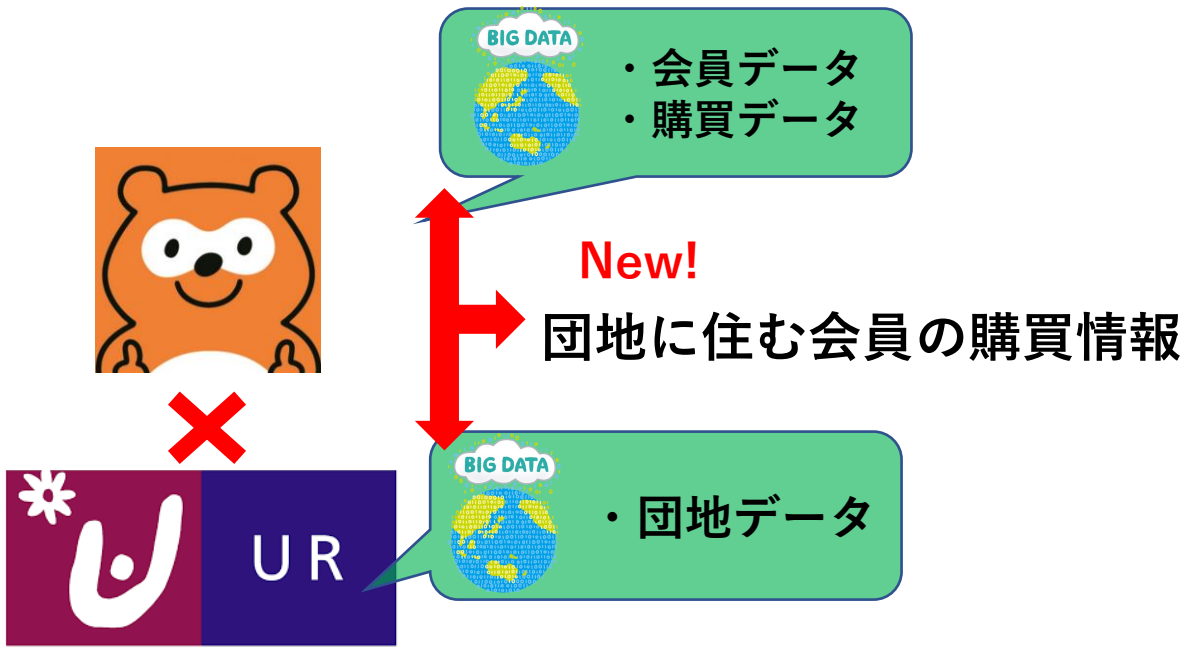
ミクロアグリゲーションを用いたk匿名化手法の提案と評価

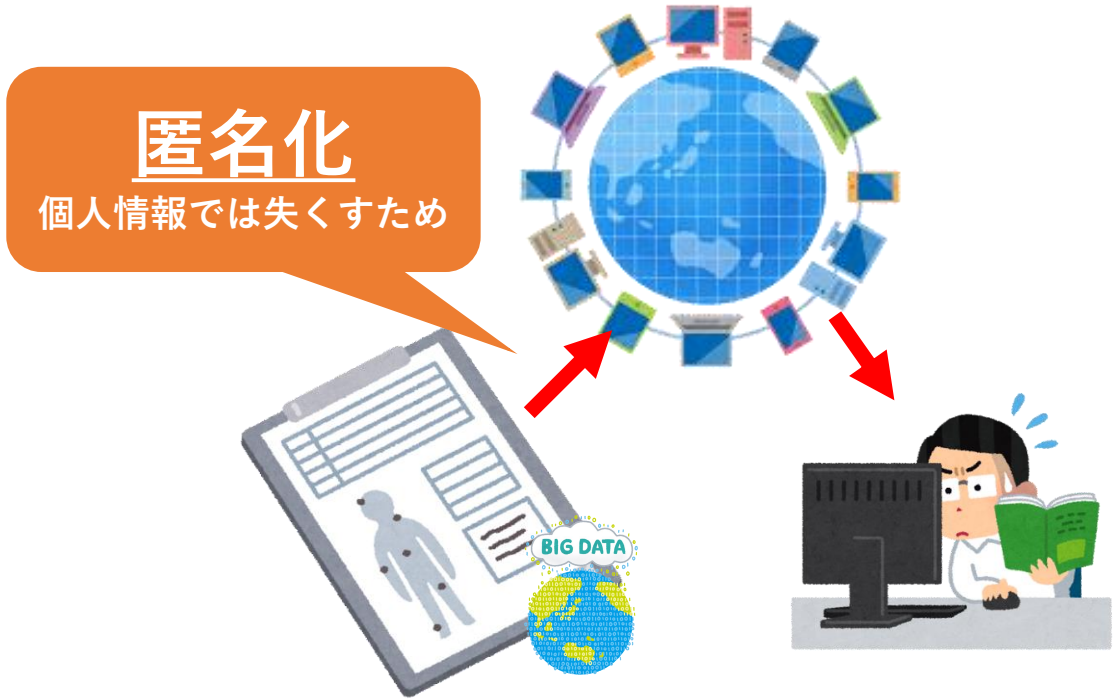
菊池研4年 入沢　響

背景(ビックデータの活用)

- 世はビックデータの加工から新たな知見を得る時代！
- 集めたデータを**<u>第三者提供</u>**をすることでマーケティングできたり新たな知見を得たりしている.



他社との連携による活用



第三者提供例

匿名化とは

使用データ：性別、年齢、身長をもつデータ
悪用を試みる人：その３列で個人を特定できる人を想定する

性別	年齢	身長	病気記録
男	20	170	○○○
男	20	170	○×△
男	20	171	○○△
男	20	172	○○○
男	20	173	○△×
男	20	173	×△×
男	20	173	×△○

加工前データ

Aくんは
・性別男
・年齢20歳
・身長171センチ
だから上から３番目だ

悪用大好きくん

性別	年齢	身長	病気記録
男	20	170	○○○
男	20	170	○×△
男	20	171	○○△
男	20	171	○○○
男	20	173	○△×
男	20	173	×△×
男	20	173	×△○

匿名加工後データ

人を特定できる列が同じ値の人が常にk以上にするような加工を**k-匿名化**と呼ぶ
（表のは2-匿名化）

性別、年齢、身長が同じ人が**２人**もいるからAくんがどれか分からないよ…

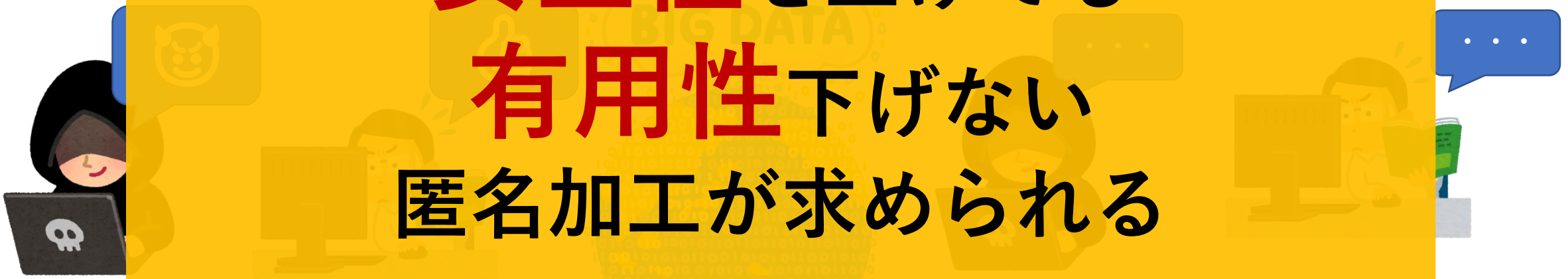
どんな匿名化が優れているのか

小　匿名加工　大

安全性down ↘
有用性up ↗

安全性up ↗
有用性down ↘

安全性を上げても
有用性下げない
匿名加工が求められる



先行研究の匿名化(レコード削除匿名*)

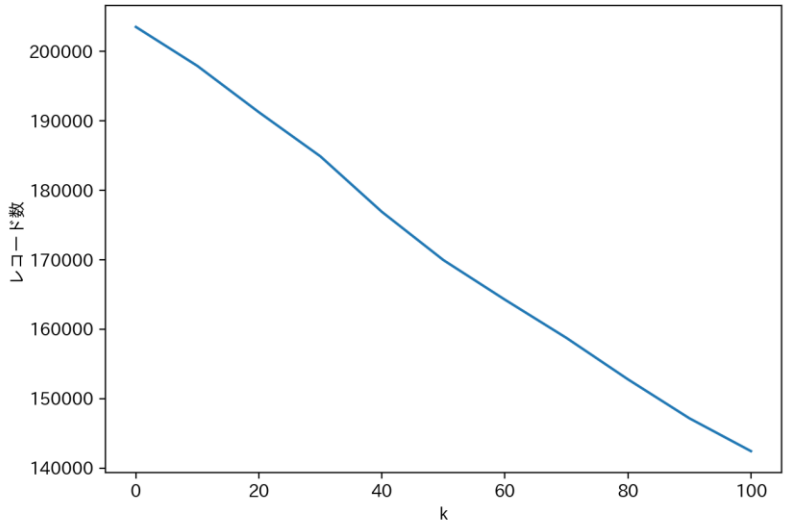
手法：識別列が同じ値がkより少ないのレコードは全て削除することでk匿名を満たす
メリット：理解が容易
デメリット:レコードが減少する(データとしての有用性が大きく落ちる)

性別	年齢	身長	病気記録
男	20	169	○○○
男	20	170	○×△
男	20	171	○○△
男	20	171	○○○
男	21	170	○△×
男	21	170	×△×

✖削除

性別	年齢	身長	病気記録
男	20	171	○○△
男	20	171	○○○
男	21	170	○△×
男	21	170	×△×

先行研究の2-匿名化の例



Kの変化によるレコード数の変化

*PWS CUP2020 のサンプル加工で提供された匿名化手法

論文概要

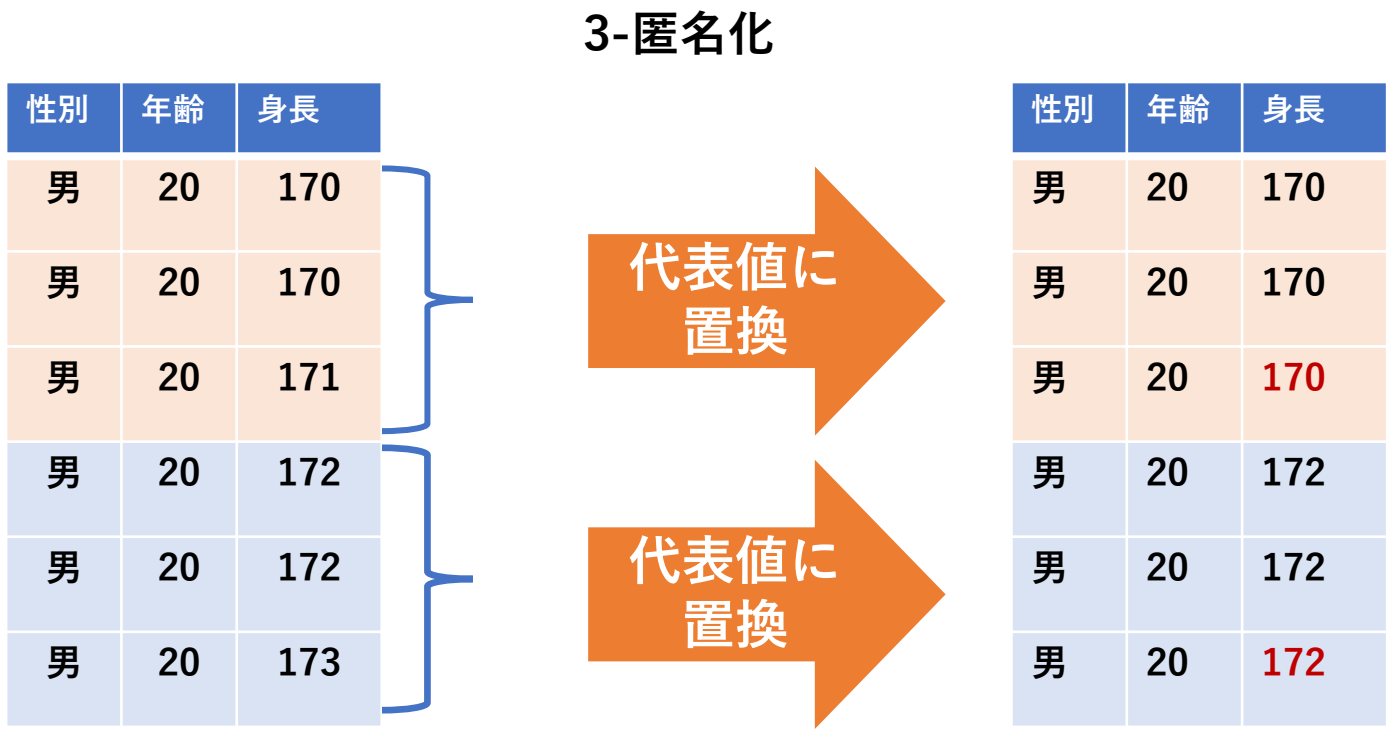
- レコードが削除されない**ミクロアグリゲーション***を2段階で用いた匿名加工手法の提案

- レコード削除匿名とともに評価し提案手法の有用性を評価

*匿名化技法としてのミクロアグリゲーションについて(2009)

ミクロアグリゲーション*とは

k個ごとにグルーピングして平均値等の代表値に置き換える→ミクロアグリゲーションによるk匿名



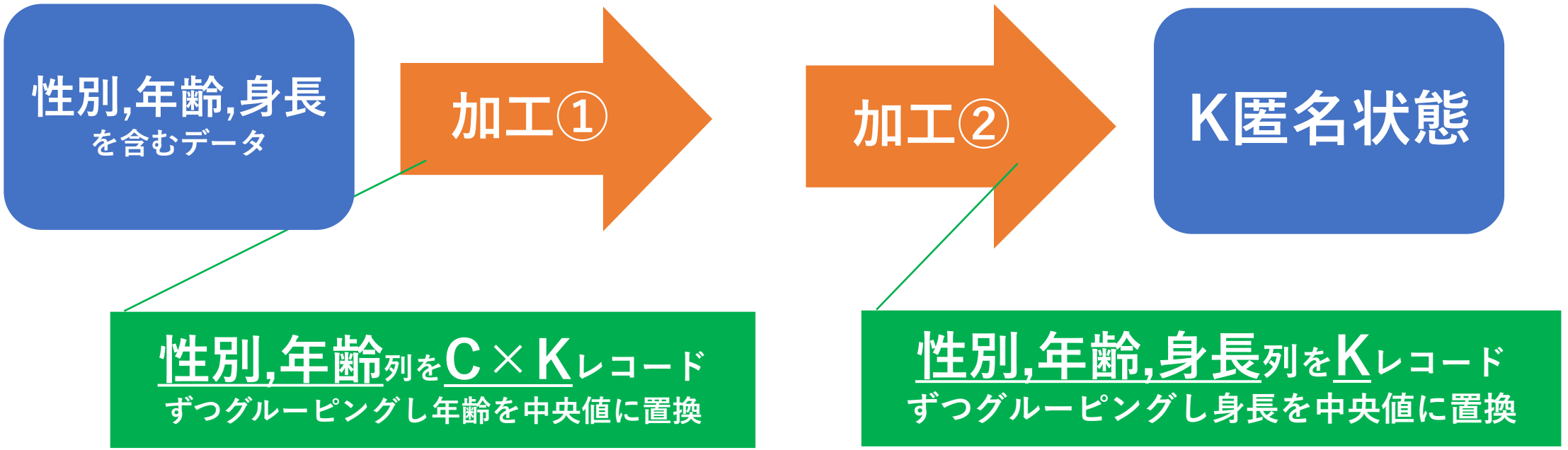
*匿名化技法としてのミクロアグリゲーションについて(2009)

入沢流匿名化手法

加工①、加工②の２段階

パラメータC,Kを利用し、識別列を「性別、年齢、身長」とする.(Cは定数)

加工①…性別,年齢でミクロアグリゲーション(年齢を加工)
加工②…性別,年齢,身長でミクロアグリゲーション(身長を加工)



提案匿名化手法の特徴

- 加工後もレコード数を保持できる
- 使用データの中からQIを「性別、年齢、**身長**」に指定している
- Kに加えて**C(1~10の整数)**をパラメータに使用し２段階での加工をしている

Aくん(160cm)　Bくん(180cm)

成人の身長は
基本的に**不変**なため、
識別リスクが高い

C

+
年齢加工増↗
身長加工減↘

-
年齢加工減↘
身長加工増↗

目的

Kごとに有用性の変化が
最も少ないCを使用することで
有用性が大きく下がることを減らす

有用性評価手法

評価手法

性別,年齢,身長の傷病への影響の変化から有用性を評価する

- 影響度は健康診断データから傷病列にロジスティック回帰した際のOR(オッズ比),p値を利用する
- 性別,年齢,身長のp値が0.05以下の傷病を利用する
- 加工前との比較を**RMSE**にて行う

性別,年齢,身長含む

3年以内の罹患歴があれば1,なければ0

レコード数	健康診断データ列	傷病列	合計列数
203,521	53	1428	1481

使用データ

例:がんの罹患歴に対するOR	
加工前	1.10
提案手法加工後	1.11
先行研究加工後	1.50

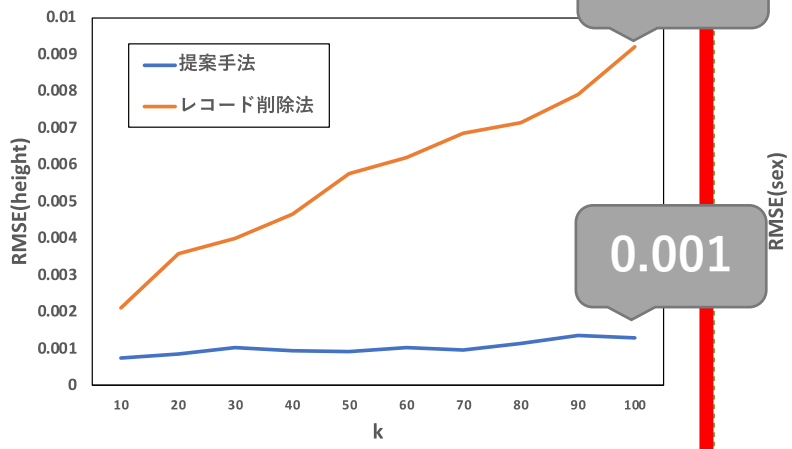
→提案手法の方が影響度の誤差が少ない

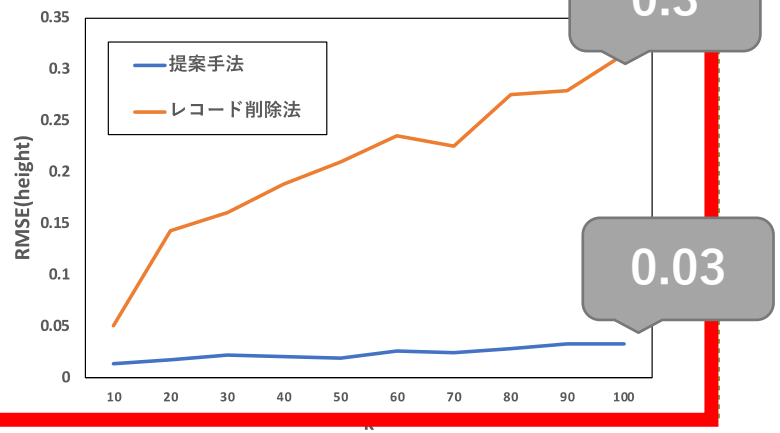
評価結果1

OR

P値

身長





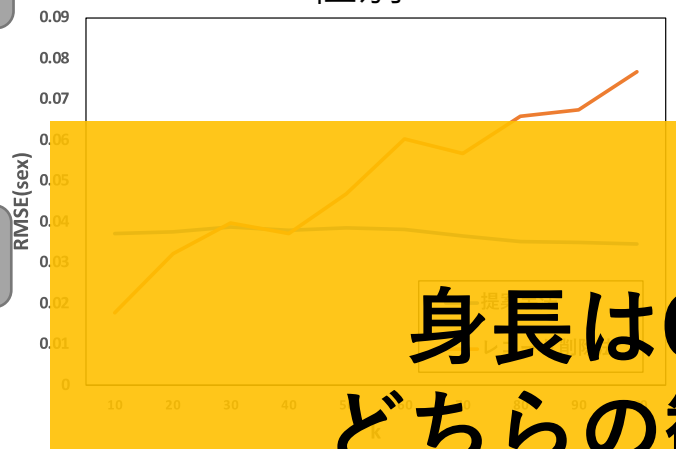
0.009

0.001

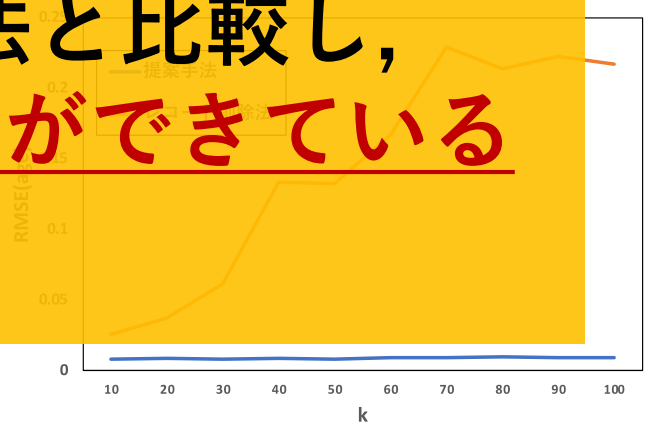
0.3

0.03

性別



年齢

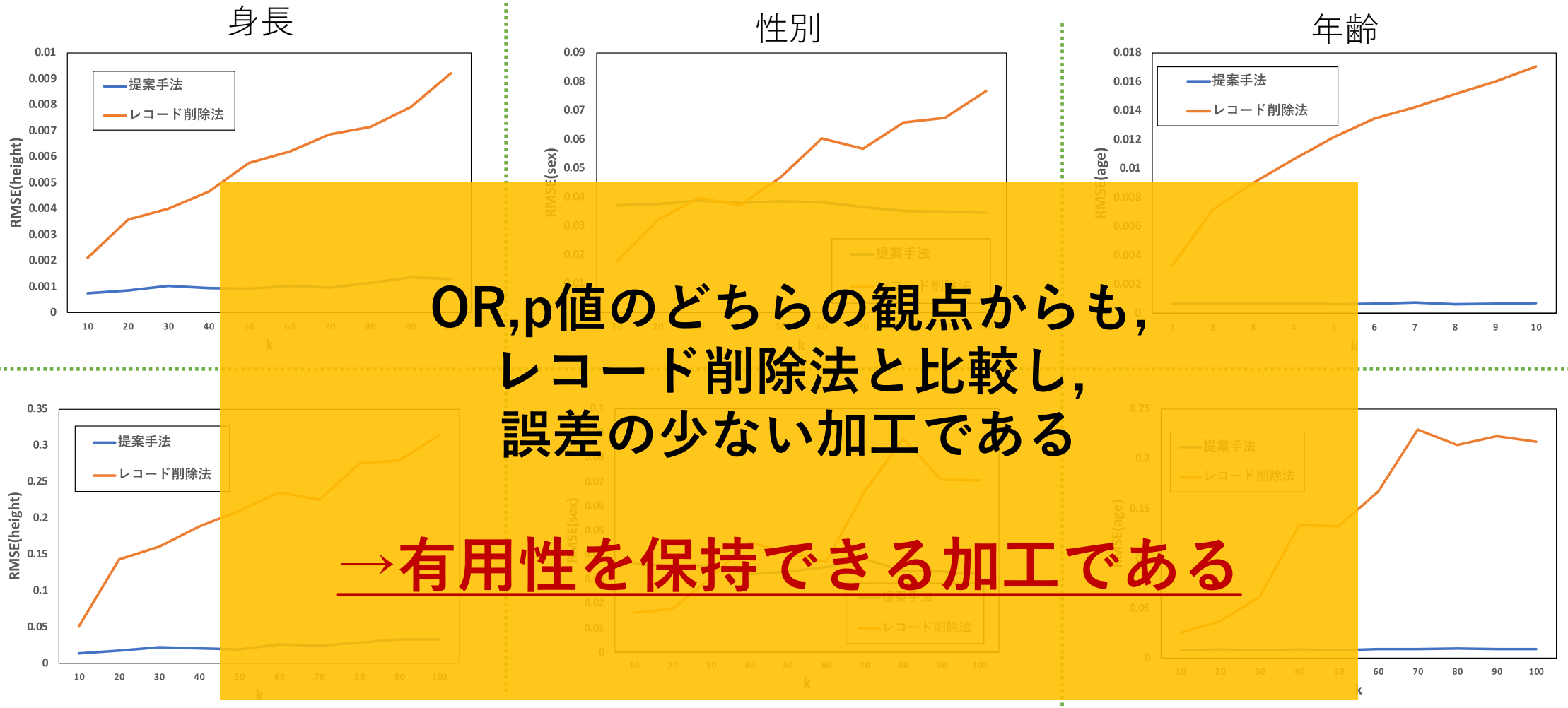


身長はOR,p値の
どちらの観点からも,
レコード削除法と比較し,
誤差を減らすことができている

評価結果2

OR

P値



OR,p値のどちらの観点からも,
レコード削除法と比較し,
誤差の少ない加工である

→有用性を保持できる加工である

考察と今後の課題

考察

- 匿名性を保ちながら, レコード数を完全に保つことができる.
- 先行研究と比較し, 傷病の予測の影響は低く, 有用性を保つ.

課題

- 比較対象である匿名化手法が１種類である
- 本稿で使ったデータに特化した匿名化手法であり,他のデータに適用するには,ロジックの調整が必要である