匿名化された健康診断と診療履歴の時系列データによる糖尿病罹患予測

明治大学総合数理学部 先端メディアサイエンス学科
菊池研究室 3年
清水 正浩　石山晴斗

背景

- 機械学習やAIの発展によるビッグデータの利活用の増加
- ビッグデータのうち、健康診断データは病気の罹患を予測する有効な情報

先行研究

- 匿名加工情報の有用性を評価
- 健康診断の情報がどの程度一意であるのか調査

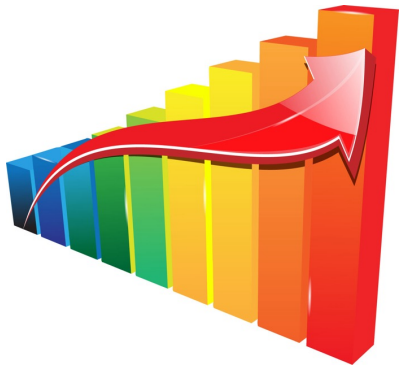
先行研究では, 各個人の診療履歴を考慮していない

⇒ 身体的特徴量の時系列変化は個人を識別する際、有効な情報

[1]池上（2020年）: 匿名加工情報の応用 (2): 各種傷病を予測する健康診断モデル
[2]伊藤（2020年）: 匿名加工情報の応用 (1): 健康診断データとレセプトデ ータの分析とプライ バシーリスク評価

研究目的

- 診療履歴を考慮していないデータと診療履歴を考慮したデータを比較
- 時系列を考慮した際の一意率への影響を考察（一意率が大きければ匿名性が弱い）



一意率・・・医療データが一意である割合

データセットについて(2.1, 2.2)

- 匿名加工した健康診断データ、基本データ（年齢、性別）、傷病レセプトデータ（通院記録）を使用
- 欠損値レコードや他の特徴量との相関が高い特徴量を削除などの前処理

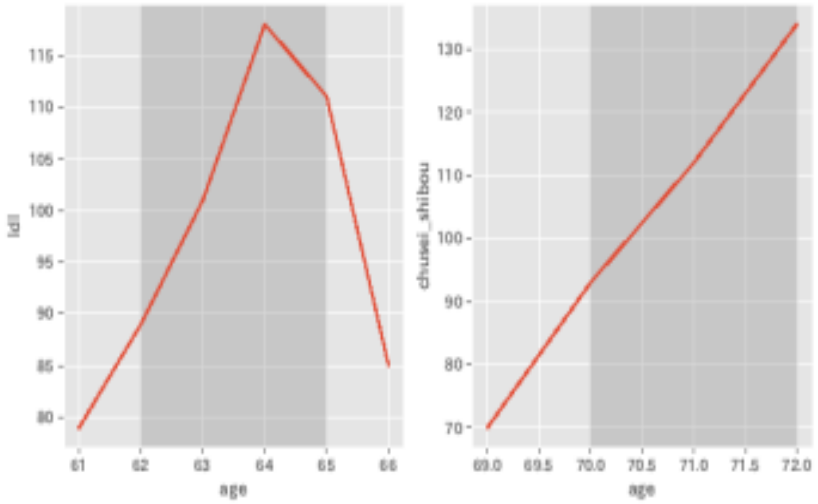
	対象年数	被験者数	身体的特徴数	問診結果数	特徴量数	レコード数	欠損値セル数
処理前	7	2,345,128	55	50	105	7,028931	439,127,300
処理後	7	172,819	11	17	28	1,858,163	0

特徴量変化の調査(2.3)

- 糖尿病に罹患した時期の特徴量の変化を10名調査

＜結果＞

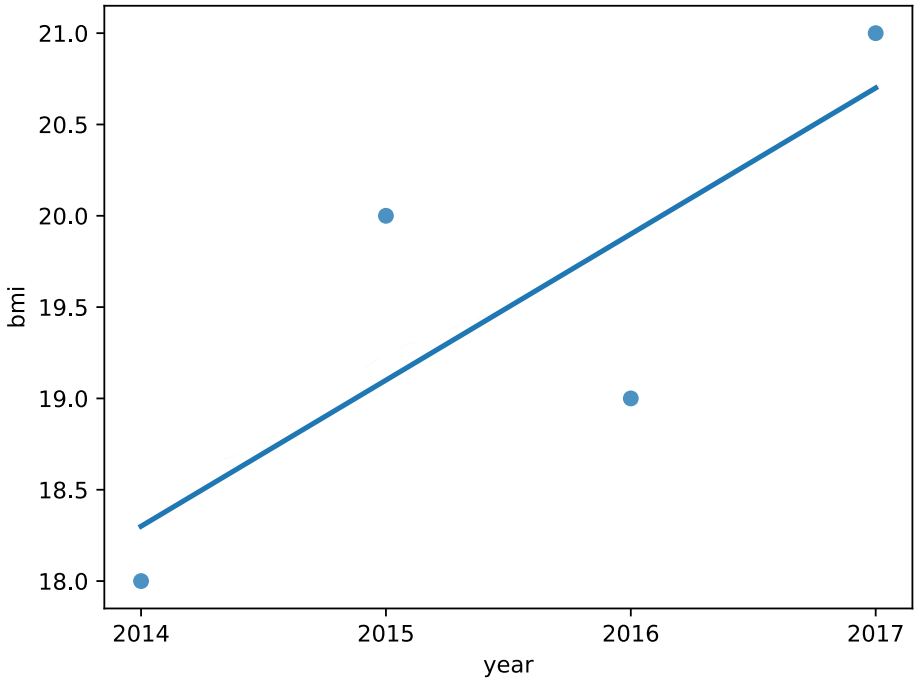
- 10名中7名の中性脂肪の値の増加
- 10名中6名の血清クレアチニンの値の減少
- 10名中5名の歩行 or 身体活動コードか，30 分以上運動コードのどちらかにおいて運動の頻度が少なくなる変化　etc.



時系列情報の付加(2.4)

1. **linearアルゴリズム**

   各特徴量の回帰直線の回帰係数を付加



分析方法(3.1)

1. ロジスティック回帰をし、時系列情報を考慮したデータの有無が診療結果の統計的有用性にどのように影響するかを調査
2. 3 年後までに糖尿病罹患する予測モデルを作成し、その精度を用いて評価
3. 診療履歴を考慮した一意率の調査

※tiltアルゴリズムはエラー発生のため、ロジスティック回帰ではlinearアルゴリズムのみ使用

1.ロジスティック回帰(4.1, 4.4)

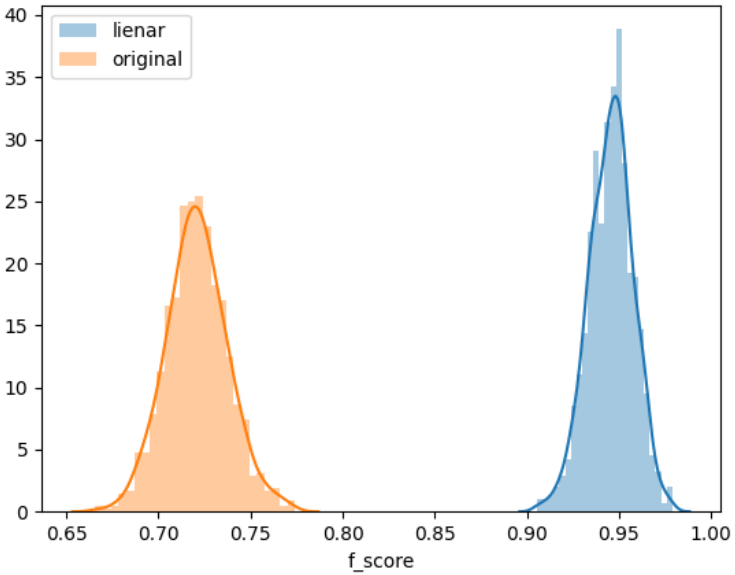
- 腹囲実測のオッズ比が0.07増加
- 拡張期血圧付加のオッズ比が0.957
- 服薬１血圧と年齢が±0.1でp値が低い
- 付加後の腹囲実測は元のp値と腹囲付加のp値が低いため不安定？
- 他3つはp値が0.5を下回るため有意

	オッズ比	提案方式	p値	提案方式
腹囲	1.037	1.106	0.645 ✖	0.362 ○
腹囲付加	-	0.992	-	0.698 ✖
拡張期血圧付加	-	0.957		0.482 ○
服薬1血圧	0.899	0.906	0.017 ○	0.056 ○
年齢	1.116	1.139	0.011 ○	0.005 ○

オッズ比･･･その変数が結果に影響与える程度
P値･･･有意であるか評価する指標（0.5以下で有意）

2.罹患予測モデル(4.2, 4.4)

- 平均・中央値・最小値・最大値でlinearアルゴリズムが最大
- 標準偏差でlinearアルゴリズムが最小



診療履歴を考慮した方が高い精度

	平均	中央値	標準偏差	最小値	最大値
Original	0.721	0.721	0.0168	0.666	0.774
Linear	0.950	0.950	0.0120	0.906	0.979

3.診療履歴を考慮した一意率(4.3, 4.4)

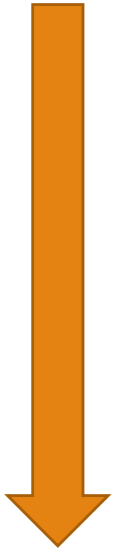
- 診療履歴の長さと一意率は比例

特徴量ベクトルが長いと、一致する確率が小さくなる

- 身体的特徴量の方がカテゴリカル変数より一意率が高い

カテゴリカル変数は離散値なため値の種類が少ない

長さ	一意率
2	0.262
3	0.424
4	0.437
5	0.687
6	0.783
7	0.795



今後について

＜結果＞

- 診療履歴を考慮した方がモデルの精度上昇
- 診療履歴の履歴期間が大きくなると一意率が上昇

＜課題>

- より安定したアルゴリズムの開発
- 匿名性が失われるという問題点の解消