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

2020. 10. 29.

CSS 2020

池上和輝　伊藤聡志　菊池浩明

明治大学大学院　先端数理科学研究科 先端メディアサイエンス専攻

背景

- 健康診断データの利用は傷病罹患の分析に有効
  - 野田らは，住民検査結果と人口動態統計死亡票分析し， 検査項目と死亡との関係を明らかにした．[1]
    - 例）女性の総コレステロール低値ならば，脳卒中による死亡リスクが1.98倍
  - 川南らは，喫煙習慣によるがん，肺 がん死亡へ影響を分析[2]
    - 毎日喫煙する集団の肺がん死亡の相対危険度が男性で 6.67 倍

病気にかかる
リスク算出

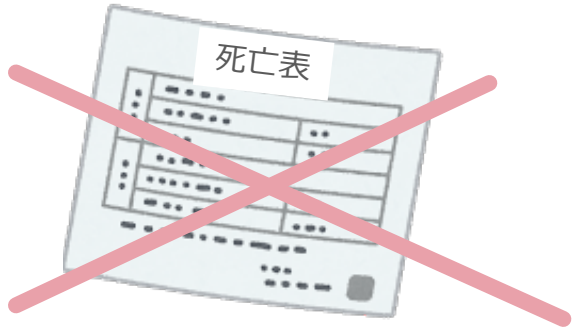
[1]野田 博之ら　住民健診 (基本健康診査) の結果に基づいた脳卒中・ 虚血性心疾患・全循環器疾 患・がん・総死亡の予測，日本公衛誌 53: 265–277，2006.
[2]川南 勝彦ら　NIPPON DATA80 研究グループ，喫煙習慣の全死因，が ん，肺がん死亡への影響に関する研究:NIPPON DATA80，日本衛生学雑誌.2003

問題点

問題点

- 2017年5月の個人情報保護法の改正に伴い，死亡票が使用できない
- 個人情報は利用目的を特定して適切に取り扱うことが定められた
- 病歴等は要配慮個人情報に分類されるため，個人の同意なく取得が禁止

これまでの大規模なコホート研究は違法



問題点の解決手法

■解決手法

ロ匿名加工された20万人分の健康診断・傷病データを使用

	野田ら	本分析
データ利用方法	人口動態統計死亡票の目的外利用	匿名加工情報
人数	92,277	68,629
説明変数	12	38
傷病数	4	274
対象期間	1993 - 2001(9年間)	2008 - 2016(9年間)

[1]野田 博之，磯 博康，西連地利己，入江ふじこ，深澤 伸子， 鳥山 佳則，大田 仁史，能勢 忠男，
住民健診 (基本健康診 査) の結果に基づいた脳卒中・ 虚血性心疾患・全循環器疾 患・がん・総死亡の予測，日本公衛誌 53: 265–277，2006.

研究目的

■健康診断データから有益な知見を得る

■リサーチクエッション

1. 健康診断と傷病罹患の関係
2. 健康診断データから傷病罹患予測モデルの作成・評価
3. 匿名加工によるの予測モデル精度の劣化

データセット

■概要

- **健康診断データ**　：10年間(2008-2018)の健康診断結果
- **傷病レセプトデータ**：患者の診断された傷病の記録
- 医薬品レセプトデータ：患者が処方された医薬品の記録

■健康診断データのクレンジング

- 分析のため欠損値などの不要なレコードを削除

	対象年	レコード数 N	欠損値セル数	説明変数の数 M
処理前	2008-2018	964,635	10,536,861	49
処理後	2008-2016	203,521	0	38

レセプトとの突合

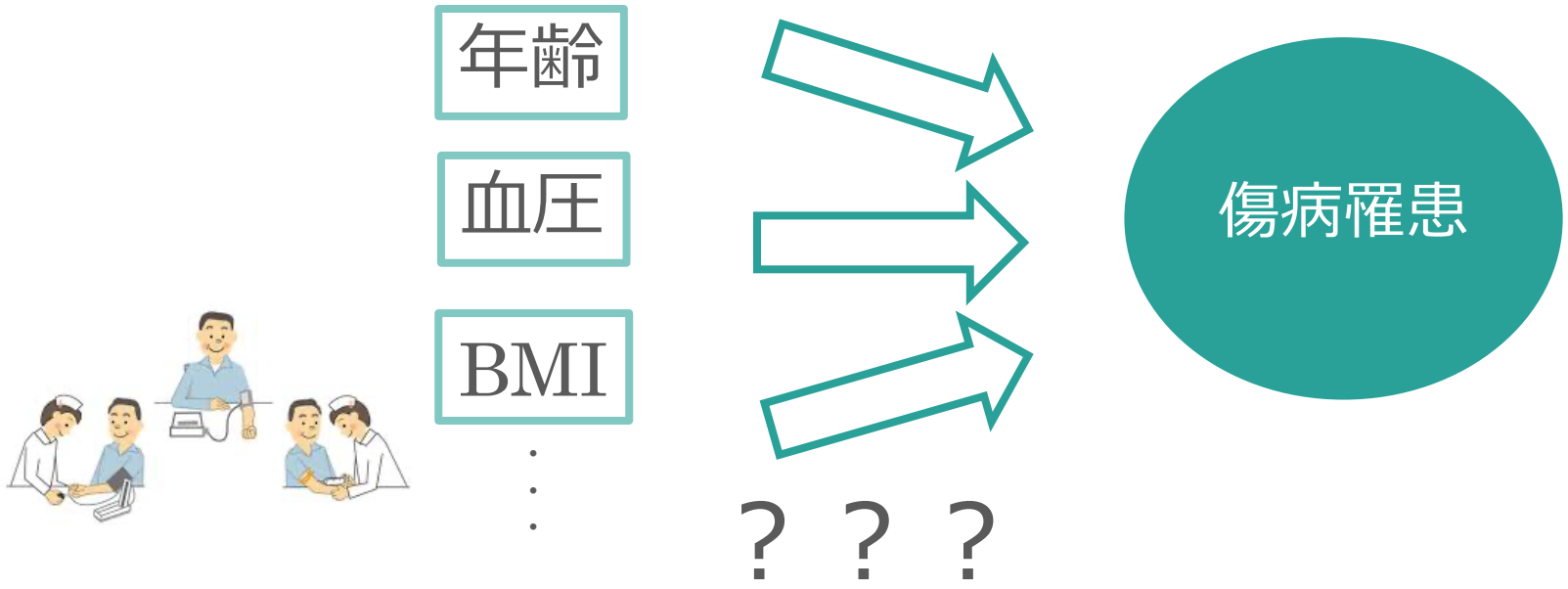
- **ユーザID**と**受診年**から突合を行う.
- 健康診断受診年から3年以内に傷病レセプトがあれば，そのユーザは罹患したと見なす.
- 罹患対象レセプトを常に健康診断から3年分確保するため，健康診断データは2008-2016年を使用する.

例）健康診断をX年に受けたときの，突合候補レセプト

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
健康診断	X		X						X		
傷病レセプト											

リサーチクエッション

1. 健康診断と傷病罹患の関係
2. 健康診断データから傷病罹患予測モデルの作成・評価
3. 匿名加工によるの予測モデル精度の劣化



分析方法１：ロジスティック回帰

- 健康診断と傷病罹患の関係をロジスティクス回帰により分析する
- 野田らのコホート研究[1]の結果と比較

	野田ら	本分析
データ利用方法	人口動態統計死亡票の目的外利用	匿名加工情報
人数	92,277	68,629
説明変数	12	38
傷病数	4	274
対象期間	1993 - 2001(9年間)	2008 - 2016(9年間)
被験者の年代	40-79	19-74
分析方法	Cox回帰	ロジスティック回帰
目的変数	死亡	三年以内の罹患

ロジスティック回帰結果（一部）

* 5%有意水準

先行研究と一致

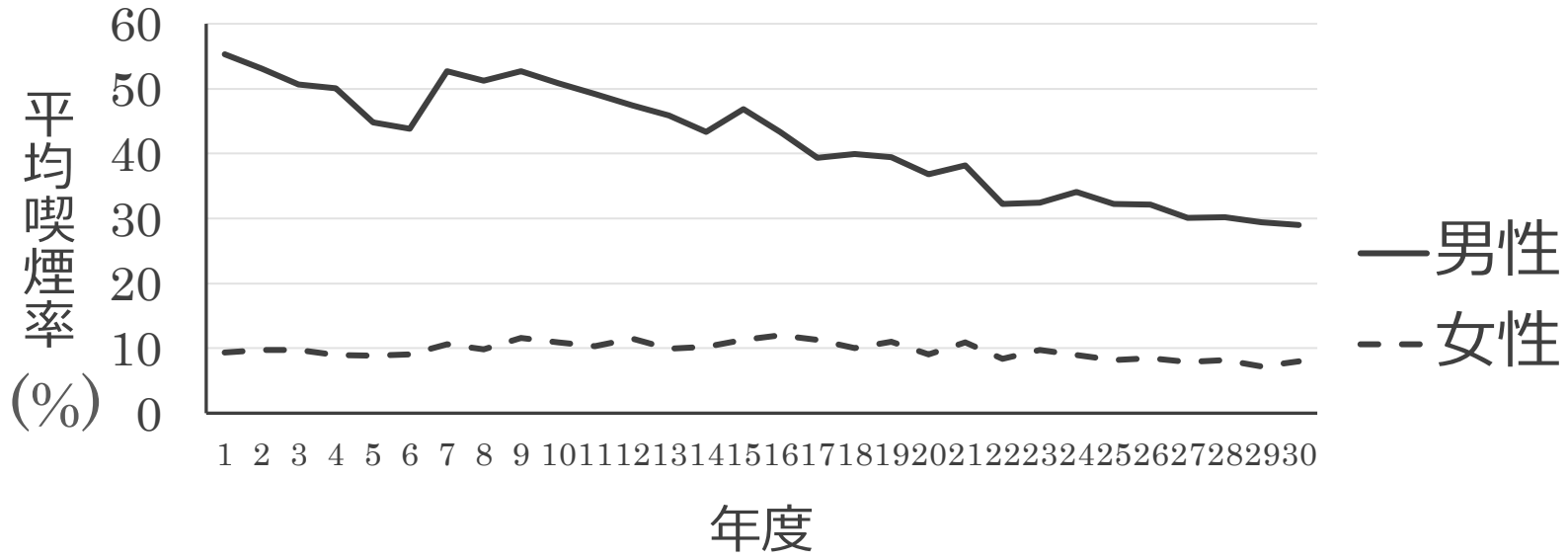
	脳卒中			がん		
	Estimate	オッズ比 OR	相対危険度 RR[1]	Estimate	オッズ比 OR	相対危険度 RR[1]
年齢（歳）	0.17 *	1.18	1.14	0.12 *	1.13	1.09
BMI	−0.02	0.99	1.00	−0.05 *	0.95	0.86
収縮期血圧	0.03	1.03	1.02	−0.02 *	0.98	-
HDLコレステロール	−0.003	1.00		−0.02 *	0.98	0.85
血圧治療	0.13 *	1.14	1.56	0.08 *	1.08	1.15
喫煙	0.01	1.01	1.27	−0.04 *	0.96	1.51
睡眠	−0.12 *	0.89		−0.06 *	0.94	
飲酒（ほとんど飲まない）	0.04 *	1.04		0.02 *	1.02	
運動習慣	−0.01	0.99		−0.03 *	0.98	

不一致

先行研究では，含まれなかった睡眠や飲酒，運動習慣と傷病の関係

喫煙因子の効果が不一致の理由

- 成人喫煙率(男性)
  - 2000年：55%，2018年：29%[2]
  - 喫煙者が，禁煙により非喫煙者となっている



- 先行研究では目的変数が死亡なのに対して，本分析では3年以内の罹患である.

[2]成人喫煙率（厚生労働省国民健康栄養調査），v

リサーチクエッション

1. 健康診断と傷病罹患の関係
2. 健康診断データから傷病罹患予測モデルの作成・評価
3. 匿名加工によるの予測モデル精度の劣化

	2008	2009	2010	2011	2012	2013	2014	2015	2016
健康診断									
傷病レセプト									

脳卒中に罹患

分析2：機械学習による罹患予測モデル

- 罹患を健康診断から274種類の予測モデルを作成
  - 傷病レセプトデータの274種類の病気

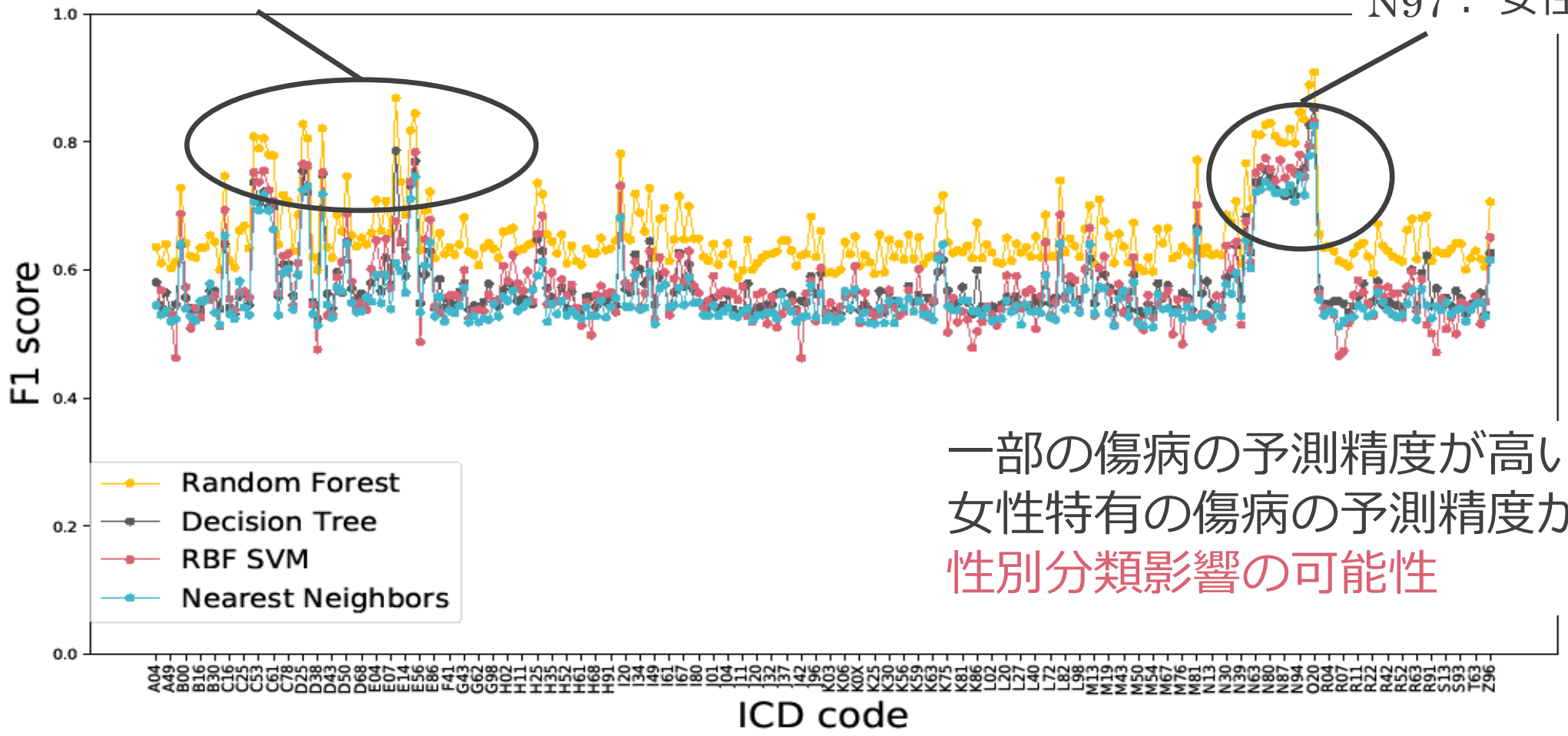
- 作成手順
  1. ４つ異なる予測アルゴリズムで学習
     - K−近傍法
     - SVM
     - 決定木
     - ランダムフォレスト

     ※ハイパーパラメータは全てデフォルト値使用
  2. ５分割交差検証，テストデータのF1 scoreにより評価

学習結果

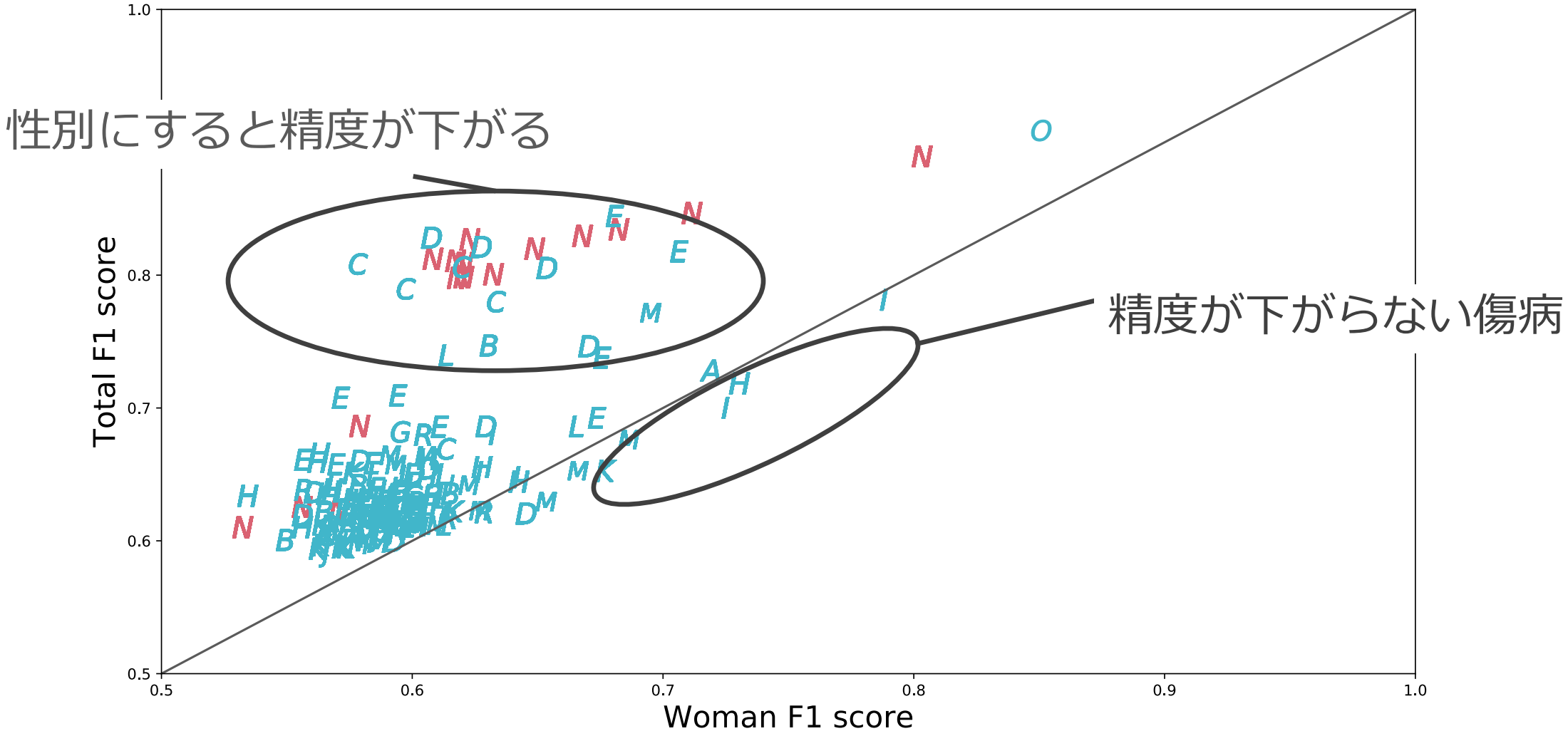
C50：乳房の悪性新生物
D41：女性生殖器の性状不詳又は不明の新生物

O20： 妊娠早期の出血
N97： 女性不妊症



一部の傷病の予測精度が高いが
女性特有の傷病の予測精度が高いため
性別分類影響の可能性

性別を考慮したモデルの評価



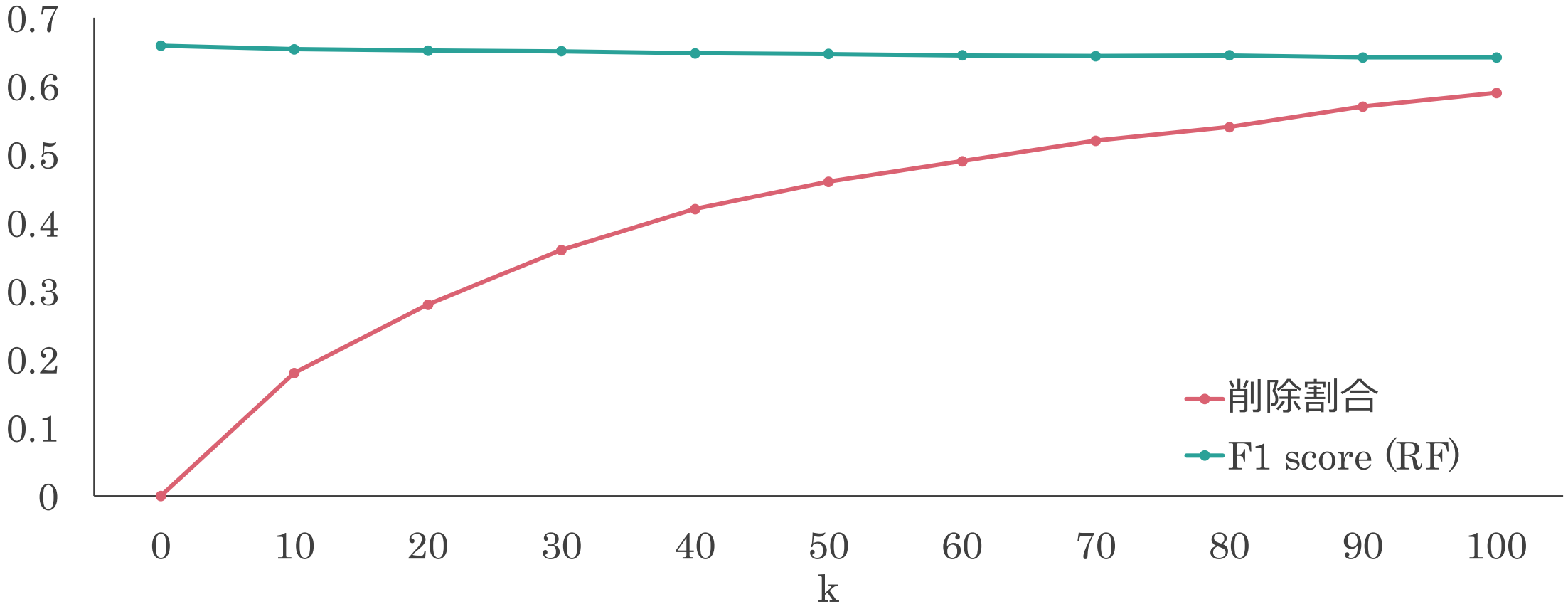
リサーチクエッション

1. 健康診断と傷病罹患の関係
2. 健康診断データから傷病罹患予測モデルの作成・評価
3. 匿名加工によるの予測モデル精度の劣化

分析方法3：k匿名

- 作成した予測モデルの精度が追加の匿名加工により、どの程度精度が劣化するかを確認する
- 方法
  1. 健康診断データの**13種類の問診結果**を疑似識別子(QI)とする
     - 喫煙，運動習慣，睡眠，食習慣，etc..
  2. QIをk=10~100匿名化する（10刻み）
     - 行削除のみ
  3. 匿名化したデータで、予測モデルの学習/評価

加工結果(RF)



k=100でレコードが約60%減少
RFで，F値が最大で0.02しか変化しない

まとめ

- 健康診断データは，非常に有効であり傷病罹患の予測などに活用可能
- 個人情報保護法の改正により，死亡票の使用不可
- 匿名加工情報である健康診断データから有益な知見を得た
  1. 健康診断と傷病の関係について匿名加工データから、既存のコホート研究と同様の結果が得られた。さらに、睡眠を十分にとることは3年以内の罹患リスクを0.89倍に下げるなど新たな知見を得た.
  2. 健康診断から3年以内の罹患を予測するモデルを279傷病について実施
     ランダムフォレストでは平均で65%の精度で予測できる
  3. K匿名化により、レコードが最大60%減少しても、F値は最大0.02しか変化しない。