実験概要

● 対象データ

患者：15,351人
- 1年以上にわたり観察されている
- 対象疾病を2つ以上診断されている

疾病：1,000種
- 対象患者内で特に出現頻度の高い、ICD-10分類

⇨ $\boldsymbol{X} \in \{0,1\}^{15,351\times1,000}$
患者-疾病行列

患者属性：5種
- 「男性」「女性」「高齢」「肥満」「痩せ」の5種

⇨ $\boldsymbol{Y} \in \{0,1\}^{15,351\times5}$
患者-患者属性行列

● 予測の評価方法

1. 患者が最後に診断された疾病を0で隠す（**「マスク疾病」**）
2. 患者ごとに発症しやすい疾病を**ランキング**で予測
3. 「マスク疾病」の予測順位を高く提示できるか、評価する

実験概要

- 評価指標
  - **Top-K Accuracy**（Top-K Acc）
    - 「マスク疾病」の予測ランキングがトップK(=10,30,50)位に入る割合
  - **Mean Reciprocal Rank**（MRR）
    - 「マスク疾病」の予測ランキングの逆数の平均

  } 1に近いほど「高精度」

- 比較手法
  - **同性別・年代によるランキング集計** (ルールベース)
  - **RandomForest**
  - **Non-Negative Matrix Factorization** (NMF)
  - **Positive Collective Matrix Factorization** (PCMF)

  } Yの有無も比較

予測精度の評価①

■ 患者全体における予測精度

手法	Top-10 Acc	Top-30 Acc	Top-50 Acc	MRR
ルールベース	0.200	0.349	0.439	0.086
RandomForest	0.241	0.415	0.505	0.108
NMF ($\boldsymbol{Y}$ なし)	0.239	0.418	0.504	0.111
NMF ($\boldsymbol{Y}$ あり)	0.231	0.409	0.498	0.107
PCMF ($\boldsymbol{Y}$ なし)	0.250	0.413	0.503	0.122
PCMF ($\boldsymbol{Y}$ あり)	0.243	0.429	0.518	0.114

※ 検証患者(15%)で最高精度であった学習済みモデルでテスト患者(15%)を評価

提案手法PCMF > 他のルールベース/モデル

予測精度の評価②

■ 疾病数が少ない(2種以下) 患者における予測精度

手法	Top-10 Acc	Top-30 Acc	Top-50 Acc	MRR
ルールベース	0.173	0.336	0.407	0.074
RandomForest	0.217	0.385	0.469	0.097
NMF ($\boldsymbol{Y}$ なし)	0.181	0.336	0.447	0.099
NMF ($\boldsymbol{Y}$ あり)	0.190	0.367	0.465	0.092
PCMF ($\boldsymbol{Y}$ なし)	0.243	0.403	0.509	0.138
PCMF ($\boldsymbol{Y}$ あり)	0.261	0.465	0.544	0.143

※ 検証患者(15%)で最高精度であった学習済みモデルでテスト患者(15%)を評価

患者-患者属性行列ありPCMF > 他のルールベース/モデル

PCMFによる特徴表現解析

- 本研究にて行った解析（時間の都合上、赤字のみ紹介）
  - ■ 因子の意味解析
  - ■ 患者属性の特徴表現解析
  - ■ 疾病の特徴表現解析
    - 疾病同士の類似性解析
    - 2次元マップへの埋め込み/ICD-10との比較
  - ■ 疾病×患者属性の特徴表現解析
    - 疾病の患者属性の類似性解析
  - ■ 患者の特徴表現解析

因子の意味解析

■ 特徴表現の12の要素において、特に値の大きい疾病を抽出

3番目の要素（因子）

ICD-10	疾病名	値
H52.2	乱視	14.465
H35.3	黄斑及び後極の変性	10.641
H40.5	その他の眼疾患に続発する緑内障	10.263
H25.0	老人性初発白内障	10.175
H33.0	白内障，詳細不明	10.131

眼に関する疾病

11番目の要素（因子）

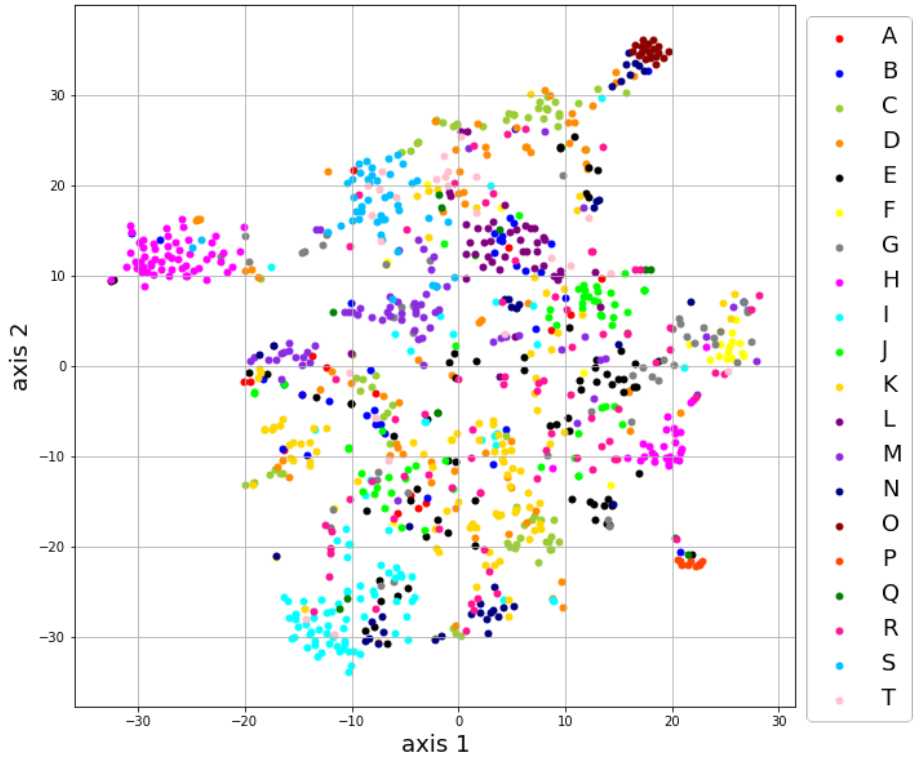
ICD-10	疾病名	値
I45.6	早期興奮症候群	11.087
I49.0	心室細動及び粗動	9.807
I47.1	上室(性)頻拍(症)	9.349
I47.2	心室(性)頻拍(症)	9.171
I42.0	拡張型心筋症	9.128

心臓に関する疾病

各要素(因子)に対し、**意味解析**を行うことができる

疾病の特徴表現解析

- 疾病の2次元マップへの埋め込み
  - 位置が近い疾病＝「同時にリスクとなりやすい」



ICD-10(疾病の分類)の大分類

各疾病に対し、一度に
類似性や関連性を可視化できる

全体の目次

- はじめに
  - 自己紹介
  - 本日の内容
- 研究紹介
  1. 「行列因子分解を使用した個別患者ごとの疾病予測および医療事象の特徴表現抽出」
  2. 「IPWを用いた医療における多種類介入のバイアス除去学習」

研究２ 目次

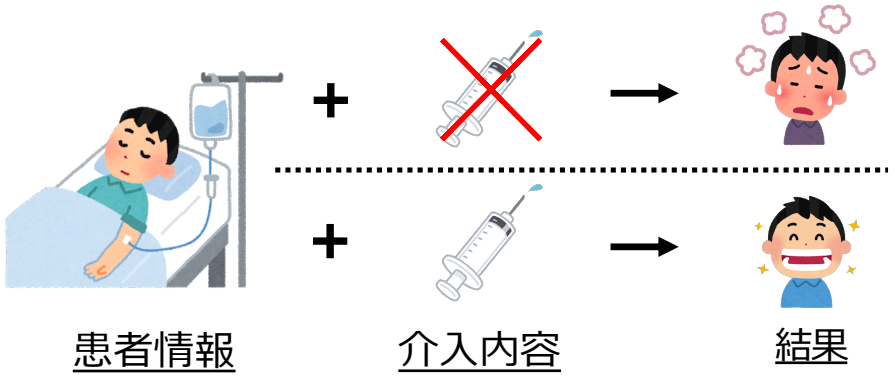
- はじめに
- 関連研究
- 提案手法
- 数値実験

研究２ 目次

- はじめに
- 関連研究
- 提案手法
- 数値実験

研究背景

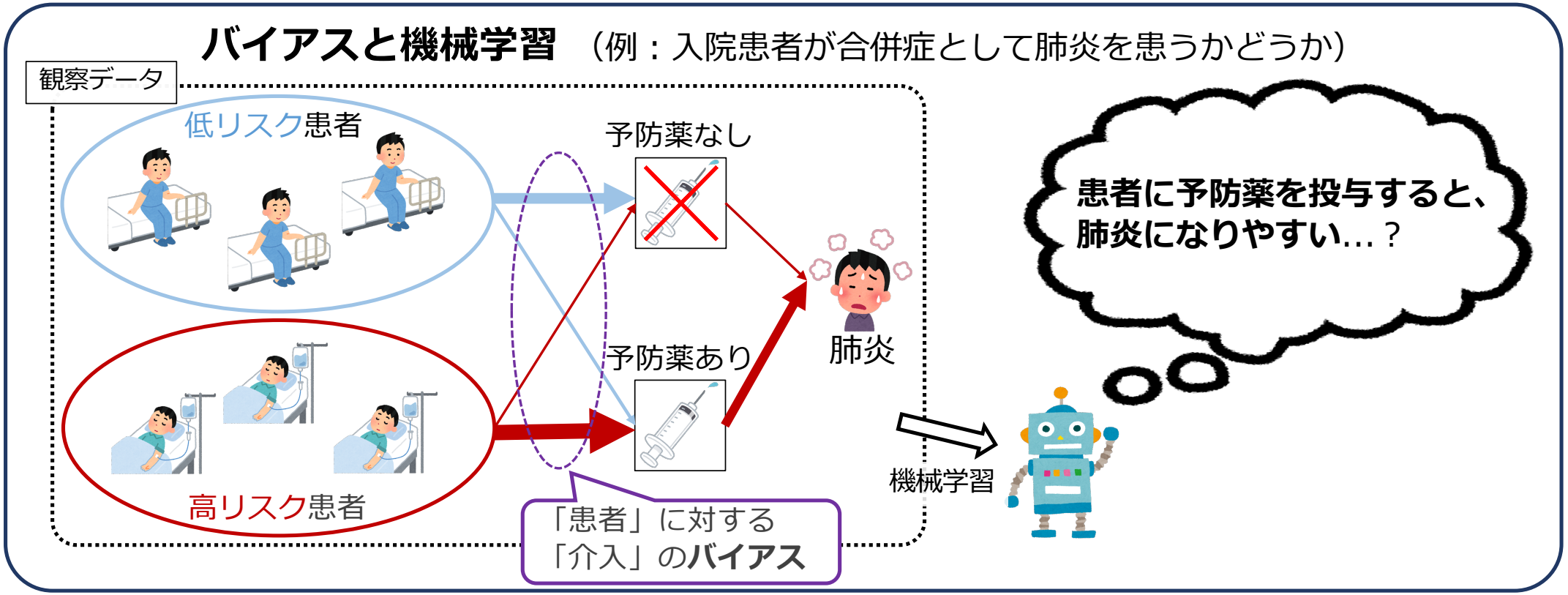
- 蓄積された電子カルテデータの2次利用が注目されている
- 機械学習による「**治療効果予測**」によって、医師へのサポートが可能



観察データでは「患者情報」に対する「介入内容」の分布は偏っている（**バイアス**）

そのまま機械学習を行うと、
元の傾向とは異なる介入に対して、**正しく予測を行えない**

研究背景



データのバイアスを除去して学習できる手法に関して検討する

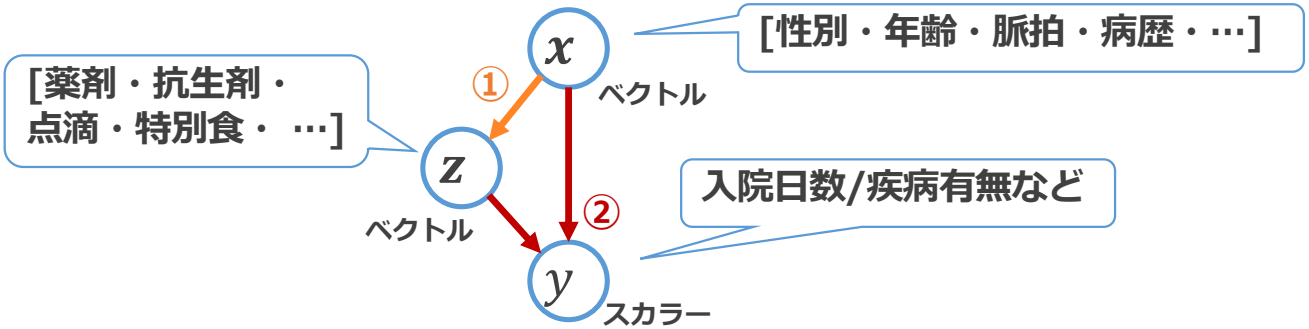
問題設定

- 前提

多種類あると想定

① 患者の「**状態**（$x \in \mathbb{R}^d$）」から、医師の「**介入**（$z \in \{0,1\}^K$）」が決まる

② 「**状態**」および「**介入**」から、「**結果**（$y \in \mathbb{R}$）」が決まる



- 問題：「状態x」 と 「介入z」 から「結果y」 を予測する

  - 「バイアス有りデータ」で学習し、「バイアス無しデータ」での予測精度を比較する
  - これにより、患者ごとに最適な治療方法を提示することに繋がる

研究２ 目次

- はじめに
- 関連研究
- 提案手法
- 数値実験

機械学習における損失関数

- 損失関数(Loss Function)とは
  - この関数を小さくする→モデルがデータにフィット
  - 具体例（回帰の場合）：

i: 各学習データの番号（$1,2,\ldots.,n$）
y_i: 実際の値、$\hat{y}_i$: モデルの予測値

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

各データの合計

実測値と予測値の差

機械学習 × IPW (Inverse Probability Weighting)

- 機械学習へのIPWの導入[Schnabel et al. 2016]
  - 学習時の損失関数の式（「介入($z \in \{0,1\}$)」が1種類の場合）

実測値と予測値の差

$$L_{IPW} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{p(z = z_i | \boldsymbol{x} = \boldsymbol{x}_i)}$$

各患者の合計

「その患者に実際に行われた介入」の発生確率（**一般化傾向スコア**）

介入の珍しさ	一般化傾向スコア	損失への影響度
起こりやすい	大	小
↕	↕	↕
起こりにくい	小	大

「起こりやすい介入」と「起こりにくい介入」を
一様的に扱い、バイアスを除去して学習できる

研究２ 目次

- はじめに
- 関連研究
- **提案手法**
- 数値実験