distributions at different time points (on a monthly basis in our dataset).

Medication Target. After diagnosing diseases that a patient is affected, a physician judges which of them need medication. To simulate this behavior, our model iteratively selects a value of a latent variable z_{rl} from diseases $\boldsymbol{d}_r$ of the MIC record r generated in the previous step. Mathematically, this step is represented by $z_{rl} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_r)$, where $\boldsymbol{\theta}_r = (\theta_{r1}, \ldots, \theta_{rD})$ is a D-dimensional parameter ($\theta_{rd} \geq 0$ and $\sum_{d=1}^{D} \theta_{rd} = 1$) indicating the disease selection distribution in the MIC record r. Note that this parameter must satisfy the constraint $\theta_{rd} = 0$ for each disease d not appearing in r (*i.e.*, $d \notin \boldsymbol{d}_r$) because such disease cannot be the cause of medication in r.

Medicine Prescription. Once identifying diseases in need of medication, a physician prescribes appropriate medicines for each of them. To simulate this behavior, our model generates a medicine for each disease in $\{z_{rl}\}_{l=1}^{L_r}$. Given $z_{rl} = d$, the l-th medicine $m_{rl} \in \boldsymbol{m}_r$ prescribed in the MIC record r is chosen from the multinomial distribution with the M-dimensional parameter $\boldsymbol{\phi}_d = (\phi_{d1}, \ldots, \phi_{dM})$, where $\phi_{dm} \geq 0$ and $\sum_{m=1}^{M} \phi_{dm} = 1$ (*i.e.*, $m_{rl} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_d)$). In our model, a disease and a medicine are not independent with each other if the disease is regarded (by the latent variable) as the cause of prescribing the medicine. Making medicine distributions dependent on (latent) diseases allows us to represent the difference in medicine prescriptions among different diseases.

In clinical practice, some diseases require many medicines for treatment. An identical medicine is sometimes prescribed for different diseases. Occasionally, no medicine are prescribed for some diseases that are diagnosed in a medical examination. Our model is flexible enough to address all of these cases because it allows many-to-many relationships between diseases and medicines in each MIC record.

B. Formulation

We formulate the aforementioned generative process of diseases and medicines in MIC data. Given the parameters $\boldsymbol{\eta}$, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$, and $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_D)$, the occurrence probability $P(\mathcal{R} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ of the MIC data $\mathcal{R} = \{(\boldsymbol{d}_r, \boldsymbol{m}_r)\}_{r=1}^{R}$ is given by

$$
\begin{aligned}
P(\mathcal{R} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= \prod_{r=1}^{R} P(\boldsymbol{d}_r \mid \boldsymbol{\eta}) P(\boldsymbol{m}_r \mid \boldsymbol{d}_r, \boldsymbol{\theta}_r, \boldsymbol{\Phi}) \\
&= \prod_{r=1}^{R} \prod_{n=1}^{N_r} P(d_{rn} \mid \boldsymbol{\eta}) \prod_{l=1}^{L_r} \sum_{z_{rl} \in \boldsymbol{d}_r} P(z_{rl} \mid \boldsymbol{\theta}_r) P(m_{rl} \mid \boldsymbol{\phi}_{z_{rl}}) \\
&= \prod_{r=1}^{R} \prod_{n=1}^{N_r} \eta_{d_{rn}} \prod_{l=1}^{L_r} \sum_{d=1}^{D} \theta_{rd} \phi_{dm_{rl}} \,. \qquad (1)
\end{aligned}
$$

In the rest of this section, we describe how to estimate these parameters when fitting our model to the MIC data and how to reproduce the prescription time series from the fitted model.

C. Inference

First, we describe the estimation of the parameter $\boldsymbol{\Theta}$. In this work, we assume that the probability of selecting a disease that needs medication is proportional to the frequency of the disease in the MIC record. This is similar to what

correspondence topic models [16], [17] assume. Under this assumption, we define the probability θ_{rd} of selecting a disease d requiring medication from a MIC record r as follows:

$$\theta_{rd} = \frac{N_{rd}}{N_r}\,, \tag{2}$$

where N_{rd} is the frequency of the disease d in the MIC record r; therefore, $\sum_{d=1}^{D} N_{rd}$ equals N_r, the number of diseases appearing in r. Clearly, θ_{rd} defined above satisfies the constraint described in Section IV-A (*i.e.*, $\theta_{rd} = 0$ if $d \notin \boldsymbol{d}_r$). This formulation is based on our rationale that the more times a patient is diagnosed as a disease, the more frequently medicines are prescribed to treat the disease.

Next, we describe the estimation of the rest parameters $\boldsymbol{\eta}$ and $\boldsymbol{\Phi}$. Taking Equation (1) as the likelihood with respect to these two parameters, the log-likelihood $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\Phi}) \equiv \log P(\mathcal{R} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ with respect to them is given by

$$\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\Phi}) = \underbrace{\sum_{r=1}^{R}\sum_{n=1}^{N_r} \log \eta_{d_{rn}}}_{\equiv \mathcal{L}(\boldsymbol{\eta})} + \underbrace{\sum_{r=1}^{R}\sum_{l=1}^{L_r} \log \sum_{d=1}^{D} \theta_{rd}\phi_{dm_{rl}}}_{\equiv \mathcal{L}(\boldsymbol{\Phi})}\,. \tag{3}$$

The parameter $\boldsymbol{\eta}$ can be estimated by maximizing $\mathcal{L}(\boldsymbol{\eta})$ with the method of Lagrange multipliers. However, the estimate of the parameter $\boldsymbol{\Phi}$ cannot be analytically obtained from $\mathcal{L}(\boldsymbol{\Phi})$. Thus, we use the EM algorithm to alternately update $\boldsymbol{\Phi}$ and a so-called responsibility $\boldsymbol{Q}$ at each iteration. The responsibility $q_{rld} \in \boldsymbol{Q}$ represents the probability of selecting a disease d as the one that needs the l-th medicinal treatment in a MIC record r and satisfies $q_{rld} \geq 0$ and $\sum_{d=1}^{D} q_{rld} = 1$. In summary, the estimates of η, ϕ_{dm}, and q_{rld} are obtained as follows:

$$\eta_d = \frac{\sum_{r=1}^{R} N_{rd}}{\sum_{d'=1}^{D}\sum_{r=1}^{R} N_{rd'}}\,, \tag{4}$$

$$\phi_{dm} = \frac{\sum_{r=1}^{R}\sum_{l=1}^{L_r} q_{rld}\mathbb{1}(m_{rl} = m)}{\sum_{m'=1}^{M}\sum_{r=1}^{R}\sum_{l=1}^{L_r} q_{rld}\mathbb{1}(m_{rl} = m')}\,, \tag{5}$$

$$q_{rld} = \frac{\theta_{rd}\phi_{dm_{rl}}}{\sum_{d'=1}^{D}\theta_{rd'}\phi_{d'm_{rl}}}\,, \tag{6}$$

where $\mathbb{1}(p)$ is an indicator function that returns 1 if the predicate p is true and 0 otherwise.

D. Time-Series Reproduction

To obtain time series $\mathcal{X}_{\mathrm{P}} \in \mathbb{R}^{D\times M\times T}$ of the number of monthly prescriptions for each disease-medicine pair, we apply our medication behavior model to individual monthly MIC datasets. More specifically, $x_{dmt} \in \mathcal{X}_{\mathrm{P}}$, the number of prescriptions of a medicine m for a disease d in a time t (on a monthly basis), is estimated as follows:

$$x_{dmt} = \sum_{r=1}^{R^{(t)}}\sum_{l=1}^{L_r^{(t)}} q_{rld}^{(t)}\ \mathbb{1}(m_{rl}^{(t)} = m)\,. \tag{7}$$

Disease time series $\mathcal{X}_{\mathrm{D}} \in \mathbb{R}^{D\times T}$ and medicine time series $\mathcal{X}_{\mathrm{M}} \in \mathbb{R}^{M\times T}$ can also be reproduced from $\mathcal{X}_{\mathrm{P}}$. Let $x_{dt} \in \mathcal{X}_{\mathrm{D}}$ be the number of medical treatments for a disease d in a time t and $x_{mt} \in \mathcal{X}_{\mathrm{M}}$ be the number of medical treatments with a medicine m in t. We estimate these counts by

$$x_{dt} = \sum_{m=1}^{M} x_{dmt}\,, \quad x_{mt} = \sum_{d=1}^{D} x_{dmt}\,. \tag{8}$$

Figure 2b shows the prescription time series that our model reproduced for the disease-medicine pairs explained in Section III-A. As shown in the figure, the prescription counts of the medicine that is not effective for hypertension were predicted to be nearly zero. For the medicine with efficacy, in contrast, we can observe that its predicted prescription time series is almost identical to the cooccurrence-based prescription time series (shown in Figure 2a).

V. Trend Change Detection

To detect the change in prescription trends of the time series reproduced by the procedure in Section IV-D, we incorporate the components described in Section III-B into a state space model.

A. Formulation

We represent the dynamics of the time series $\{x_{qt}\}_{t=1}^{T}$ of either diseases ($q = d$), medicines ($q = m$) or prescriptions ($q = (d, m)$) with the following state space model:

$$\begin{aligned}
x_{qt} &= \mu_{qt} + \gamma_{qt1} + \lambda_q w_{qt} + \epsilon_{qt}\,,\\
\mu_{q,t+1} &= \mu_{qt} + \xi_{qt}\,,\\
\gamma_{q,t+1,s} &= \begin{cases} -\sum_{s'=1}^{11}\gamma_{qts'} + \omega_{qt} & (s = 1)\,,\\ \gamma_{qt,s-1} & (s \in \{2,\ldots,11\})\,,\end{cases}\\
\epsilon_{qt} &\sim N(0,\sigma_\epsilon^2),\ \xi_{qt} \sim N(0,\sigma_\xi^2),\ \omega_{qt} \sim N(0,\sigma_\omega^2)\,,
\end{aligned} \tag{9}$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. The observation equation (the first in Equations (9)) decomposes the given time series into four components: level ($\mu.$), seasonality ($\gamma.$), intervention ($\lambda.w.$), and irregularity ($\epsilon.$). We explain each component below.

Level. This is intended to express slow change in time series, which cannot be explained by the other three. This level is similar to the intercept in linear regression, but it may vary over time.

Seasonality. This captures periodic change in time series. Remember that diseases in Figure 3a have a 12-month periodicity in their prescription time series. Thus, we express the dynamics of the seasonality component with 11 state equations.

Intervention. This component is added to capture structural change in time series. This component consists of two variables: $w.$, a dummy value indicating the presence or absence of the change, and $\lambda.$, the scale of the change. As described in Section III-B, we are interested particularly in the structural change due to the new medicine and new indication effects, both of which usually occur at most once and, if so, cause an increase in the slope of a time series (Figures 3b and 3c). Thus, we allow for a single change point for each time series and use the slope shift [22] to model its structural change.

Algorithm 1 Find exact change point of time series $\{x_{qt}\}_{t=1}^{T}$ with exhaustive search.

```
best_point ← NULL, best_aic ← ∞
for each change point t ∈ {1, ..., T, ∞} do
    aic ← AIC({x_qt}_{t=1}^{T}, t)  ▷ AIC value of our model fitted with t
    if aic ≤ best_aic then
        best_point ← t, best_aic ← aic
return best_point
```

Algorithm 2 Find approximate change point of time series $\{x_{qt}\}_{t=1}^{T}$ with binary search.

```
function FindWithin(left, right)
    if right − left ≤ 1 then
        return arg min_{t∈{left,right}} AIC({x_qt}_{t=1}^{T}, t)
    middle ← (left+right)/2
    if AIC({x_qt}_{t=1}^{T}, left) < AIC({x_qt}_{t=1}^{T}, right) then
        return FindWithin(left, middle)
    else
        return FindWithin(middle, right)
best ← FindWithin(1, T)
return arg min_{t∈{best,∞}} AIC({x_qt}_{t=1}^{T}, t)
```

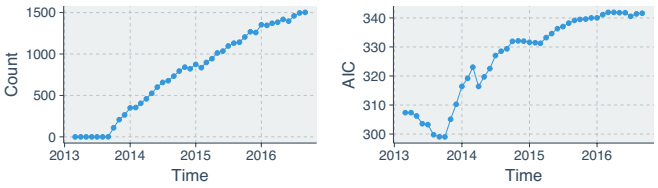
More specifically, we define w_{qt}, a dummy value at a time point t for time series having the change point t_{CP} (defined as ∞ if no change point exists), as $t - t_{\mathrm{CP}} + 1$ if $t \geq t_{\mathrm{CP}}$) and 0 otherwise. In this work, we assume that the scale of the structural change is constant over time for simplicity.

Irregularity. The last terms in the observation and state equations represent the irregularity of these components. These terms allow the value of each component to vary over time, which improves the flexibility the model. In addition, outliers may exist in time series data, as mentioned in Section III-B. As these abnormal values are absorbed into the irregularity term $\epsilon_{\cdot}$ of the observation equation, our trend change detection is robust to noise.

B. Inference

Given the time series $\{x_{qt}\}_{t=1}^{T}$ and its change point t_{CP}, we can efficiently estimate the parameters of our state space model by using the Kalman filter [21]. Thus, the remaining problem to be solved is finding the change point for given time series. In particular, we would like to find it using a fully automated approach without any hyperparameter, because human intervention is unrealistic for the massive number of time series (*e.g.*, more than 200 thousand prescriptions in our experiments).

To this end, we use Akaike's Information Criterion (AIC) [27] as a criterion for the automatic change point detection. AIC measures the quality of statistical models on the basis of both the likelihood and the number of parameters of the model. A lower AIC value indicates better quality. Our choice of this criterion is based on the following: it allows for a fair comparison between multiple models with different numbers of parameters [22], has been commonly used for model selection [18], [28], and performs at least as well as its alternatives (*e.g.*, the Bayesian Information Criterion (BIC) [29]) [30]. Note, however, that our solutions presented below can work with other criteria for model selection.



(a) Time series with change point in September 2013

(b) AIC values of models fitted with different interventions

Fig. 5: The effect of intervention selection on model performance.

Exact Solution. In this approach, we find the change point of the given time series $\{x_{qt}\}_{t=1}^{T}$ by Algorithm 1. This algorithm regards each time point $t \in \{1, \ldots, T\}$ as a candidate change point of $\{x_{qt}\}_{t=1}^{T}$ and fits our model with this assumption. In this way, we obtain the best candidate change point that minimizes the AIC value of the learned model within the entire period. Finally, we compare the AIC values of this model and a model without the intervention component to decide whether the change point really exists.

Approximate Solution. While Algorithm 1 can find the exact solution for the change point of the given $\{x_{qt}\}_{t=1}^{T}$, it conducts an exhaustive search for T. Thus, the computational cost increases linearly with respect to the length of the whole period for which MIC records have been considered. To reduce the search space, we focus on the sensitivity of AIC over change points. Figure 5 shows the AIC values of our models fitted with different intervention points, together with an original time series having the slope change in September 2013. This figure indicates that models with a intervention point near the true change point yield lower AIC values than those far from it. This observation leads us to an idea that we can skip fitting a model with unlikely change points, which is summarized in Algorithm 2. This algorithm behaves similarly to the binary search: it halves the search space at each iteration.

Time Complexity. Let $\mathcal{C}_{\mathrm{KF}}$ be the computational time required for the Kalman filter to fit our model to the given time series. Algorithm 1 requires $\mathcal{O}(\mathcal{C}_{\mathrm{KF}}T)$ time to find the exact solution, while Algorithm 2 requires $\mathcal{O}(\mathcal{C}_{\mathrm{KF}}\log(T))$ time to find the approximate solution. We evaluate the cost-effectiveness of these algorithms in Section VIII.

VI. Dataset

We use the medical insurance claim data of *all* the elderly citizens over 75 years of age who live in the Mie Prefecture, Japan. This dataset was recorded from March 2013 through September 2016 (*i.e.*. 43 months), and then anonymized and disclosed to the authors by Mie Prefectural Association of Medical Care Services for Older Senior Citizens by contract. This association is the only healthcare insurance organization in service to all the elderly citizens over 75 years of age who reside in Mie, which enables us to perform *universal* analysis for the regional elderly healthcare.

Records in our dataset are created on a monthly basis, as described in Section III-A. On average, 3,347 medical institutes, 202,972 patients, 332,167 MIC records, 9,173 diseases, and 9,346 medicines are contained in each monthly MIC dataset $\mathcal{R}^{(t)}$ $(t \in \{1, \ldots, 43\})$. We selected geriatric patients as our

target age group because of the following three reasons. The first reason is the coverage and homogeneity of the data as mentioned above.[4] Second, as elderly people are likely to visit hospitals more frequently than younger people do, we can expect to obtain more MIC records from the former. Third, the duration of this dataset is the longest out of all the ones available for us.[5]

Before fitting our probabilistic medication model to each dataset $\mathcal{R}^{(t)}$, we omitted diseases and medicines that appear less than five times in $\mathcal{R}^{(t)}$, as was done in the existing work on topic modeling [17], [31]. Similarly, when fitting our state space model to the reproduced time series, we omitted those whose total frequency during the said period is less than 10 to ensure the reliability of the model fitting. These filtering processes reduced the numbers of diseases, medicines, and prescriptions for which the trend change detection is applied to 3,978, 7,474, and 206,829, respectively.

VII. Applications

This section introduces three applications for our universal medical analysis.

A. Temporal Prescription Change Detection

As described in Section I, detecting the change in prescription trends has various practical applications. We demonstrate the effectiveness of our state space model in detecting important trend changes in reproduced time series data. Figures 6 and 7 show the fitting results for six time series reproduced from MIC records with Equations (7) and (8). For each case example, the top figure shows the original (*i.e.*, x_{qt}) and fitted (*i.e.*, $x_{qt} - \epsilon_{qt}$) time series, while the middle contains three components (*i.e.*, μ_{qt}, γ_{qt1}, and $\lambda_q w_{qt}$) decomposed by our model. For comparison, we also plotted several time series related to the original one at the bottom.

Seasonality and outliers. Our model successfully identified the seasonality of influenza, as shown in the middle graph in Figure 6a. This time series contains the spike around the winter in 2015 due to its outbreak in that season. As this sudden increase is a temporal effect, our model treated it as an outlier for better fitting. Figure 6b shows the time series of diarrhea, which is diagnosed frequently as seasons change. The figure demonstrates that our model was able to capture the seasonality having more than one peak per year.

Medicine-derived structural changes. Figure 6c shows the fitting result for a new medicine for osteoporosis, which has been released in August 2013. We can observe from the middle graph that our model detected the release date accurately. The bottom plot draws the time series of medicines with the same indication that were less used after the release of the new medicine. Figure 6d is an example of the sudden decreasing trend in medicine time series. This is attributed to the release of generic medicines, whose usage started to increase around the same time as shown in the bottom.

Prescription-derived structural changes. Our model also found the new indication effect on Lewy body dementia, as shown in Figure 7a. Another type of prescription-derived structural changes detected by ours is shown in Figure 7b. The top graph draws the time series of a medicine for oral feeding difficulty. This is not a new medicine because it was prescribed to other diseases as shown in the bottom graph. Remarkably, the time series of dehydration (denoted as "related1") in this graph exhibits the opposite trend, suggesting that patients with the same or similar symptoms might be diagnosed as different diseases according to times.

B. Geographical Prescription Spread Visualization

While the application in Section VII-A focuses on the *temporal* change in prescription trends, understanding *geographical* spread of prescriptions also plays an important role in local governments' managing medicine supply and demand. To investigate the geographical prescription spread, we divided our dataset on the basis of the city of each hospital where MIC records are created. Then, we learned our probabilistic medication model for MIC records in each city separately.

Taking the original anti-platelet medicine in Figure 6d as an example again, we visualized the prescription counts of the original and generic medicines at each city to see how the latter spread geographically after their release. We can observe from Figure 8 that Generic-3 was used most frequently among the three in the first month after the release and still kept its popularity at a high level one year later. This is probably because Generic-3 is an authorized generic,[6] which may be more acceptable for patients than other generics. Another finding from this analysis is that the original medicine still dominated the market in some cities (*e.g.*, the northernmost area) even after the release of the generics. As generics are typically cheaper than their original medicines, this application could be used to find areas where medical costs can be reduced by encouraging local hospitals to use generics more frequently.

C. Inter-Hospital Prescription Gap Analysis

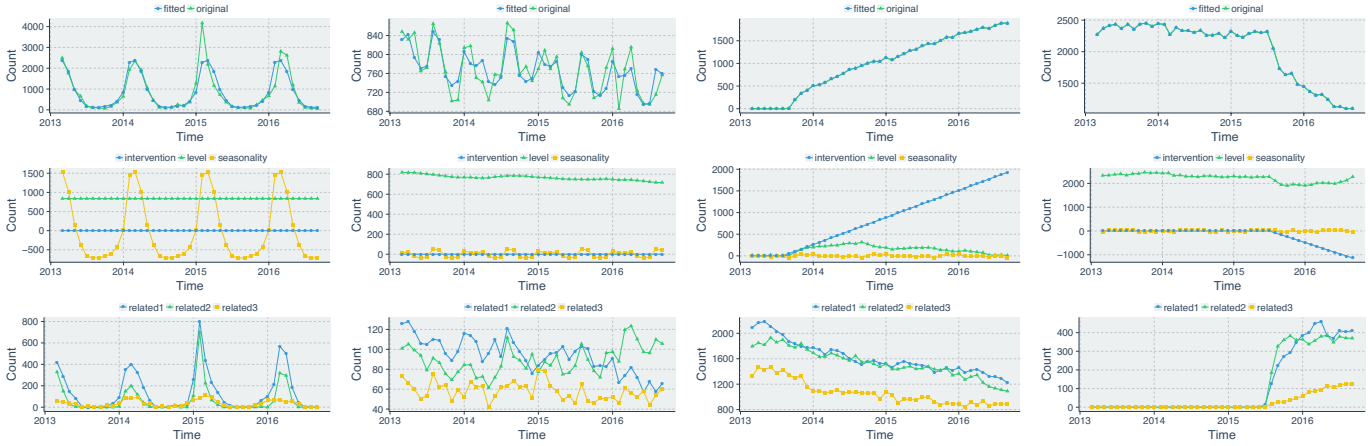
The number of big hospitals that provide advanced medical treatments is limited. There are many small hospitals at which only a few physicians diagnose local residents. Understanding difference in prescriptions between these hospitals is of practical importance to close a inter-hospital gap in the quality of medical care. As an application for this purpose, we analyzed the prescription trends of different-sized hospitals. More specifically, we grouped hospitals into three classes based on the number of beds: *small* for $[0, 20)$, *medium* for $[20, 400)$, and *large* for $[400, \infty)$.[7] Then, we learned our probabilistic

[4]The other citizens, below 75 years of age, enroll in different public healthcare insurances according to their residential addresses, employers, and income levels.

[5]We also applied our probabilistic medication model to a smaller dataset that contain the MIC records of younger patients. As a result, the frequently prescribed medicines for major diseases (such as influenza, hay fever, and hypertension) that our model learned were consistent between the different datasets. In this paper, we only report our results for the larger dataset due to space limitation.

[6]Authorized generics are identical to their original medicine in not only active ingredients but also inactive ones and manufacturing process, *etc*.

[7]In Japan, hospitals in the small and large classes are called clinics and advanced treatment hospitals, respectively.



(a) Influenza (b) Diarrhea (c) New medicine for osteoporosis (d) Original anti-platelet medicine

Fig. 6: Fitting results of our state space model for disease and medicine time series. Top: original vs. fitted time series. Middle: components decomposed by our model. Bottom: time series related to original one.

(a) New indication of Lewy body dementia (b) Possible trend change in diagnostics

Fig. 7: Fitting results of our state space model for prescription time series, arranged in the same way as Figure 6.

medication model for MIC records in each class separately.

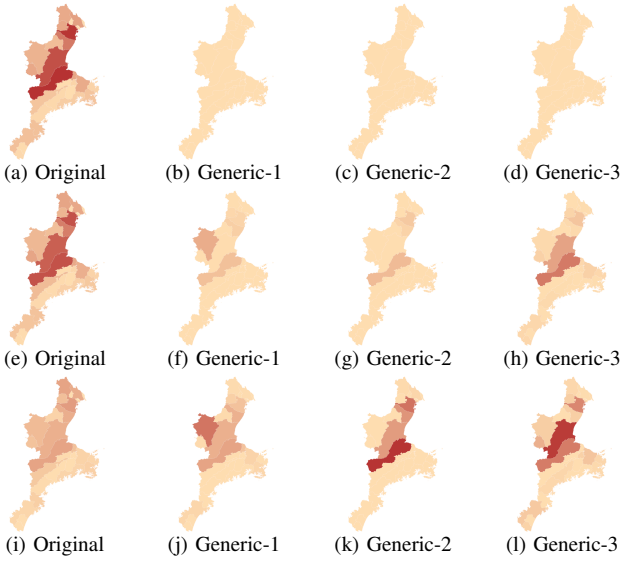
Table II shows the top 10 frequent diseases for which an antibiotic was prescribed at hospitals in each class. Despite the fact that antibiotics have efficacy for bacteria-caused diseases, small hospitals more frequently prescribed this medicine for cold syndrome (*e.g.*, acute upper respiratory inflammation) and influenza, both of which are mostly caused by viruses. The abuse of antibiotics increases not only medical costs but also risk for drug resistance. this application could help national governments decide to which hospitals they should announce the proper use of medicines.

VIII. Experiments

To evaluate our models proposed in Sections IV and V, we conducted experiments using real MIC data. Our experiments were designed to answer the following questions:

Q1 *Accuracy*: How accurately can our probabilistic medication model predict prescription links missing from the MIC data?

Q2 *Usefulness*: How well can our state space model explain the dynamics of the prescription time series?



(a) Original (b) Generic-1 (c) Generic-2 (d) Generic-3

(e) Original (f) Generic-1 (g) Generic-2 (h) Generic-3

(i) Original (j) Generic-1 (k) Generic-2 (l) Generic-3

Fig. 8: Geographical spread of anti-platelet medicines (same as Figure 6d). Top: one month before the release of generic medicines. Middle: one month later. Bottom: one year later.

Q3 *Efficiency*: How efficient is our state space model with binary search and how does the detection performance change?

All the experiments were run on a Linux machine with two 10-Core 2.40GHz Intel Xeon (E7-2870) CPUs and 512GB of memory. In what follows, significant effects are reported on the significant level $\alpha = 0.05$.

A. Q1: Accuracy

To directly answer the first question Q1, which is about the accuracy of our probabilistic medication model, we need true prescription links between diseases and medicines in each MIC record, which are, however, difficult to obtain. Therefore, we decided to evaluate our model from different perspectives and conducted two experiments about predictive performance and prescription relevance.

Baselines. The common baseline (denoted as Cooccurrence) we used in both experiments is to predict prescriptions be-

TABLE II: Top 10 frequent diseases for which an antibiotic is prescribed at small, medium, and large hospitals.

(a) Small hospitals

Disease	Ratio (%)
acute bronchitis	31.996
bronchitis	12.059
acute upper respiratory inflammation	9.750
allergic rhinitis	5.198
other	4.975
influenza	3.303
pharyngitis	3.202
acute pharyngolaryngitis	2.667
chronic sinusitis	2.481
acute pharyngitis	2.400

(b) Medium hospitals

Disease	Ratio (%)
acute bronchitis	24.526
chronic bronchitis	12.632
bronchitis	7.233
nontuberculous mycobacterial infection	5.923
chronic sinusitis	5.260
allergic rhinitis	5.059
Helicobacter pylori infection	4.494
bronchiectasis	4.059
pharyngitis	2.150
pneumonia	1.602

(c) Large hospitals

Disease	Ratio (%)
chronic sinusitis	10.626
nontuberculous mycobacterial infection	9.336
acute bronchitis	8.536
bronchiectasis	6.707
allergic rhinitis	3.997
chronic bronchitis	3.976
pneumonia	3.779
Helicobacter pylori infection	3.568
pharyngitis	2.725
diffuse panbronchiolitis	1.418

TABLE III: Mean (and SD) of predictive performance (measured by medicine perplexity) and prescription relevance (measured by AP@10 and NDCG@10).

	Perplexity		AP@10	NDCG@10
Unigram	2315.083	(103.395)	—	—
Cooccurrence	168.241	(7.408)	0.304 (0.243)	0.450 (0.260)
Proposed	**112.436**	(4.480)	**0.787** (0.298)	**0.835** (0.288)

tween diseases and medicines on the basis of disease-medicine cooccurrences, as mentioned in Section III-A. More specifically, instead of Equation (5), this method uses the following equation to predict the parameter $\Phi_d = (\phi_{d1}, \ldots, \phi_{dM})$ of the medicine generation distribution for each disease d:

$$\phi_{dm} = \frac{\sum_{r=1}^{R} \sum_{l=1}^{L_r} \mathrm{Cooc}_r(d, m)}{\sum_{m'=1}^{M} \sum_{r=1}^{R} \sum_{l=1}^{L_r} \mathrm{Cooc}_r(d, m')}, \qquad (10)$$

where $\mathrm{Cooc}_r(d, m)$ is the frequency of cooccurrences between a disease d and a medicine m in a MIC record r. In the experiment about predictive performance, we also used the unigram model of medicines [32] (denoted as Unigram) as another baseline. The comparison with Unigram and Cooccurrence allows us to validate the effectiveness of the disease-dependent medicine prescription and the medication target identification in our model, respectively.

1) Predictive Performance: The first experiment aims to evaluate the predictive performance of our probabilistic medication model.

Settings. In this experiment, we sampled 90% medicines from each MIC record to train the proposed and baseline models and tested these models with the remaining 10% of medicines. We used perplexity as our evaluation measure, which is widely used to evaluate the predictive performance of statistical models (*e.g.*, [16], [17]). A lower perplexity indicates better predictive performance. Letting $\boldsymbol{m}'_r = \{m'_{rl}\}_{l=1}^{L'_r}$ be a bag of test medicines in each MIC record r, the perplexity $\mathrm{PPL}(\mathcal{M})$ of a trained model $\mathcal{M}$ is given by

$$\mathrm{PPL}(\mathcal{M}) = \exp\left(-\frac{\sum_{r=1}^{R} \sum_{l=1}^{L'_r} \log P(m'_{rl} \mid \mathcal{M})}{\sum_{r=1}^{R} L'_r}\right). \qquad (11)$$

Results. For each monthly MIC dataset, we trained individual models and measured the perplexity for these models. Table III shows the mean and standard deviation (SD) of the perplexity scores for each model. Unigram performed poorly. Its mean perplexity was 20 times higher than ours. On average, our model achieved about two-thirds as much perplexity as the Cooccurrence model. In fact, ours beat this baseline in terms of perplexity for every monthly dataset. We also conducted a paired t-test, which revealed a significant difference between these models ($t(42) = -103.670,\ p < 0.001,$ Cohen's $d = -15.810$). In summary, this experiment uncovered that our model simulating physicians' medication behavior was the most effective among the three.

2) Prescription Relevance: The second experiment aims to compare our model with the baseline model in terms of the relevance of estimated prescriptions.

Settings. In this experiment, we first selected 100 most frequent diseases over the entire period for which the MIC records have been obtained. Then, for each model and each frequent disease d, we ranked medicines in the descending order of $x_{dm} = \sum_{t=1}^{T} x_{dmt}$, which is the total prescription count of a medicine m for d. Finally, we evaluated the relevance of the resulting ranking at the cutoff $K = 10$ with the two measures common in the information retrieval community: Average Precision (AP) [33], [34] and Normalized Discounted Cumulative Gain (NDCG) [35]. Note that a higher score indicates a better ranking for both measures.

Ground Truth. The aforementioned evaluation required ground truth data about the relevance of prescriptions (1,591 in total). To prepare such data, one author of this paper assessed the package insert of each medicine m and judged its relevance to each disease d on the basis of the following objective criterion: *the prescription is relevant if d or d's hypernym is described in the indications and/or therapeutic category fields in the package insert of m.* The assessment resulted in 1,154 prescriptions that were judged as relevant ($= 1$) or irrelevant ($= 0$). However, the remaining 437 prescriptions were left untouched because of the difficulty in judging relevance for those without domain knowledge. To overcome this issue, we asked a medical professional to investigate the relevance of these prescriptions on the basis of his expertise.[8] As a result, we obtained 1,528 prescriptions being labeled. The remaining 63 unlabeled prescriptions were treated as irrelevant.

Results. Figure III shows the mean and SD of AP@10 and NDCG@10 scores for each model. Our model largely improved the prescription relevance over the Cooccurrence model in either measure. Examining individual evaluation scores, we found that this baseline beat ours only once in terms of AP@10 and only twice in terms of NDCG@10 out of 100 ranking pairs. The differences between these two models were shown to be significant by paired t-tests

[8]We also shared the already assessed prescriptions with the medical professional for the purpose of double checking.

TABLE IV: Mean (and SD) of fitting quality (measured by AIC) for disease, medicine, and prescription time series.

	disease	medicine	prescription
Local Level (LL)	326.350 (64.010)	277.238 (80.352)	119.305 (76.138)
LL + Seasonality (S)	254.018 (44.937)	218.295 (56.216)	103.594 (54.999)
LL + Intervention (I)	316.905 (71.306)	268.716 (83.155)	107.579 (80.608)
LL + S + I (proposed)	**244.603** (44.937)	**208.396** (56.441)	91.888 (50.618)
ARIMA	286.416 (55.975)	242.001 (70.856)	**87.888** (302.026)

($t(99) = 15.398,\ p < 0.001,$ Cohen's $d = 1.540$ for AP@10; $t(99) = 14.374,\ p < 0.001,$ Cohen's $d = 1.437$ for NDCG@10). The advantage of our model that we illustrated with an example in Section III-A was empirically demonstrated by this experiment.

B. Q2: Usefulness

This subsection answers the second question Q2 about the usefulness of our state space model. To this end, we evaluated the fitting quality and forecast performance of our approach.

1) Fitting Quality: First, we measured the fitting quality of our model by AIC. To identify the contribution of each component in our full model, we used its simpler variants for comparison: Local Level (LL), where neither the seasonal nor intervention component exists, LL with seasonality (LL + S), and LL with intervention (LL + I). We fitted these models to each time series. In addition, we used as a baseline the ARIMA model, where we determined the optimal parameters by using AIC. Table IV summarizes the mean and SD of the AIC values of those models.

LL, which is the simplest variant of our model, achieved the least performance consistently. Both LL + S and LL + I contributed to the improvement of the fitting quality: the seasonality component was in particular effective for disease time series while the intervention component decreased the AIC values for all types of time series to an equal degree. This supports our observation that the seasonality is a disease-dependent factor while both diseases and medicines affect the structural change of prescription time series (Section III-B).

Our full model (LL + S + I) achieved the best performance for disease and medicine time series. Compared with the second best model (LL + S), it decreased the mean AIC value by about 10 for each type of time series. Conducting paired t-tests, we found significant differences between these models ($t(3977) = -36.619,\ p < 0.001,$ Cohen's $d = -0.581$ for diseases; $t(7473) = -49.829,\ p < 0.001,$ Cohen's $d = -0.576$ for medicines), indicating the importance of capturing both the seasonality and structural change in these kinds of time series data. In fact, our model identified change points in the time series for 12%, 28%, and 10% of diseases, medicines, and prescriptions, respectively.

For prescription time series, our full model achieved the second best performance, which was comparable to the best model (*i.e.*, ARIMA). While the difference between these models was shown to be significant ($t(206800) = 6.107,\ p < 0.001$), the effect size was negligible (Cohen's $d = 0.013$). The AIC variance of the ARIMA model was much larger, indicating that its fitting quality was not as stable as ours. In addition, the ARIMA model, unlike our model, has no ability to explain the cause of prescription trend changes.

TABLE V: Computational time (in minutes) required to fit models for all time series. Values in parentheses indicate the increased computation rate from our model without the intervention variables.

	disease	medicine	prescription
Exact Solution	8.529 (27.878)	17.565 (29.900)	562.614 (35.492)
Approximate Solution	1.832 (5.989)	3.678 (6.260)	117.308 (7.400)

2) Forecast Performance: While our main focus in this paper is *detecting* prescription trend changes from given time series, *forecasting* future prescriptions is also a problem of practical importance. Thus, we also investigated the forecast performance of our state space model. Again, we used as a baseline the ARIMA model with the AIC-based optimal parameters. We used the data from the first 31 months for training and the remaining 12-month data for forecasting.

Overall, the forecast error of these two models was comparable to each other: the median of Root Mean Squared Errors (RMSEs) for (normalized) disease time series was 0.169 for the ARIMA model and 0.187 for our model. However, we found that ARIMA made less stable forecasts than ours. Figure 9 shows the forecasting results of the two models for five such time series, of which two have seasonality and three have structural breaks. ARIMA failed to forecast seasonal patterns in the testing period. It also worked unstably for time series having structural breaks near the end of the training period. In contrast, our model made accurate forecasts for both cases, indicating that its seasonal and intervention components are effective for forecasting as well as fitting.

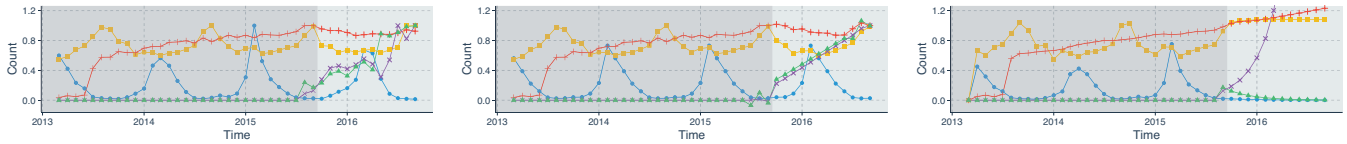
C. Q3: Efficiency

To answer the last question Q3, which is about the efficiency of our state space model, we compared the performance of our model with the exact change point detection (Algorithm 1) and that with the approximate change point detection (Algorithm 2) in terms of the cost-effectiveness.

1) Computational Time: First, we investigated the computational cost of these models. More specifically, we measured the total computation time required to fit these models for the entire set of time series.

Table V shows the results together with the increased computation rate against the computational time of our model without the intervention variables. It is observed that the approximate change point detection greatly decreased the computation time compared with the exact change point detection. Theoretically, the exact and approximate solutions take $\mathcal{O}(\mathcal{C}_{\mathrm{KF}} T)$ and $\mathcal{O}(\mathcal{C}_{\mathrm{KF}} \log(T))$ times, respectively (Section V-B). Note that the computational cost of the Kalman filter can be regarded as constant given the fixed duration of the time series ($T = 43$ in our experiments). Thus, the expected increase rates of the exact and approximate solutions would be 43 and $\log_2(43) \approx 5.426$, respectively. Our experimental results almost agreed with these theoretical values.

2) Approximate Accuracy: Next, we evaluated the accuracy of our approximate solution that detects change points. This was done by comparing the change points detected by this algorithm with those by the exact algorithm.



(a) Original time series (normalized) (b) Forecast by our model (c) Forecast by ARIMA

Fig. 9: Forecasting results. We used the first 31 months for training and the remaining 12 months for forecasting.

TABLE VI: Change point consistency between our exact and approximate methods.

(a) Disease

		Approximate	
		pos.	neg.
Exact	pos.	423	40
	neg.	0	3,515

(b) Medicine

		Approximate	
		pos.	neg.
Exact	pos.	1,944	154
	neg.	0	5,376

(c) Prescription

		Approximate	
		pos.	neg.
Exact	pos.	19,106	2,079
	neg.	0	185,644

Table VI shows the change point consistency between these two algorithms. It is worth noting that no false-positive case exists in the table for every type of time series, which is due to the nature of Algorithm 2. In addition, the rate of false-negative discoveries is also very low (8.639% for diseases, 7.340% for medicines, and 9.814% for prescriptions). The measured Cohen's κ values were 0.949 for diseases, 0.948 for medicines, and 0.943 for prescriptions, indicating strong agreement in the change point detection by the exact and approximate algorithms.

We also measured RMSE values between the exact change points and the approximate ones. The RMSE values for disease, medicine, and prescription time series were 3.862, 7.154, and 4.481, respectively. Given the period of our dataset (*i.e.*, $T = 43$), the approximate algorithm found change points with a reasonable degree of accuracy especially for diseases and prescriptions.

When we used Algorithm 2 for change point detection, the mean AIC values of our model for disease, medicine, and prescription time series were 244.742, 209.076, and 92.099, respectively. Our approximate solution achieved a comparable fitting quality to the exact solution by Algorithm 1 (the last row in Table IV).

IX. Discussion

In this paper, we proposed a probabilistic medication model to predict missing links in MIC data. It was shown to be accurate by our experiments. In terms of both subjective and objective evaluations, our model significantly outperformed the cooccurrence-based baseline. This suggests that the generative process defined by our model was reasonable to capture the medication behavior by real physicians. Achieving high accuracy in this step is crucial to reliably detect prescription trend changes from reproduced time series. While we trained our model for each monthly dataset separately, modeling the evolution of disease and medicine distributions at consecutive months (as Dynamic Topic Model [36] and Topic Tracking Model [37] do) and/or geographical differences in those distributions (by applying location-aware topic models [38]) could further improve the performance of missing link prediction, which would be a promising direction to extend this study.

We also proposed a state space model with seasonal and intervention components to detect prescription trend changes. Our full model achieved the better fitting quality in terms of AIC than a model without these components. As shown in Section VII-A, our model successfully identified the various types of prescription trend changes (*e.g.*, periodic changes due to disease seasonality and structural breaks due to new medicines and new indications). An interesting finding from this analysis is that some time series have the *early signs* of structural breaks (*e.g.*, in Figure 7, the small number of initial prescriptions exist before the prevalence). Can we predict the future growth of a prescription from its initial behavior? While this paper focuses mainly on detecting time series having prescription trend changes, building a forecast model for prescription time series would also be worth exploring in future. Our model could be used as a foothold for this purpose as we exemplified in Section VIII-B2.

There are several limitations that we should acknowledge for this work. First, we assume at most one change point for each time series because our main focus is structural changes due to the new medicine and new indication effects, both of which usually occur at most once. In reality, however, more than one change point can exist in time series. This may be a possible explanation of why our model did not outperform the ARIMA model for prescription time series, which tended to have a zigzag shape due to data sparsity. It is worth examining whether the fitting quality improves by allowing for multiple change points. As state space models can accept more than one intervention variable, we can extend our model in that way. Second, we formulated our state space model with linear equations and Gaussian distributions to make our problem simple and tractable. To capture more realistic trends, more sophisticated techniques (*e.g.*, non-linear and non-Gaussian state space models [39] and deep learning models [40]) are worth considering. Other directions for future work include modeling the co-evolution [41] of medicines, making our solution more efficient, and experimenting with other datasets for different populations and countries.

X. Conclusions

In this paper, we addressed the problem of detecting the change in prescription trends and identifying its cause. To our knowledge, this work is the first attempt to use MIC data for this purpose. We proposed a two-fold approach and evaluated its effectiveness through extensive experiments with the real data consisting of 3.5-year MIC records. Our approach was shown to be

1) *accurate*: Our probabilistic medication model performed significantly better than a cooccurrence-based baseline

in terms of both predictive capability and prescription relevance;

2) *useful*: Our state space model successfully detected the change in prescription trends due to new medicine and new indication effects, *etc*. and could be used for practical applications such as geographical prescription trend visualization and inter-hospital prescription gap analysis;
3) *efficient*: Our approximate algorithm reduced the computational cost while detecting most change points correctly.

In addition to the temporal prescription change detection, we also demonstrated the geographical prescription spread visualization and the inter-hospital prescription gap analysis as promising applications for MIC big data.

Our future directions include improving our model so that it can discover more complex changes in prescription trends and accurately forecasting the spatiotemporal growth of prescriptions from their initial trends. We are also interested in leveraging the MIC data for other challenging tasks for population-scale healthcare.

ACKNOWLEDGMENTS

This work has been in part supported by Health Labour Sciences Research Grant (Ministry of Health Labour and Welfare, Japan), Funding Program for World-Leading Innovative R&D on Science and Technology (Cabinet Office, Japan) and Impulsing Paradigm Change through Disruptive Technologies Program (Cabinet Office, Japan).

REFERENCES

[1] Y. Matsubara, Y. Sakurai, W. G. van Panhuis, and C. Faloutsos, "FUNNEL: Automatic mining of spatially coevolving epidemics," in *KDD*, pp. 105–114, 2014.

[2] T. T. Ashburn and K. B. Thor, "Drug repositioning: Identifying and developing new uses for existing drugs," *Nature Reviews Drug Discovery*, vol. 3, pp. 673–683, 2004.

[3] H. Iwata, R. Sawada, S. Mizutani, and Y. Yamanishi, "Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data," *Journal of Chemical Information and Modeling*, vol. 55, no. 2, pp. 446–459, 2015.

[4] R. Sawada, H. Iwata, S. Mizutani, and Y. Yamanishi, "Target-based drug repositioning using large-scale chemical-protein interactome data," *Journal of Chemical Information and Modeling*, vol. 55, no. 12, pp. 2717–2730, 2015.

[5] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun, "Rubik: Knowledge guided tensor factorization and completion for health data analytics," in *KDD*, pp. 1265–1274, 2015.

[6] U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger, "X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words," *IEEE Transactions on Medical Imaging*, vol. 30, no. 3, pp. 733–746, 2011.

[7] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Y.-X. Wang, P.-X. Lu, and C. J. McDonald, "Automatic tuberculosis screening using chest radiographs," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, 2014.

[8] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation," in *CVPR*, pp. 2497–2506, 2016.

[9] J. Paparrizos, R. W. White, and E. Horvitz, "Detecting devastating diseases in search logs," in *KDD*, pp. 559–568, 2016.

[10] N. Mishra, R. W. White, S. Ieong, and E. Horvitz, "Time-critical search," in *SIGIR*, pp. 747–756, 2014.

[11] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: Detecting influenza epidemics using Twitter," in *EMNLP*, pp. 1568–1576, 2011.

[12] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[13] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98–101, 2008.

[14] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[15] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *KDD*, pp. 542–550, 2008.

[16] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR*, pp. 127–134, 2003.

[17] T. Iwata, T. Yamada, and N. Ueda, "Modeling noisy annotated data with application to social annotation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1601–1613, 2013.

[18] A. J. Bagnall and G. J. Janacek, "Clustering time series from arma models with clipped data," in *KDD*, pp. 49–58, 2004.

[19] B. Biller and B. L. Nelson, "Modeling and generating multivariate time-series input processes using a vector autoregressive technique," *ACM Transactions on Modeling and Computer Simulation*, vol. 13, no. 3, pp. 211–237, 2003.

[20] A. Jain, E. Y. Chang, and Y.-F. Wang, "Adaptive stream resource management using Kalman filters," in *SIGMOD*, pp. 11–22, 2004.

[21] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[22] J. Commandeur and S. J. Koopman, *An Introduction to State Space Time Series Analysis*. Oxford University Press, 2007.

[23] V. Guralnik and J. Srivastava, "Event detection from time series data," in *KDD*, pp. 33–42, 1999.

[24] J. Kleinberg, "Bursty and hierarchical structure in streams," in *KDD*, pp. 91–101, 2002.

[25] K. Yamanishi and J.-i. Takeuchi, "A unifying framework for detecting outliers and change points from non-stationary time series data," in *KDD*, pp. 676–681, 2002.

[26] Y. Zhu and D. Shasha, "Efficient elastic burst detection in data streams," in *KDD*, pp. 336–345, 2003.

[27] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *ISIT*, pp. 267–281, 1973.

[28] R. Jones and K. L. Klinkner, "Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs," in *CIKM*, pp. 699–708, 2008.

[29] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistics. Springer, 2009.

[31] K. Tsukuda, M. Hamasaki, and M. Goto, "Why did you cover that song?: Modeling n-th order derivative creation with content popularity," in *CIKM*, pp. 2239–2244, 2016.

[32] F. Song and W. B. Croft, "A general language model for information retrieval," in *CIKM*, pp. 316–321, 1999.

[33] T. Sakai, "Alternatives to bpref," in *SIGIR*, pp. 71–78, 2007.

[34] S. Robertson, "A new interpretation of average precision," in *SIGIR*, pp. 689–690, 2008.

[35] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[36] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*, pp. 113–120, 2006.

[37] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Topic tracking model for analyzing consumer purchase behavior," in *IJCAI*, pp. 1427–1432, 2009.

[38] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *WWW*, pp. 247–256, 2011.

[39] H. Tanizaki, "Nonlinear and non-Gaussian state space modeling using sampling techniques," *Annals of the Institute of Statistical Mathematics*, vol. 53, no. 1, pp. 63–81, 2001.

[40] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," in *SIGIR*, pp. 95–104, 2018.

[41] Y. Matsubara, Y. Sakurai, and C. Faloutsos, "The Web as a jungle: Non-linear dynamical systems for co-evolving online activities," in *WWW*, pp. 721–731, 2015.