大規模レセプトデータからの投薬トレンドの変化検知

梅本　和俊†　　合田　和生†　　満武　巨裕††　　喜連川　優†

† 東京大学 生産技術研究所　〒 153-8505 東京都目黒区駒場 4-6-1
†† 医療経済研究機構　〒 105-0003 東京都港区西新橋 1-5-11 第 11 東洋海事ビル 2F
E-mail: †{umemoto,kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp, ††mitsutake@ihep.jp

あらまし　本稿では，医療機関が保険者への医療費の請求を目的として毎月作成するレセプトの大規模データから，時間の経過にともない処方の傾向が大きく変化する疾患と医薬品のペアを発見する問題に取り組む．我々はまず，レセプトデータ中で欠損している疾患と医薬品との間の処方関係を潜在変数モデルによって推定することで，疾患と医薬品の各ペアに対する月次処方数の時系列を正確に復元する．次に，干渉変数を含む状態空間モデルによって処方数の時系列を長期変動，周期変動，構造変化，外れ値に分解することで，膨大な数の疾患と医薬品の中から処方傾向に経時変化が生じているものを検出する．三重県の後期高齢者の約 3 年半にわたる診療記録から構成される大規模電子レセプトデータを用いた評価実験によって，正確性，有用性，効率性の観点から提案手法の優位性を示した．

キーワード　レセプト，時系列解析，変化検知，処方推定，医療情報処理

1. はじめに

医薬品医療機器総合機構の報告 [34] によると，我が国では年間 100 件以上の新医薬品 (注1) に関する申請が厚生労働大臣によって承認されている．また，多くの疾患には季節性や流行といった時間変動要因が存在する [21]．そのため，各疾患に対する投薬の傾向は時間の経過とともに変化し得る．こうした投薬トレンドの変化の早期検知は，医療や行政，経済など多様な分野への応用の可能性を秘めた重要な問題である．例えば，新薬の処方傾向の早期把握は，製薬会社による製品の適正な供給戦略の立案や最新の処方知識の普及に役立つ．また，医療現場における投薬傾向の把握は，行政機関が医薬品の適正利用の監視や医療費の適正化を行う上で欠かせない．他にも，トレンド解析を通じて既存薬の新規適応事例が発見できれば，従来の生命情報学的なアプローチ [13], [27] とは異なる，臨床データ駆動のドラッグリポジショニング (注2) の実現可能性を見出だせる．

近年，医療情報解析の対象として電子レセプトへの期待が高まっている [40]．レセプトとは，医療機関が保険者に医療費を請求するために毎月作成する医療報酬の明細書である．我が国では，カルテと同様に電子化が推進されており，平成 29 年 9 月時点での電子レセプトの普及率は医療機関数ベースで 93.2%，請求件数ベースで 98.2%に達している [37]．厚生労働省は匿名化済みの電子レセプト情報等を継続的に収集することでレセプト情報・特定検診等情報データベース（NDB）を構築しており，平成 25 年度からは研究目的での電子レセプトデータの第三者提供が本格的に実施されている．同様の制度は，韓国や台湾といった諸外国でも導入されている [40].

レセプト情報の特筆すべき特徴の 1 つとして悉皆性があげられる．我が国では国民皆保険制度が採用されているため，電子レセプトデータの解析を通じて，個々の医療機関レベルでなく，都道府県や全国レベルでの投薬トレンドの把握が可能となる．従来研究で利用されてきた電子カルテ [31], [41] や X 線画像 [5], [15], [28] といった医療情報では，こうした網羅的な解析を行うことは容易ではない．それは，前者は電子レセプトに比べ普及率が低く (注3)，後者は X 線撮影が不要な疾患については解析不能であることに起因する．さらに，含まれる情報の機密性の高さゆえにデータを提供可能な医療機関の数が限られることも，従来の医療情報による悉皆分析を困難なものとしている．

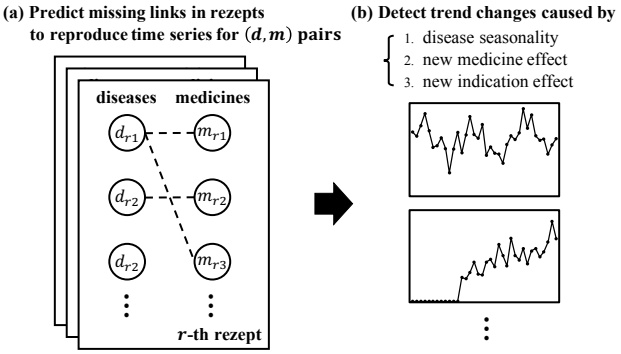


図 1　提案する 2 段階アプローチの概要．(a) 処方数を正確に復元するためにレセプト中の疾患・医薬品間の欠損リンクを推定する．(b) 処方数の時系列を季節変動や構造変化等に分解することで新薬効果や新適応効果に起因する投薬トレンドの変化を検出する

本稿では，時間の経過にともない投薬トレンドが変化する疾患，医薬品，およびその組み合わせを大規模レセプトデータから発見する問題に取り組む．図 1 に示すように，我々は 2 段階のアプローチで本問題を解決する．まず，レセプトデータ中で欠損している疾患・医薬品間の処方関係を推定するための潜在変数モデルを提案する．このモデルによって疾患と医薬品の各ペアに対する月次処方数の時系列を正確に復元する（図 1a）．次に，干渉変数を含む状態空間モデルによって処方数の時系列を長期変動，周期変動，構造変化，外れ値に分解する．この分解結果に基づき，膨大な数の疾患と医薬品の中から処方傾向に経時変化が生じているものを検出し，変化の原因を分類する（図 1b）．三重県の後期高齢者の約 3 年半にわたる診療記録か

(注1) 新有効成分含有医薬品，新医療用配合剤，新投与経路医薬品，新効能医薬品，新剤型医薬品，新用量医薬品などを指す [34].

(注2) ある疾患に有効な既存薬に対して別の疾患に有効な効能を発見すること．安全試験が簡略化でき，開発・製造コストも低減できるため，創薬の代替手段として注目を集めている [4], [36].

(注3) 平成 29 年度の電子カルテの普及率は医療機関全体の約 30%にとどまる [39].

ら構成される大規模電子レセプトデータを用いた評価実験により，提案手法の正確性，有用性，効率性を示した．

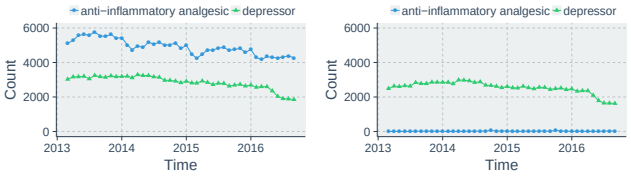
本研究の主な貢献は以下の 4 点である．

- 大規模電子レセプトデータの 1 つの活用として，学術のみならず医療や行政，経済などの分野にも大きな影響を持つ投薬トレンドの変化検知というタスクについて論じた点．我々の知る限り，当該データを用いて本タスクに取り組んだ研究は過去に存在しない．
- レセプトデータ中で欠損している疾患・医薬品間の処方関係を推定する確率的投薬モデルを提案し，提案モデルの予測性能と処方推定精度が共起頻度に基づく手法よりも優れていることを実験的に示した点．
- 干渉変数を含む状態空間モデルを用いることで，膨大な数の疾患，医薬品，およびその組み合わせの中から，季節性や新薬効果，新適応効果などに起因して投薬トレンドが変化しているものを発見可能であることを実証した点．
- 投薬トレンド変化点を効率的に発見する近似解法を提案し，その費用対効果を厳密解法との比較を通じて検証した点．

2. 関 連 研 究

医療情報解析． 医療の質や水準の向上を目的として，医療情報解析に関する研究が数多く行われてきた [5], [15], [28], [31], [41]．近年では，医療分野のユーザ理解や知識発見のための分析対象として，電子カルテ [31], [41] や X 線画像 [5], [15], [28] といった機密性の高い医療情報だけでなく，その他のデータ資源も利用され始めている．Paparrizos ら [24] は，検索ログを利用して，重大な疾患に関する自己診断を行うユーザの予測に取り組んでいる．Mishra ら [22] は，救命救急センターの検索のような，健康に関する緊急性の高い情報要求を持つユーザの発見に同様のデータを利用している．Aramaki ら [3] は，インフルエンザの流行検知のために，Twitter をソーシャルセンサとして使うことを提案している．本研究では，医療情報解析のための新たなデータ資源として，悉皆性を有するレセプトデータに着目する．Matsubara ら [21] は，大規模疫病データの非線形モデル解析手法として FUNNEL を提案している．FUNNEL は，既存の疫病感染モデル [2] を一般化した柔軟なモデルであり，疫病，場所，期間からなる所与の 3 階テンソルデータに対して，季節性，ワクチンによる減少効果，地域性，突発的な流行などといった疫病の重要な特徴を自動で発見することができる．本稿で提案する投薬トレンドの変化検知モデルは，Matsubara ら [21] が考慮した疾患の季節性に加えて，本研究の問題設定に特有の要因である，医薬品の発売や適応拡大などによる処方数の変動を捉える．

リンク予測． 2 つのノード間にリンクが存在するかという問題は，人間関係の推定や商品の購買予測といった応用範囲の広さから，精力的に研究されている [10], [20], [23]．Blei ら [8] は，文書に付与されるタグのような補助情報を含むデータのモデリング手法として，対応トピックモデルを提案している．対応トピックモデルでは，ある文書に付与されたタグのトピックは同一文書中の単語に割り当てられたトピックの中から選択される．Iwata ら [14] は，対応トピックモデルを発展させ，文書の内容とは関係のないタグが付与される状況にも対処可能な，ノイズ有り対応トピックモデルを提案している．本研究で利用するレセプトでは，診断された各疾患に対してどの医薬品を処方されたかという情報が欠落している．この処方関係を考慮せずに投薬数の正確な時系列を復元することは困難である（3.1 節）．そこで我々は，対応トピックモデルにおける単語・タグ間の関係性と，本問題における疾患・医薬品間の関係性との共通点に着目し，当該モデルと類似した構造を有する確率的投薬モデルを提案することで，疾患・医薬品間の処方関係を推定する．



(a) 共起頻度（ベースライン）　(b) 確率的投薬モデル（提案手法）

図 2　処方関係の欠損が月次投薬数の推定結果に与える影響．共起頻度のみに基づく推定では疾患に関係のない医薬品の処方数まで高くなる．確率的投薬モデル（4 節）を用いると誤差が低減される

時系列解析． 自己回帰モデル（AR）や自己回帰和分移動平均モデル（ARIMA），状態空間モデルは，代表的な時系列解析手法として知られており，これらの技術を礎として多くの研究が行われてきた [6], [7], [16]．特に，本研究でも投薬トレンドのモデル化に用いる状態空間モデルは，AR や ARIMA を包含した技術であり，カルマンフィルタ [18] を用いることで効率的なパラメータ推定や将来予測が可能となる．また，時系列の変動要因を複数の要素に分解して表現するため，対象ドメインに関する知識や仮定をモデルに柔軟に組み込むことができ，得られた結果の解釈も容易であるという特徴を持つ [11]．こうした利点を考慮して，本研究では状態空間モデルを採用し，投薬トレンド変化を干渉変数としてモデルに組み込む．変化検知やバースト検知はデータマイニング分野で長年にわたり研究されてきた [12], [19], [32], [33]．本研究では，ハイパーパラメータ不要の完全自動的な方法で投薬トレンドの変化点を発見するために，学習した状態空間モデルの相対的な質の高さを利用する．

3. 研 究 課 題

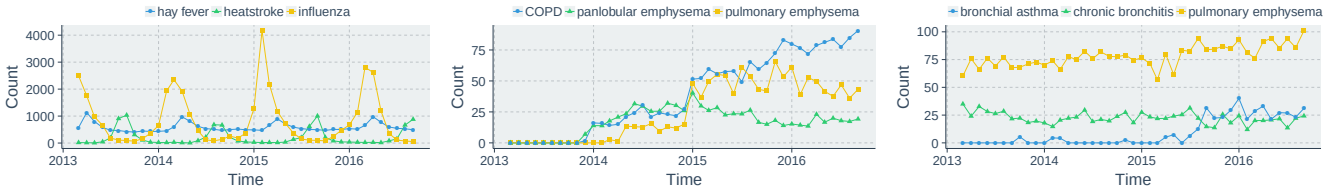
本研究で解決すべき課題として，電子レセプトデータを使う上での課題と投薬トレンドに関する課題が存在する．それぞれについて本節で整理する．

3.1 レ セ プ ト

本研究で利用する医科用レセプト[注4] には，請求者である医療機関，被保険者である患者，診断された疾患名，提供された診療行為，使用された医薬品，診療報酬等に関する情報が含まれている．各医療機関は，当該機関を受診した患者のそれぞれについて，これらの情報を月毎にまとめて保険者に送付する．保険者は，受領したレセプトの内容を審査した上で，請求された医療費を医療機関に支払う．紙面の制約上，レセプトの仕様の仔細 [35], [38] については説明を省略する．

処方関係の欠損． 電子レセプトデータを調べることで，疾患の診断回数や医薬品の使用回数については容易に把握することができる．一方で，疾患と医薬品の各ペアに対して処方回数を直接的に求めることはできない．これは，診断された各疾患に対してどの医薬品が使用されたかという処方関係に相当する情報がレセプトデータ中から欠損しているためである．前述したように，レセプトは月単位での受診記録から構成されている．そのため，異なる疾患の治療目的で 1 か月に何度か同じ病院に通った場合など，複数回の受診記録が同一のレセプトにまとめられることも多い．実際，本研究で利用するデータセット（6.1 節）では，各レセプトに含まれる疾患および医薬品の平均数はそれぞれ，7.435 および 4.788 である．信頼性の高い投薬トレンドの変化検知のためには，レセプト中で欠損している疾患と

(注4) 病院や診療所にて患者が外来診療もしくは入院診療を受けた際に発生する．その他に DPC 用，歯科用，調剤用レセプトが存在する [35], [38]．



(a) 季節性と外れ値　(b) 医薬品に起因する構造変化（新薬効果）　(c) 疾患・医薬品ペアに起因する構造変化（新適応効果）

図 3　本研究で考慮する投薬トレンドの変動要因

医薬品の処方関係を推定し，各ペアの処方回数を可能な限り正確に復元する必要がある．

この問題に対する単純な解決策として，レセプトデータ中における疾患と医薬品の共起回数を当該ペアの処方回数とみなすという方法が考えられる．図 2a は，この方法によって推定された，高血圧症に対する 2 種類の医薬品（経皮鎮痛消炎剤および血圧降下剤）の月次処方回数の時系列を表している．このうち，高血圧症に対する効能を有する医薬品は後者のみである．それにもかかわらず，図 2a 中の処方回数は両者ともに高くなっている．このように，共起頻度のみを用いた場合，レセプト中で出現頻度の高い医薬品の処方回数が不当に高く推定されてしまう．

3.2　投薬トレンド

疾患に対する医薬品の処方数はさまざまな要因に影響を受けて時間変化する．実データの観察を通じて，処方数のモデル化に必要となる要素を整理する．なお，以下で実例として示す処方数の時系列は，電子レセプトデータに関する前述の課題を解決するモデル（4 節）を学習することで推定した．

疾患の季節性．疾患の中には特定の季節に流行しやすいものが存在する [21]．図 3a は，花粉症，熱中症，およびインフルエンザに関する医薬品の処方件数を示している．同図より，花粉症は春頃，熱中症は夏頃，インフルエンザは冬頃にピークが存在することが分かる．このように，疾患の季節性は投薬数に周期的な変化をもたらすため，長期的な投薬トレンドの解析においてはその影響を区別して扱う必要がある．

医薬品に起因する構造変化．季節性は疾患の発生件数に影響する要素であるのに対して，医薬品に固有の時間変動要因も存在する．その最たるものが，新製品の発売にともなう処方数の変化である．図 3b は，2013 年 11 月に発売が開始された気管支拡張剤の処方件数を示している．同医薬品が効能を持つ複数の疾患に対して，処方数が発売時期を起点に急激に増加していることが同図より読み取れる．

疾患と医薬品の組み合わせに起因する構造変化．特定の疾患に対する治療薬として処方されていた医薬品が，別の疾患に対しても有効であることが発見され，処方対象の疾患の数が増加することもある．これは適応拡大と呼ばれる．その実例として，（前例とは別の）気管支拡張剤に関する処方数の時系列を図 3c に示す．同医薬品は，以前から慢性閉塞性肺疾患（慢性気管支炎や肺気腫の総称）に対して処方されていたが，新適応の取得が発表された 2014 年末頃を境として気管支喘息に対する処方数が大きく増加していることが同図より確認できる．

外れ値．処方数の時系列には，一時的な流行などにともなうランダムな変動も存在し得る．例えば図 3a では，2015 年初頭におけるインフルエンザに関する処方数が他の年の同時期における処方数と比べて非常に多いことが観測できる．実データにはこのような外れ値が存在するため，処方数の時系列解析モデルはノイズに対して頑健である必要がある．

4.　疾患と医薬品の処方関係の推定

前述したように，本研究で扱う電子レセプトには個人の治療記録を 1 か月単位で集約および要約した情報のみが含まれており，診断された各々の疾患に対して実際にどの医薬品が処方されたかという情報は欠損している．そこで本稿では，処方対象の疾患を潜在変数として表現した確率的投薬モデルを提案する．提案モデルは，医師の投薬行動を模倣することで，レセプト中で欠損している疾患・医薬品間の処方関係を推定する．

表 1　記号の表記

記号	説明
R	レセプトの総数
D	疾患の種類数
M	医薬品の種類数
$\boldsymbol{d}_r$	レセプト r に含まれる疾患の（多重）集合（$\{d_{rn}\}_{n=1}^{N_r}$）
$\boldsymbol{m}_r$	レセプト r に含まれる医薬品の（多重）集合（$\{m_{rl}\}_{l=1}^{L_r}$）
$\boldsymbol{\eta}$	疾患を生成する D 次元多項分布のパラメータ
z_{rl}	レセプト r 中で l 番目に投薬の対象となった疾患
$\boldsymbol{\theta}_r$	z_{rl} を生成する D 次元多項分布のパラメータ
$\boldsymbol{\phi}_d$	疾患 d に対する医薬品を生成する M 次元多項分布のパラメータ
q_{rld}	レセプト r 中で l 番目に投薬対象となる疾患が d である確率

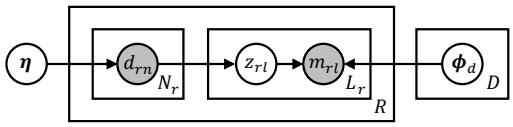


図 4　医師の投薬行動を模倣する確率的投薬モデルにおける疾患と医薬品の生成過程

本稿で用いる記号の表記を表 1 にまとめる．以降では，全レセプトの数を R と表記し，その中で出現するユニークな疾患の数および医薬品の数をそれぞれ D および M とする．また，$r \in \{1, \ldots, R\}$ 番目のレセプトを $(\boldsymbol{d}_r, \boldsymbol{m}_r)$ と表現する．ここで，$\boldsymbol{d}_r = \{d_{rn}\}_{n=1}^{N_r}$ は r 番目のレセプトにおいて診断された疾患名の（多重）集合を，$\boldsymbol{m}_r = \{m_{rl}\}_{l=1}^{L_r}$ は当該レセプトにおいて処方された医薬品の（多重）集合を表す．

4.1　生 成 過 程

提案モデルにおける疾患と医薬品の生成過程を図 4 に示す．各レセプト r について，提案モデルはまず疾患を生成する．これは，医師が患者に対する診察を通して疾患名に関する診断を下す行為に相当する．診断候補となる疾患は，病院や患者，あるいは時期によって異なることもあり得るが，本研究では簡単のため，全レセプトに共通の疾患分布が存在し，各疾患がその分布に従って選択されると仮定する．具体的には，$\eta_d \geqq 0$ かつ $\sum_{d=1}^{D} \eta_d = 1$ を満たす D 次元パラメータ $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_D)$ で構成される多項分布から各疾患 d_{rn} を生成する $\left(d_{rn} \sim \mathrm{Multinomial}(\boldsymbol{\eta})\right)$．

次に，生成された疾患集合 $\boldsymbol{d}_r = \{d_{rn}\}_{n=1}^{N_r}$ の中から，潜在変数 z_{rl} の実現値を（重複を許して）反復的に選択する．この過程は，診断を下した疾患の中から投薬治療が必要なものを判断するという医師の行為を模倣している．具体的には，先述の過程と同様に，D 次元のパラメータ $\boldsymbol{\theta}_r = (\theta_{r1}, \ldots, \theta_{rD})$ の下で $z_{rl} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_r)$ によって生成を行う．ここで，$\theta_{rd} \geqq 0$ かつ $\sum_{d=1}^{D} \theta_{rd} = 1$ である．ただし，関係のない疾患と医薬品が対応付けられることを防ぐため，レセプト r 中に出現してい

ない疾患 $d \notin \boldsymbol{d}_r$ に対しては $\theta_{rd} = 0$ という条件を設ける．

最後に，各潜在変数 z_{rl} ($l \in \{1, \ldots, L_r\}$) に対して医薬品 m_{rl} を生成する．これは，投薬が必要な疾患に対して適切な医薬品を選択するという医療行為に相当する．具体的には，$z_{rl} = d$ である時に，M 次元のパラメータ $\boldsymbol{\phi}_d = (\phi_{d1}, \ldots, \phi_{dM})$ の下で $m_{rl} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_d)$ として生成する．ここで，$\phi_{dm} \geqq 0$ かつ $\sum_{m=1}^{M} \phi_{dm} = 1$ である．

4.2 定 式 化

上述した投薬行動の生成過程を定式化する．パラメータ $\boldsymbol{\eta}$，$\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$，および $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_D)$ が与えられた際の，レセプト集合 $\mathcal{R} = \{(\boldsymbol{d}_r, \boldsymbol{m}_r)\}_{r=1}^{R}$ の生起確率 $P(\mathcal{R} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ は次式で与えられる．

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

以降では，所与のレセプトデータに対する提案モデルのパラメータの推定方法と，学習した提案モデルを用いた処方数の時系列の復元方法について述べる．

4.3 推　　定

まず，パラメータ $\boldsymbol{\Theta}$ の推定方法について述べる．本研究では，投薬の必要な疾患は各レセプト中でその疾患が出現した数に比例した確率で選択されると仮定する．これは，単語とタグのような複数種類の情報を含む文書集合に対するトピックモデルの中で用いられる仮定と同様のものである [8], [14]．この仮定に基づき，r 番目のレセプトにおいて投薬が必要と判断される疾患が d である確率 θ_{rd} を次式で定める．

$$
\theta_{rd} = \frac{N_{rd}}{N_r} \,. \qquad (2)
$$

ここで，N_{rd} は r 番目のレセプトにおける疾患 d の出現回数であり，$\sum_{d=1}^{D} N_{rd} = N_r$ となる．このようにして定義される θ_{rd} は，4.1 節で述べた「疾患 d がレセプト r 中に出現していない場合は $\theta_{rd} = 0$」という条件を明らかに満たす．

次に，残りのパラメータである $\boldsymbol{\eta}$ および $\boldsymbol{\Phi}$ の推定方法について述べる．式 (1) をパラメータの尤度関数と見ると，対数尤度 $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\Phi}) = \log P(\mathcal{R} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ は次式で与えられる．

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

ラグランジュの未定乗数法を用いると，$\mathcal{L}(\boldsymbol{\eta})$ を $\sum_d \eta_d = 1$ という制約の下で最大化するパラメータ $\boldsymbol{\eta}$ として以下が得られる．

$$
\eta_d = \frac{\sum_{r=1}^{R} N_{rd}}{\sum_{d'=1}^{D} \sum_{r=1}^{R} N_{rd'}} \,. \qquad (4)
$$

パラメータ $\boldsymbol{\Phi}$ については，最尤推定値が解析的に得られない．そこで，EM アルゴリズムを用いることで，対数尤度の下界の更新とその下でのパラメータの最大化を反復的に行う．イェンセンの不等式より，当該パラメータに関する対数尤度 $\mathcal{L}(\boldsymbol{\Phi})$ について以下が成り立つ．

$$
\mathcal{L}(\boldsymbol{\Phi}) \geqq \sum_{r=1}^{R} \sum_{l=1}^{L_r} \sum_{d=1}^{D} q_{rld} \log \frac{\theta_{rd} \phi_{dm_{rl}}}{q_{rld}} \equiv \mathcal{L}_{\mathrm{LB}}(\boldsymbol{Q}, \boldsymbol{\Phi}) \,. \qquad (5)
$$

ここで，$q_{rld} \in \boldsymbol{Q}$ はいわゆる負担率であり，r 番目のレセプトにおいて l 番目に投薬が必要と判断される疾患が d である確率を表す．負担率は $q_{rld} \geqq 0$ かつ $\sum_{d=1}^{D} q_{rld} = 1$ を満たす．EM アルゴリズムは，以下で述べる E ステップと M ステップを繰り返す．E ステップでは，パラメータ $\boldsymbol{\Phi}$ を固定した上で，対数尤度の下界 $\mathcal{L}_{\mathrm{LB}}(\boldsymbol{Q}, \boldsymbol{\Phi})$ を負担率 $\boldsymbol{Q}$ に関して $\sum_d q_{rld} = 1$ という制約の下で最大化する．M ステップでは，負担率 q_{rld} を固定した上で，$\mathcal{L}_{\mathrm{LB}}(\boldsymbol{Q}, \boldsymbol{\Phi})$ をパラメータ $\boldsymbol{\Phi}$ に関して $\sum_m \phi_{dm} = 1$ という制約の下で最大化する．ラグランジュの未定乗数法を用いると，各ステップにおける推定量は次式で与えられる．

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

$$
\phi_{dm} = \frac{\sum_{r=1}^{R} \sum_{l=1}^{L_r} q_{rld} \mathbb{1}(m_{rl} = m)}{\sum_{m'=1}^{M} \sum_{r=1}^{R} \sum_{l=1}^{L_r} q_{rld} \mathbb{1}(m_{rl} = m')} \,. \qquad (7)
$$

ここで，$\mathbb{1}(\cdot)$ は指示関数である．

4.4 処方時系列の復元

データセットに T か月分のレセプトが含まれるとする．疾患と医薬品の各ペアに対する月次処方数の時系列 $\mathcal{X}_{\mathrm{P}} \in \mathbb{R}^{D \times M \times T}$ を復元するために，各月のレセプト集合に対して上記で提案した確率的提案モデルを学習する．得られた学習モデルを用いて，疾患 d に対して医薬品 m が時刻 $t \in \{1, \ldots, T\}$ に処方された回数 $x_{dmt} \in \mathcal{X}_{\mathrm{P}}$ を次式で推定する．

$$
x_{dmt} = \sum_{r=1}^{R^{(t)}} \sum_{l=1}^{L_r^{(t)}} q_{rld}^{(t)} \mathbb{1}(m_{rl}^{(t)} = m) \,. \qquad (8)
$$

ここで，各変数に付与されている上付き文字 (t) は，時刻 t におけるレセプト集合に対して上記の計算を行うことを意味する．

得られた $\mathcal{X}_{\mathrm{P}}$ を用いることで，各疾患に対する月次処方数の時系列 $\mathcal{X}_{\mathrm{D}} \in \mathbb{R}^{D \times T}$ と，各医薬品の月次処方数の時系列 $\mathcal{X}_{\mathrm{M}} \in \mathbb{R}^{M \times T}$ も復元が可能となる．時刻 t において，疾患 d に医薬品が処方された回数を x_{dt}，医薬品 m が疾患に処方された回数を x_{mt} とおくと，これらの値は次式で推定される．

$$
x_{dt} = \sum_{m=1}^{M} x_{dmt} \,, \quad x_{mt} = \sum_{d=1}^{D} x_{dmt} \,. \qquad (9)
$$

3.1 節で例示した疾患と医薬品のペアに対して，提案手法を用いて推定した処方数の時系列を図 2b に示す．提案手法による推定結果では，疾患に対する効能のない医薬品については処方数がゼロに近い一方で，効能のある医薬品については共起頻度に基づく処方数（図 2a）と同様の傾向が維持されていることが確認できる．

5. 投薬トレンドの変化検知

4 節の確率的投薬モデルによって復元された処方数の時系列から処方数のトレンドの経時変化を検出するために，3.2 節で述べた要素を組み込んだ状態空間モデルを提案する．

5.1 定 式 化

処方時系列 $\{x_{qt}\}_{t=1}^{T}$（変数 q は，疾患 d に対する処方数であれば d，医薬品 m の処方数であれば m，両者のペアに関する処方数であれば (d, m) を表す）の振る舞いを以下の状態空間モデルで表現する．

$$
\begin{aligned}
x_{qt} &= \mu_{qt} + \gamma_{qt1} + \lambda_q w_{qt} + \epsilon_{qt} \,, \\
\mu_{q,t+1} &= \mu_{qt} + \xi_{qt} \,, \\
\gamma_{q,t+1,s} &= \begin{cases} -\sum_{s'=1}^{11} \gamma_{qts'} + \omega_{qt} & (s = 1) \,, \\ \gamma_{qt,s-1} & (s \in \{2, \ldots, 11\}) \,, \end{cases} \qquad (10) \\
\epsilon_{qt} &\sim N(0, \sigma_\epsilon^2), \, \xi_{qt} \sim N(0, \sigma_\xi^2), \, \omega_{qt} \sim N(0, \sigma_\omega^2) \,,
\end{aligned}
$$

解法 1　時系列 $\{x_{qt}\}_{t=1}^{T}$ の厳密変化点の発見

best_point ← NULL, best_aic ← ∞
for each change point $t \in \{1, \ldots, T, -1\}$ **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

ここで，$N(\mu, \sigma^2)$ は平均が μ で分散が σ^2 の正規分布である．処方回数 $x_{\cdot}$ は，上式 1 段目の観測方程式によって，レベル項 $\mu_{\cdot}$，季節項 $\gamma_{\cdot}$，干渉項 $\lambda_{\cdot} w_{\cdot}$，誤差項 $\epsilon_{\cdot}$ に分解される．各要素について以下で説明する．

レベル項．他の要素では説明されない，時系列の緩やかな長期変動を表現する．本項は古典的な線形回帰モデルにおける切片と類似した役割を持つが，その値は時刻とともに変化し得る．

季節項．季節性にともなう時系列の周期的な変動を捉える．多くの季節性疾患に対する処方数には 12 か月を周期とする変動が存在するため（図 3a），本研究では 11 本の状態方程式を用いて季節項の変動をモデル化する [注5]．

干渉項．ある時点に発生したイベントによる時系列の構造的な変化を表現する．本項は $\lambda_{\cdot}$ と $w_{\cdot}$ の 2 つの変数から構成される．前者は変化の有無を表す疑似変数であり，後者は変化の規模を表す数値である．3.2 節で述べたように，本研究の主な関心は，新薬効果や新適応効果等に起因する投薬トレンドの変化であり，これらの効果は処方数の時系列の傾きに変化をもたらす（図 3b および図 3c）．そこで，本項で捉える変化の種類として傾きシフト [11] を採用する．すなわち，時系列 $\{x_{qt}\}_{t=1}^{T}$ が時刻 t_{CP} において変化点を持つとすると，時刻 t における疑似変数 w_{qt} の値を，$t \geqq t_{\mathrm{CP}}$ であれば $t - t_{\mathrm{CP}} + 1$ とし，それ以外であれば 0 とする．また，簡単のため，構造的な変化の規模 λ_q は時刻によらず一定であると仮定する．

誤差項．式 (10) の 1 段目の観測方程式と 2 段目以降の状態方程式はともに要素の誤差変動を最終項に含む．後者の誤差項は各要素の値が時刻とともに変化することを許容するため，モデルの柔軟性が向上する．前者の誤差項は時系列成分のうち各要素に分解不可能なものを吸収する役割を持つ．3.2 節で述べたように，処方数の時系列には外れ値が存在し得る．誤差項の存在によって，ノイズに対して頑健な解析が可能となる．

5.2　推　　定

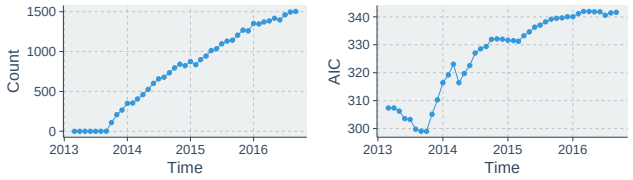
処方数の時系列 $\{x_{qt}\}_{t=1}^{T}$ とその変化点 t_{CP} が与えられれば，カルマンフィルタ [18] を用いて，状態空間モデルのパラメータを効率的に推定できる．そのため，残された課題は，所与の時系列の変化点を発見することとなる．時系列の変化点は，ハイパーパラメータを必要としない完全自動的な方法で発見することが望ましい．これは，膨大な数の時系列（6.1 節で述べるように，我々の実験では 20 万件以上の疾患・医薬品ペアを扱う）に対して，変化点を人手で発見することやハイパーパラメータを調節することが現実的ではないという理由による．

厳密解法．本研究では，変化点の自動発見のための指標として，赤池情報量規準（AIC） [1] を採用する．AIC は，統計モデルの質の高さを，データに対する適合度とモデル自体の複雑度から評価する指標であり，その値が低いほど質の高いモデルであることを示す．我々は，解法 1 を用いて，所与の時系列の厳密な変化点を発見する．本解法は，時系列の各時点を変化点とみなしたモデルを学習し，その AIC 値が最小となる候補点を発見する．次に，候補点に対するモデルと変化点が存在しないモデル（式 (10) から干渉項を除去したもの）とで AIC 値を比較し，変化点の有無を決定する．

[注5] 誤差項を除いた季節変動は各周期においてその総和がゼロとなる．

解法 2　時系列 $\{x_{qt}\}_{t=1}^{T}$ の近似変化点の発見

function FindWithin(left, right)
　if right − left ≦ 1 **then**
　　return $\arg\min_{t \in \{\mathrm{left}, \mathrm{right}\}} \mathrm{AIC}(\{x_{qt}\}_{t=1}^{T}, t)$
　middle ← $\frac{\mathrm{left}+\mathrm{right}}{2}$
　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 $\arg\min_{t \in \{\mathrm{best}, -1\}} \mathrm{AIC}(\{x_{qt}\}_{t=1}^{T}, t)$



(a) 2013 年 9 月に変化点を持つ時系列　(b) 各月を変化点とみなした際の AIC 値

図 5　真の値に近い点を変化点とみなすモデルほど AIC 値が低くなる．解法 2 はこの観測に基づき探索範囲を効率的に絞り込む

近似解法．解法 1 は時系列の全数探索を行うことで変化点を発見するため，その計算時間はレセプトデータの期間の長さに比例して増加する．本研究では，無駄な探索を回避し計算時間を短縮するために，異なる干渉変数の値に対する AIC 値の感度に着目する．図 5 は，2013 年 9 月に変化点を持つ時系列に対して，各月を変化点をみなして学習したモデルの AIC 値の変化を示している．同図から，真の変化点に近い値で学習したモデルほど AIC 値が低くなっている様子が観察される．そこで我々は，解法 2 を用いることで，所与の時系列の変化点の近似値を効率的に発見する．本解法の振る舞いは二分探索と類似しており，反復のたびに探索範囲が半減される．

時間計算量．カルマンフィルタを用いたモデルの学習に要する時間計算量を $\mathcal{C}_{\mathrm{KF}}$ とする．解法 1 は変化点の厳密解の発見に $\mathcal{O}(\mathcal{C}_{\mathrm{KF}} T)$ を要する．一方，解法 2 は変化点の近似解の発見に $\mathcal{O}(\mathcal{C}_{\mathrm{KF}} \log(T))$ を要する．両者の費用対効果を 6 節で評価する．

6.　評 価 実 験

4 節および 5 節で提案したモデルを評価するために，レセプトの実データを用いた実験を行った．本節では，評価実験を通じて以下の項目を検証する．

- 確率的投薬モデルに基づく処方関係推定の正確性
- 状態空間モデルに基づく投薬トレンドの変化検知の有用性
- 近似解法に基づく時系列の変化点発見の効率性

以降では，有意水準 $\alpha = 0.05$ の下で統計的有意性を報告する．

6.1　データセット

本実験では大規模電子レセプトデータとして，三重県の後期高齢者の診療記録に関する医科用レセプトを利用した．当該レセプトの請求期間は 2013 年 3 月から 2016 年 9 月までの約 3 年半にわたる（$T = 43$）．各月のデータ $\mathcal{R}^{(t)}$ ($t \in \{1, \ldots, T\}$) には平均して，3,347 か所の医療機関，202,972 人の患者，332,167 件のレセプト，9,173 種類の疾患，9,346 種類の医薬品が含まれ

表 2　医薬品の予測性能（パープレキシティ）と処方関係の適合性（AP@10 と NDCG@10）の平均（と標準偏差）．全指標において提案モデルがベースラインに比べて有意に良い結果を示した

	パープレキシティ	AP@10	NDCG@10
共起頻度	168.241 (7.408)	0.304 (0.243)	0.450 (0.260)
確率的投薬モデル	**112.436** (4.480)	**0.787** (0.298)	**0.835** (0.288)

ていた．

確率的投薬モデルの学習時には，トピックモデルに関する既存研究 [14],[30] と同様に，各月のレセプトデータ $\mathcal{R}^{(t)}$ から出現回数が 5 回未満の疾患および医薬品を除外した．状態空間モデルの学習時には，不安定な学習を避けるために，全期間における総処方回数が 10 回未満の時系列を除外した．これらの除外操作の結果，トレンド変化検知対象の疾患，医薬品，および両者のペアはそれぞれ，3,978 件，7,474 件，および 206,829 件となった．

6.2 正　確　性

提案モデルの正確性を評価するために 2 種類の実験を実施した．両者の実験では以下の手法をベースラインとして利用した．

ベースライン． 3.1 節で述べた，共起頻度に基づき疾患・医薬品間の処方数を推定する手法をベースラインに採用した．本手法は，各疾患 d に対する医薬品の生成分布のパラメータ $\Phi_d = (\phi_{d1}, \ldots, \phi_{dM})$ を，式 (7) に代わって次式で推定する．

$$\phi_{dm} = \frac{\sum_{r=1}^{R} \sum_{l=1}^{L_r} \mathrm{Cooc}_r(d, m)}{\sum_{m'=1}^{M} \sum_{r=1}^{R} \sum_{l=1}^{L_r} \mathrm{Cooc}_r(d, m')} . \tag{11}$$

ここで，$\mathrm{Cooc}_r(d, m)$ はレセプト r における疾患 d と医薬品 m の共起回数を表す．

6.2.1 医薬品の予測性能

まず，確率的投薬モデルと上述のベースラインモデルの予測性能を評価した．

設定． 各レセプトに含まれる医薬品のうち，90%を用いて各モデルを訓練し，残りの 10%に対する予測性能をパープレキシティで評価した．パープレキシティは，統計モデルの予測性能の評価のために広く利用されている評価指標であり [8],[14]，その値が低いほどモデルの予測性能が高いことを示す．各レセプト r に含まれる評価用の医薬品の（多重）集合を $\boldsymbol{m}'_r = \{m'_{rl}\}_{l=1}^{L'_r}$ とすると，モデル $\mathcal{M}$ のパープレキシティは次式で計算される．

$$\mathrm{PPL}(\{\boldsymbol{m}'_r\}_{r=1}^{R} \mid \mathcal{M}) = \exp\left(-\frac{\sum_{r=1}^{R} \sum_{l=1}^{L'_r} \log P(m'_{rl} \mid \mathcal{M})}{\sum_{r=1}^{R} L'_r} \right) . \tag{12}$$

結果． 各月のレセプトデータ $\mathcal{R}^{(t)}$ $(t \in \{1, \ldots, T\})$ に対して，提案モデル $\mathcal{M}^{(t)}_{\mathrm{prop}}$ とベースラインモデル $\mathcal{M}^{(t)}_{\mathrm{cooc}}$ を学習し，両者のパープレキシティを計測した．その平均（および標準偏差）を表 2 にまとめる．平均すると，提案モデルのパープレキシティはベースラインモデルの 3 分の 2 程度となった．個々の結果を調べたところ，全ての月のレセプトデータ $\mathcal{R}^{(t)}$ に対して提案モデルのパープレキシティがベースラインよりも低いことが判明した．対応あり t 検定を行った結果，両者のパープレキシティには有意な差が存在することが確認された（$t(42) = -103.670$, $p < 0.001$, Cohen's $d = -15.810$）．参考として，医薬品のユニグラムモデル [29] のパープレキシティを計測したところ，その平均値は 2315.083 となり，提案モデルと比べて性能が 20 倍悪化することが分かった．以上の実験から，医師の投薬行動を模倣する提案モデルが比較対象の中で最も優れた予測性能を発揮することが示された．

6.2.2 処方関係の適合性

次に，各モデルによって関係が強いと推定された疾患・医薬品ペアの適合性を評価した．

表 3　各種の処方数の時系列に対するモデルの AIC 値の平均（と標準偏差）．干渉項を含む提案モデルが干渉項のないモデルに比べて有意に良い結果を示した．近似変化点（解法 2）を用いたモデルは厳密変化点（解法 1）を用いた場合と同等の性能を達成した

	疾患	医薬品	疾患・医薬品ペア
干渉項なし	254.018 (44.937)	218.295 (56.216)	103.594 (54.999)
干渉項あり + 厳密変化点	**244.603** (44.937)	**208.396** (56.441)	**91.888** (50.618)
干渉項あり + 近似変化点	**244.742** (44.915)	**209.076** (55.638)	**92.099** (50.614)

設定． レセプトデータの全期間において出現頻度が最も高い 100 件の疾患を評価対象として用意した．各モデルについて，これらの各疾患 d に対する医薬品 m の関連度に基づくランキングを，両者の推定処方回数（$x_{dm} = \sum_{t=1}^{T} x_{dmt}$）の降順によって生成した．こうした得られたランキングの上位 10 件の適合性を，情報検索分野で広く利用されている評価指標である Average Precision（AP） [25],[26] と Normalized Discounted Cumulative Gain（NDCG） [9],[17] を用いて評価した．いずれの指標も値が高いほど優れたランキングであることを示す．

正解セット． ランキングの評価には，そこに含まれる全ての疾患・医薬品ペア（合計 1,591 件）に関する適合性の正解セットが必要となる．そこで，本稿の第 1 著者が，評価対象の各医薬品 m と各疾患 d との間の適合性を「d もしくは d の上位語が m の添付文書の薬効分類名と効能・効果の少なくとも一方に記述されていれば適合（$= 1$），そうでなければ不適合（$= 0$）」という基準で判定した．この手続きによって，1,154 件については適合値が得られたが，残りの 437 件については専門知識なしでの判定が困難であった．これらの未判定事例については，医療専門家に専門知識に基づく判定を依頼した (注6)．最終的に，1,528 件について適合値が得られた．残りの 63 件については不適合としてランキングの評価を実施した．

結果． 各ランキングに対する AP@10 と NDCG@10 の平均（および標準偏差）を表 2 に示す．同図から，いずれの指標においても，提案モデルによってランキングの精度が大きく改善されていることが確認できる．全 100 件のランキングのうち，ベースラインモデルが提案モデルを上回ったのは，AP@10 では 1 件，NDCG@10 では 2 件のみであった．両者の差は対応あり t 検定によって有意であることが判明した（AP@10：$t(99) = 15.398$, $p < 0.001$, Cohen's $d = 1.540$．NDCG@10：$t(99) = 14.374$, $p < 0.001$, Cohen's $d = 1.437$）．本実験によって，図 2 で例示した提案モデルの優位性が定量的に実証された．

6.3 有　用　性

提案モデルの有用性を定量的な観点と定性的な観点からそれぞれ評価した．

6.3.1 定量的評価

まず，提案モデルの質の高さを AIC を用いて定量的に評価した．ベースラインには，提案モデルから干渉項を除外したモデルを採用した．これらのモデルを疾患，医薬品，および両者のペアに関する処方数の時系列データに対して学習した．

各モデルの AIC 値の平均（および標準偏差）を表 3 に示す．平均すると，干渉項を含む提案モデルは，ベースラインに比べて 10 程度低い AIC 値を達成した．対応あり t 検定を行った結果，全ての種類の時系列に対して両モデルの間には有意な差が存在することが判明した（疾患：$t(3,977) = -36.619$, $p < 0.001$, Cohen's $d = -0.581$．医薬品：$t(7,473) = -49.829$, $p < 0.001$, Cohen's $d = -0.576$．疾患・医薬品ペア：$t(206,830) = -412.520$, $p < 0.001$, Cohen's $d = -0.907$）．こ

(注6)　二重チェックのため，評価依頼時に判定済みの事例も医療専門家と共有した．

の結果は，処方数の時系列解析において，構造的な変化を捉えることの重要性を示唆している．実際に，提案モデルは，12%の疾患時系列，28%の医薬品時系列，10%の疾患・医薬品ペア時系列について変化点を検出した．

6.3.2 定性的評価

次に，定性的な有用性を評価するために，トレンド変化が確認された事例を分析した．6 種類の時系列に対する提案モデルの学習結果を図 6 および図 7 に示す．各事例について，原時系列と学習時系列を上段に，モデルによって分解された各要素を中段に，原時系列に関係する時系列を下段に並べている．

季節性と外れ値．図 6a はインフルエンザに対する学習結果を示している．中段のグラフから，毎年冬に流行するというインフルエンザの季節性を提案モデルが精度良く学習できていることが分かる．インフルエンザは 2015 年の初頭に例年以上の流行が観測された（上段）．この変化は一時的なものであるため，提案モデルはそれを外れ値として扱うことで安定した学習結果を実現している．季節性疾患に関する別の事例として，下痢症に対する学習結果を図 6b に示す．下痢症は季節の変わり目に起こりやすいが，こうした年 1 回以上のピークが存在する疾患についても，提案モデルは季節性をうまく捉えている．

医薬品に関する構造変化．図 6c は，2013 年秋に発売が開始された骨粗鬆症の新薬に関する学習結果を表している．提案モデルは発売時期とその後の処方数の伸びを正確に検出している（中段）．下段の時系列は，骨粗鬆症に対する効能を有する既存の医薬品の処方数の変遷を表している．同図から，新薬の発売によって，骨粗鬆症に対して処方される医薬品の傾向に大きな変化が生じたことが確認できる．図 6d は，提案モデルによって処方数の急な減少が検出された医薬品の 1 例である．この医薬品は，同じ効能を持つ複数のジェネリック医薬品が 2015 年中頃に発売されたため（下段），その時期を境に処方数が大きく落ち込んだものと予想される．

疾患・医薬品ペアに関する構造変化．提案モデルは，新適応に起因する投薬トレンドの変化検知にも成功した．図 7a は，レビー小体型認知症への適応が追加された医薬品の学習事例である．最後に，疾患と医薬品の組み合わせに起因する別の種類の構造変化の事例として，ある医薬品の経口摂取困難に対する処方数の学習結果を図 7b に示す．同図の上段のグラフから，2014 年中頃を境に処方数が突然上昇していることが分かる．下段のグラフは，同医薬品の別の疾患に対する処方数の時系列を示しており，この医薬品が新薬ではないことが分かる [注7]．同グラフでもう 1 つ注目すべきは，経口摂取困難の時系列と正反対の振る舞いをしている脱水症（“related1”）の時系列である．この結果は，同様の症状に対する診断名が時期によって異なる可能性を示唆している．

6.4 効 率 性

提案モデルの有用性を定量的な観点と定性的な観点からそれぞれ評価した．提案モデルの効率性を評価するために，変化点検出の厳密解法と近似解法の費用対効果を検証した．

6.4.1 計 算 時 間

まず，両解法の計算時間を評価した．具体的には，疾患，医薬品，両ペアの時系列集合に対して，これらの解法を用いた学習に要する合計時間を計測した．

その結果を表 4 に示す．同表には比較として，変化点検出を行わないモデルに対する学習時間の増加率も含めている．近似解法を用いることで，計算時間が大幅に減少することが確認できる．理論的には，厳密解法および近似解法はそれぞれ，$\mathcal{O}(\mathcal{C}_{\mathrm{KF}}T)$ および $\mathcal{O}(\mathcal{C}_{\mathrm{KF}}\log(T))$ の計算時間を要する（5.2 節）．ここで，カルマンフィルタの実行に要する計算時間は時系列の期間（$T = 43$）に関して定数であるとみなせる．そのため，計算時間の増加率の期待値はそれぞれ，43 および $\log_2(43) \approx 5.426$ となる．実験結果は理論値に概ね一致しているといえる [注8]．

表 4　各種の時系列の学習に要する合計時間（分単位）．括弧内の値は，干渉項のないモデルに対する学習時間の増加率を表す．変化点検出の近似解法により学習時間が大きく減少した

	疾患	医薬品	疾患・医薬品ペア
干渉項あり＋厳密変化点	8.529 (27.878)	17.565 (29.900)	562.614 (35.492)
干渉項あり＋近似変化点	1.832 (5.989)	3.678 (6.260)	117.308 (7.400)

表 5　厳密解法と近似解法とで検出された変化点の有無の内訳．近似解法は偽陽性率と偽陰性率ともに十分に低い値を達成した

(a) 疾患

		近似	
		pos.	neg.
厳密	pos.	423	40
	neg.	0	3,515

(b) 医薬品

		近似	
		pos.	neg.
厳密	pos.	1,944	154
	neg.	0	5,376

(c) 疾患・医薬品ペア

		近似	
		pos.	neg.
厳密	pos.	19,106	2,079
	neg.	0	185,644

6.4.2 近 似 精 度

次に，近似解法の精度を評価するために，厳密解法と近似解法によって検出された変化点を比較した．

両解法によって検出された変化点の有無の分割表を表 5 に示す．近似解法は，その定義から，変化点を誤検出することはなく，偽陽性率は常に 0%である．さらに，偽陰性率も低いことが同表より分かる（疾患：8.639%．医薬品：7.340%．疾患・医薬品ペア：9.814%）．疾患，医薬品，および疾患・医薬品ペアの結果に関するカッパ係数はそれぞれ，0.949，0.948，0.943 であり，両解法によって検出される変化点の一致度は高いといえる．

近似精度を評価する別の指標として，両解法が検出した変化点の平均二乗誤差（RMSE）を計測した．疾患，医薬品，および疾患・医薬品ペアの結果に関する RMSE 値はそれぞれ，3.862，7.154，および 4.481 となった．本実験で用いた時系列の期間が 43 か月であることを考慮すると，医薬品以外の時系列については，近似解法は比較的高い精度で変化点を検出したといえる．

また，表 3 の下段には，近似解法を用いて学習したモデルの AIC 値の平均（および標準偏差）が示されている．同表から，近似解法によるモデルは，厳密解法を用いた場合と同等の質の高さを達成していることが確認できる．

7. おわりに

本稿では，投薬トレンドの検知とその原因の分類という問題に取り組んだ．本問題に対してレセプトデータを用いた研究は，我々の知る限り，過去に存在しない．3.5 年分の大規模なレセプトの実データを用いて，提案した 2 段階アプローチの優位性を以下の観点から示した．

（1） **正確性：**提案した確率的投薬モデルが，共起頻度に基づく手法に比べて，医薬品の予測力と処方関係の適合性の双方の点において有意に良い性能を達成することを示した．

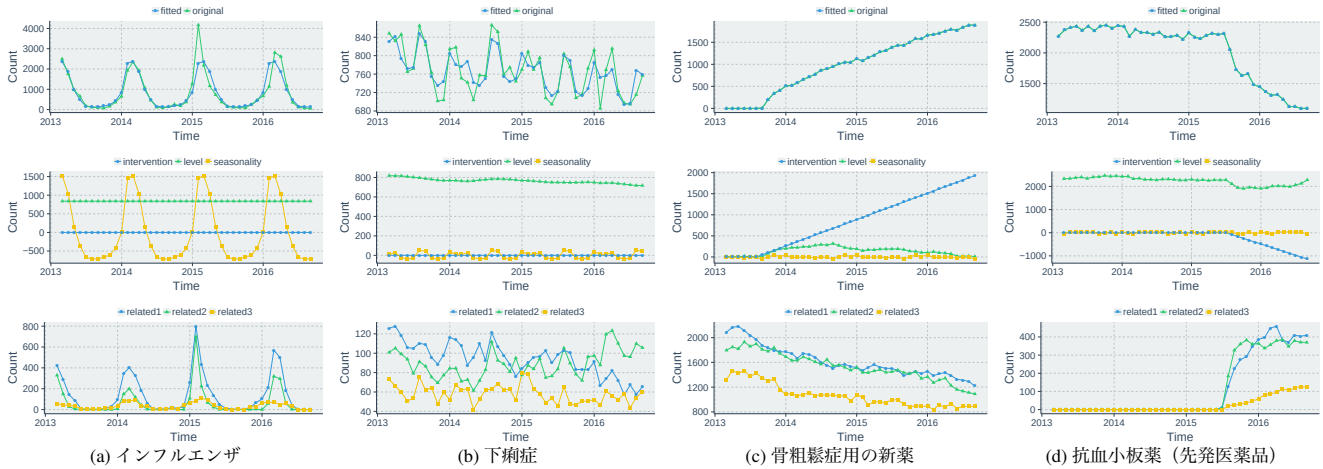
（2） **有用性：**提案した状態空間モデルが，膨大な数の時系列の中から，新薬効果や新適応効果等に起因する投薬トレンドの変化を発見可能であることを示した．

（3） **効率性：**提案した近似解法が，投薬トレンド変化点の検出精度を保ちつつ，計算時間を短縮可能であることを示した．

今後の展開として，より複雑な投薬トレンド変化を捕捉可能

[注7] この医薬品の実際の発売年月は 1986 年 1 月である．

[注8] 近似解法の増加率は理論値をやや上回っている．その一因として，同解法はまず時系列の両端を変化点をみなして計算が行うことがあげられる．



(a) インフルエンザ　(b) 下痢症　(c) 骨粗鬆症用の新薬　(d) 抗血小板薬（先発医薬品）

図 6　疾患の処方数時系列および医薬品の処方数時系列に対する提案モデルの学習結果．上段：原系列と学習時系列．中段：提案モデルによって分解された時系列の各要素．下段：原系列に関係する別の時系列

(a) レビー小体型認知症の適応追加　(b) 診断トレンドの変化の可能性

図 7　疾患・医薬品ペアの処方数時系列に対する提案モデルの学習結果

なモデルの考案や，全国規模のレセプトデータを用いたトレンド地域性の検証が考えられる．また，医療関係者向けの投薬トレンドの探索的検索用インタフェースの開発も検討している．

謝辞　本研究の一部は，厚生労働科学研究費政策科学推進研究「汎用性の高いレセプト基本データセット作成に関する研究」，厚生労働科学特別研究事業戦略研究「レセプト情報・特定健診等情報データベースを利用した医療需要の把握・整理・予測分析および超高速レセプトビッグデータ解析基盤の整備」，内閣府最先端研究開発支援プログラム（FIRST）「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」，内閣府革新的研究開発推進プログラム（ImPACT）「社会リスクを低減する超ビッグデータプラットフォーム」，日本医療研究開発機構（AMED）臨床研究等 ICT 基盤構築研究事業「エビデンスの飛躍的創出を可能とする超高速・超学際次世代 NDB データ研究基盤構築に関する研究」の助成に依る．電子レセプト情報の第三者提供に掛かる手続きおよび作業に関しては，三重県国民健康保険団体連合会に多大な尽力を頂いた．

文　献

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *ISIT*, pp. 267–281, 1973.
[2] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.
[3] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *EMNLP*, pp. 1568–1576, 2011.
[4] T. T. Ashburn and K. B. Thor. Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3:673–683, 2004.
[5] 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*, 30(3):733–746, 2011.
[6] A. J. Bagnall and G. J. Janacek. Clustering time series from arma models with clipped data. In *KDD*, pp. 49–58, 2004.
[7] 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*, 13(3):211–237, 2003.
[8] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR*, pp. 127–134, 2003.
[9] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pp. 89–96, 2005.
[10] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 05 2008.
[11] J. Commandeur and S. J. Koopman. *An Introduction to State Space Time Series Analysis*. Oxford University Press, 2007.
[12] V. Guralnik and J. Srivastava. Event detection from time series data. In *KDD*, pp. 33–42, 1999.
[13] 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*, 55(2):446–459, 2015.
[14] T. Iwata, T. Yamada, and N. Ueda. Modeling noisy annotated data with application to social annotation. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1601–1613, 2013.
[15] 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*, 33(2):233–245, 2014.
[16] A. Jain, E. Y. Chang, and Y.-F. Wang. Adaptive stream resource management using kalman filters. In *SIGMOD*, pp. 11–22, 2004.
[17] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.
[18] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82:35–45, 1960.
[19] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD*, pp. 91–101, 2002.
[20] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.
[21] Y. Matsubara, Y. Sakurai, W. G. van Panhuis, and C. Faloutsos. Funnel: Automatic mining of spatially coevolving epidemics. In *KDD*, pp. 105–114, 2014.
[22] N. Mishra, R. W. White, S. Ieong, and E. Horvitz. Time-critical search. In *SIGIR*, pp. 747–756, 2014.
[23] 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.
[24] J. Paparrizos, R. W. White, and E. Horvitz. Detecting devastating diseases in search logs. In *KDD*, pp. 559–568, 2016.
[25] S. Robertson. A new interpretation of average precision. In *SIGIR*, pp. 689–690, 2008.
[26] T. Sakai. Alternatives to bpref. In *SIGIR*, pp. 71–78, 2007.
[27] 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*, 55(12):2717–2730, 2015.
[28] 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.
[29] F. Song and W. B. Croft. A general language model for information retrieval. In *CIKM*, pp. 316–321, 1999.
[30] 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.
[31] 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.
[32] 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.
[33] Y. Zhu and D. Shasha. Efficient elastic burst detection in data streams. In *KDD*, pp. 336–345, 2003.
[34] 医薬品医療機器総合機構. 平成 28 年度承認品目一覧（新医薬品）. https://www.pmda.go.jp/review-services/drug-reviews/review-information/p-drugs/0026.html. 参照日: 2018 年 1 月 8 日.
[35] 合田和生, 山田浩之, 喜連川優, 満武巨裕. 我が国の公的医療保険の悉皆分析を可能とする高速レセプト解析システムの開発と今後の展望. 第 9 回データ工学と情報マネジメントに関するフォーラム, 2017.
[36] 齊藤聡, 山本由美, 猪原匡史. ドラッグ・リポジショニングの新展開. 日本老年医学会雑誌, 52(3):200–205, 2015.
[37] 社会保険診療報酬支払基金. レセプト請求形態別の請求状況（平成 29 年度）. http://www.ssk.or.jp/tokeijoho/tokeijoho_rezept/tokeijoho_04_h29.html. 参照日: 2018 年 1 月 9 日.
[38] 社会保険診療報酬支払基金. 電子レセプトの作成. http://www.ssk.or.jp/seikyushiharai/rezept/iryokikan/iryokikan_02.html. 参照日: 2018 年 1 月 9 日.
[39] 総務省. 平成 29 年版 情報通信白書 | 医療・介護・健康分野における ict 利活用の推進. http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h29/html/nc275120.html. 参照日: 2018 年 1 月 9 日.
[40] 満武巨裕. 日本のレセプト情報・特定健診等データベース（ndb）の有効活用. 情報処理, 56(2):140–144, jan 2015.
[41] 森田祐司, 吉川正俊, 濱崎暁洋, 杉山治, 岡本和也, 黒田知宏. 投薬歴の構築と医師の多様性を考慮した投薬パターンマイニング. 第 9 回データ工学と情報マネジメントに関するフォーラム, 2017.