CSS 2020

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

伊藤聡志，池上和輝，菊池浩明
明治大学

研究背景

- 健康診断は非常に有用なデータであり，個人がこれから罹患する病気を予測できる可能性がある
- 野田ら(2006)[1]は健康診断データと住民健診データを分析することにより，検査項目と死亡との関係を明らかにした

研究目的

- 健康診断データや疾病と生活習慣の相関関係を明らかにし，疾病予防，生活改善，健康施策づくりに有益な知識を得ること

問題点

- 個人情報保護法の改正により，健康診断結果や病歴などは要配慮情報に分類され，利活用の際に特別な措置が必要になった

解決手法

- 2017年5月に施行された改正個人情報保護法で定義された匿名加工情報に注目する
- あるヘルスケア企業が収集・匿名加工した３つのデータを分析する（健康診断データ，傷病レセプトデータ，医薬品レセプトデータ）

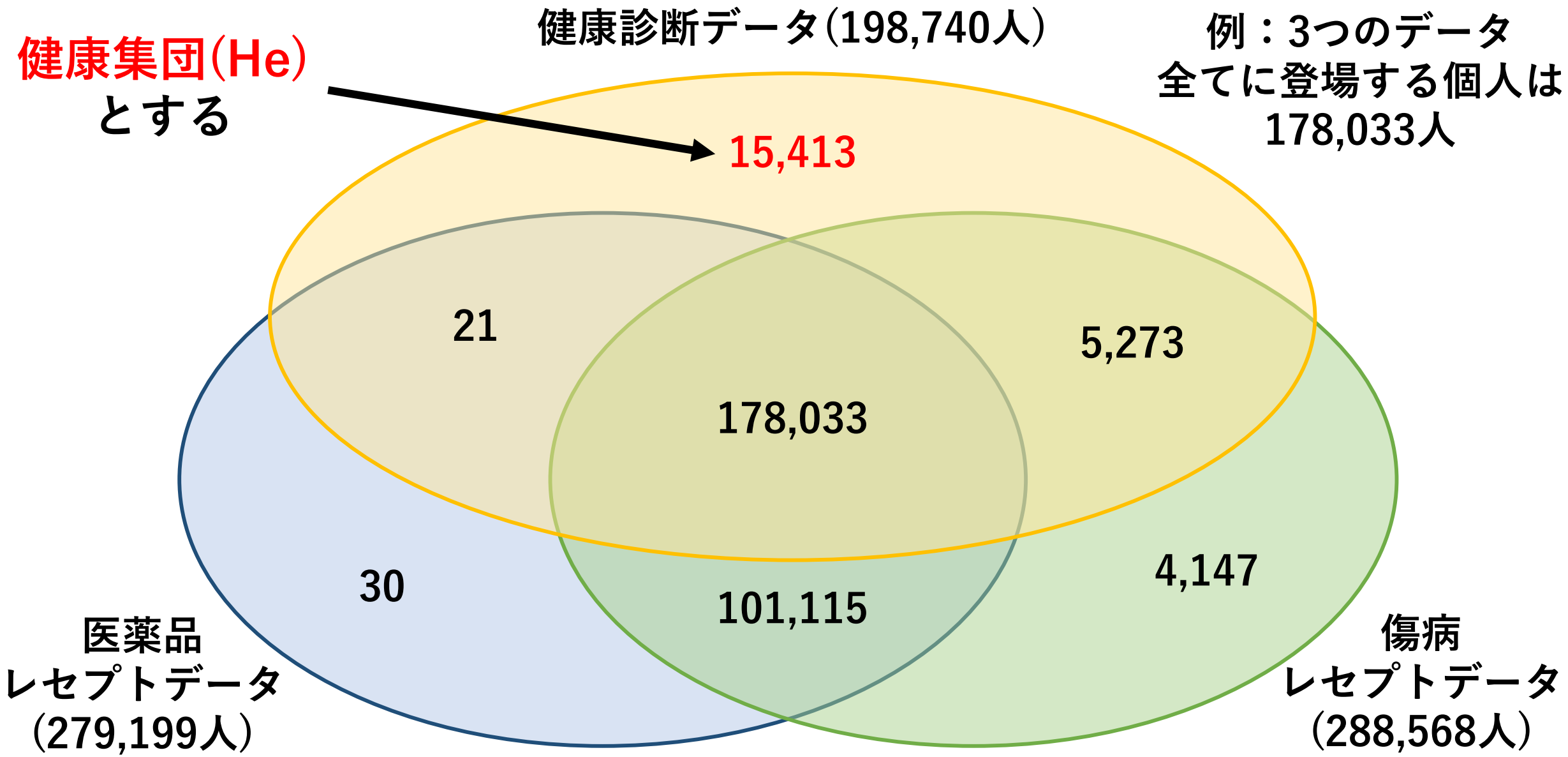
リサーチクエスチョン

1. データ中で特異なふるまいをしている記録はあるか？
2. 健康診断結果でこれから罹患する病気を予測できるか？
3. 患者の病歴/処方歴や健康診断結果はどれくらい一意であるか？
4. 病歴をk-匿名化することによって，データの安全性・有用性はどう変化するか？

データ概要

データ名	個人数	詳細	レコード数	属性数	属性例	レセプト枚数
健康診断データ	198,740	健康診断結果が記録されている	964,636	49	身長，体重 健康分布	-
傷病レセプトデータ	288,568	患者が診断された傷病が記録されている	39,363,878	15	傷病分類コード	11,912,236
医薬品レセプトデータ	279,199	患者が処方された医薬品が記録されている	31,465,504	21	医薬品分類コード	9,000,249

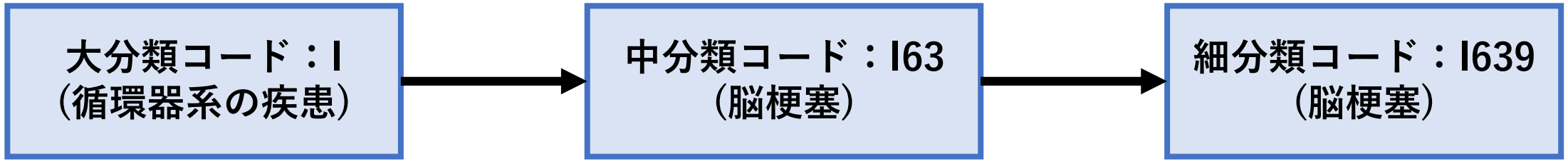
データ詳細（３データ間の包含関係と健康集団）



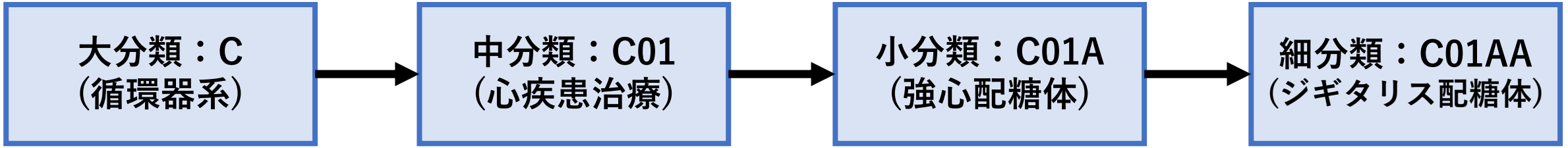
データ詳細（医薬品/傷病コード）

処方される医薬品や診断される傷病には分類コードが割り当てられている
（医薬品レセプト：第14~17属性，傷病レセプト：第7~12属性）
医薬品の分類には**ATC分類**が，傷病の分類には**ICD10**が用いられている
解剖治療化学分類　　国際疾病分類第10版

例：脳梗塞(I639)



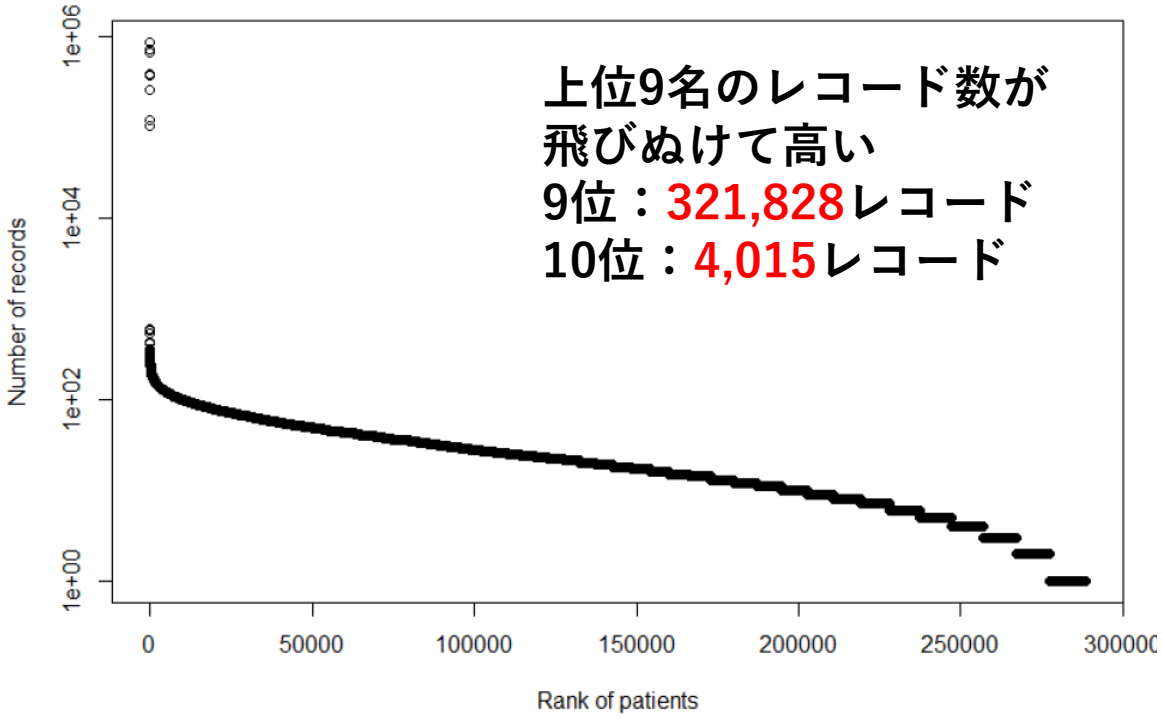
例：ジギタリス配糖体(C01AA)



データ中の特異な値

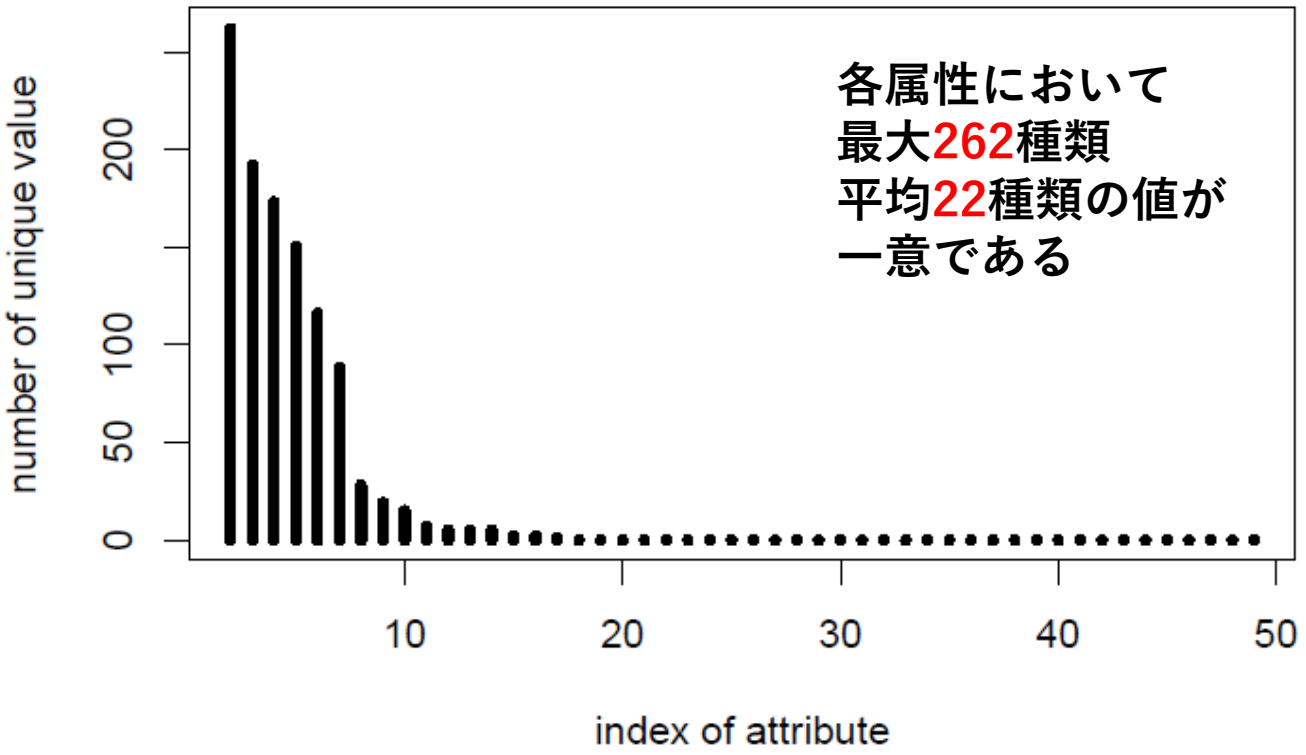
RQ1: データ中で特異なふるまいをしている記録はあるか？

傷病レセプトデータでの患者のレコード数分布



上位9名のレコード数が
飛びぬけて高い
9位：321,828レコード
10位：4,015レコード

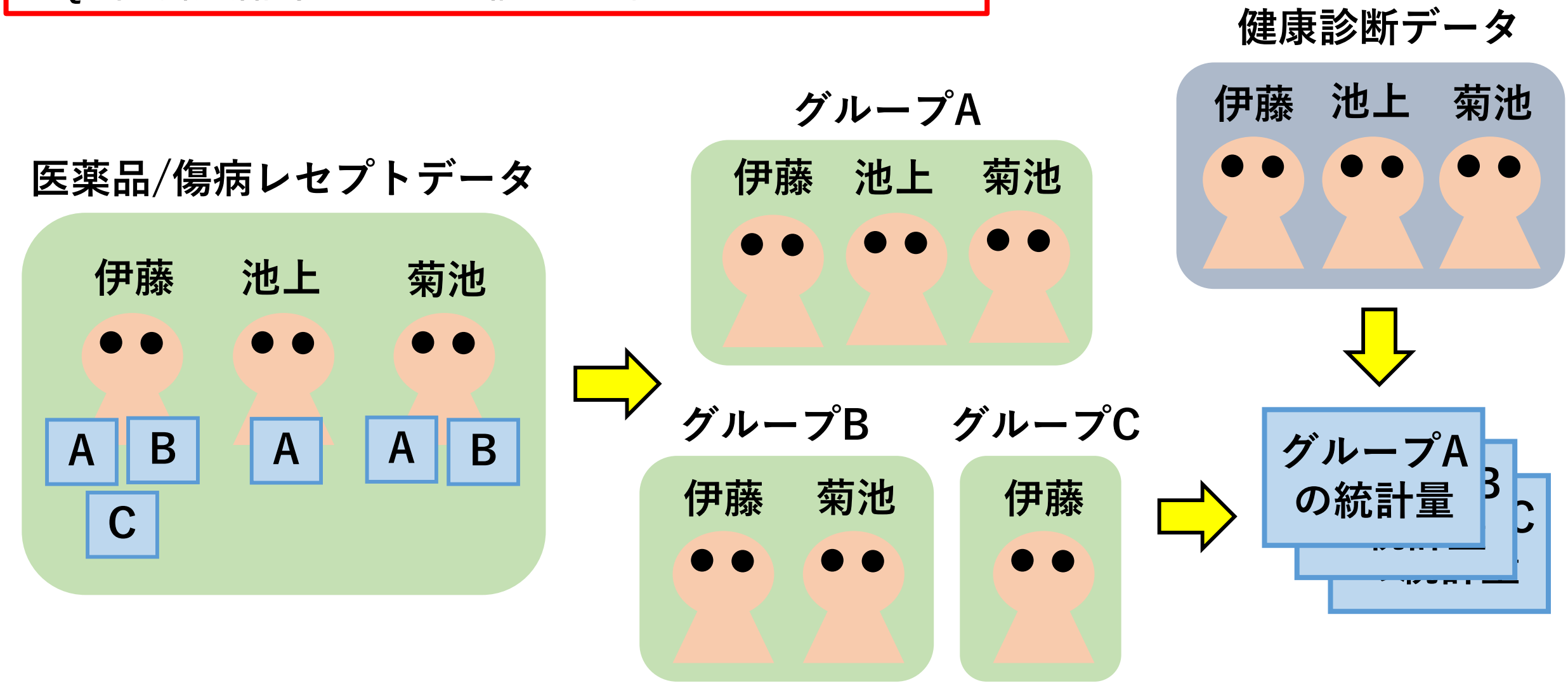
健康診断データ各属性での一意な値数



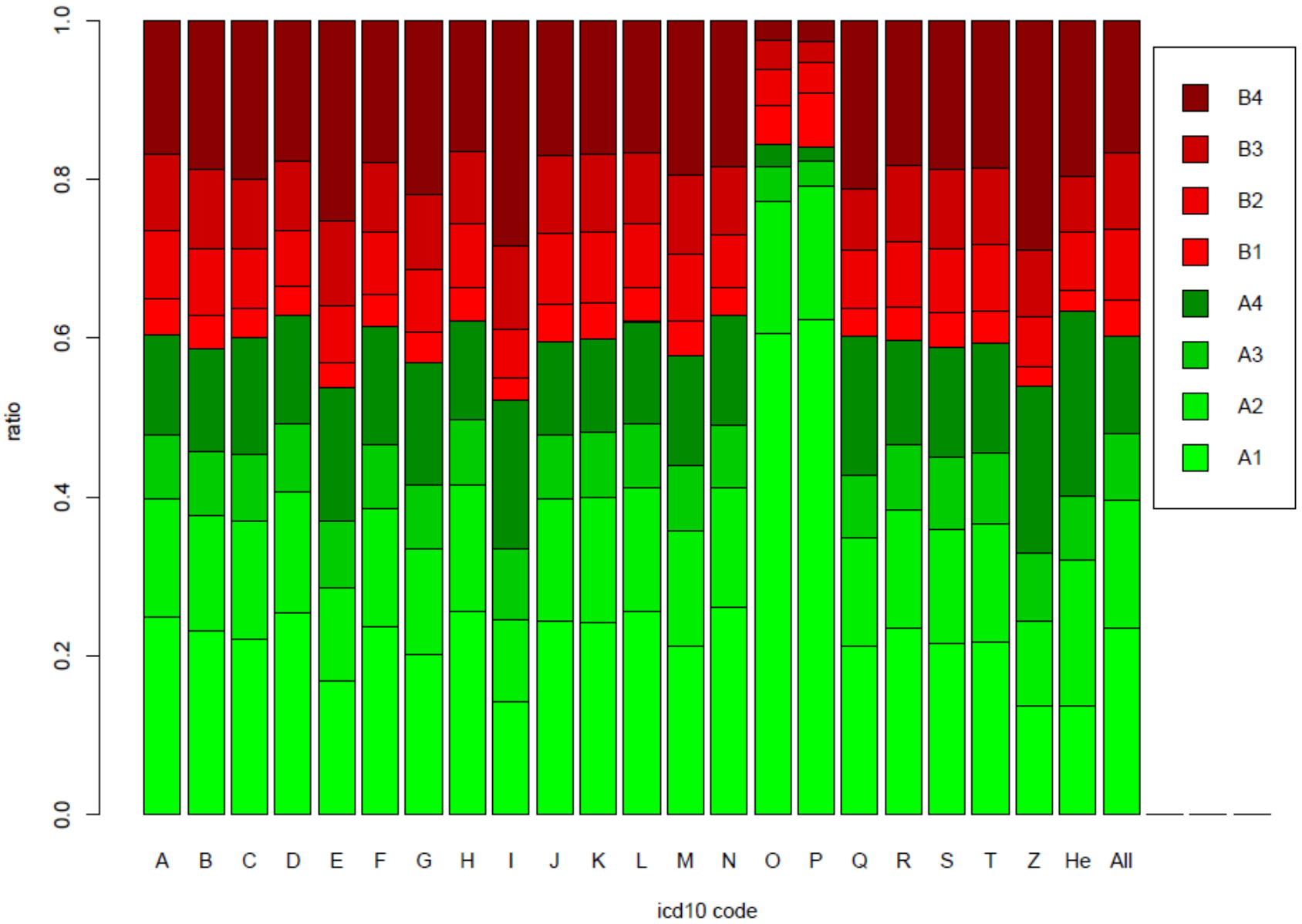
各属性において
最大262種類
平均22種類の値が
一意である

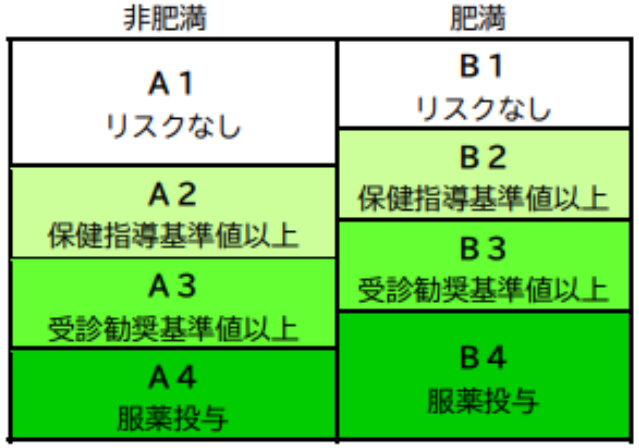
傷病/医薬品グループごとの分析

RQ2:健康診断結果でこれから罹患する病気を予測できるか？



傷病グループごとの違い(健康分布)



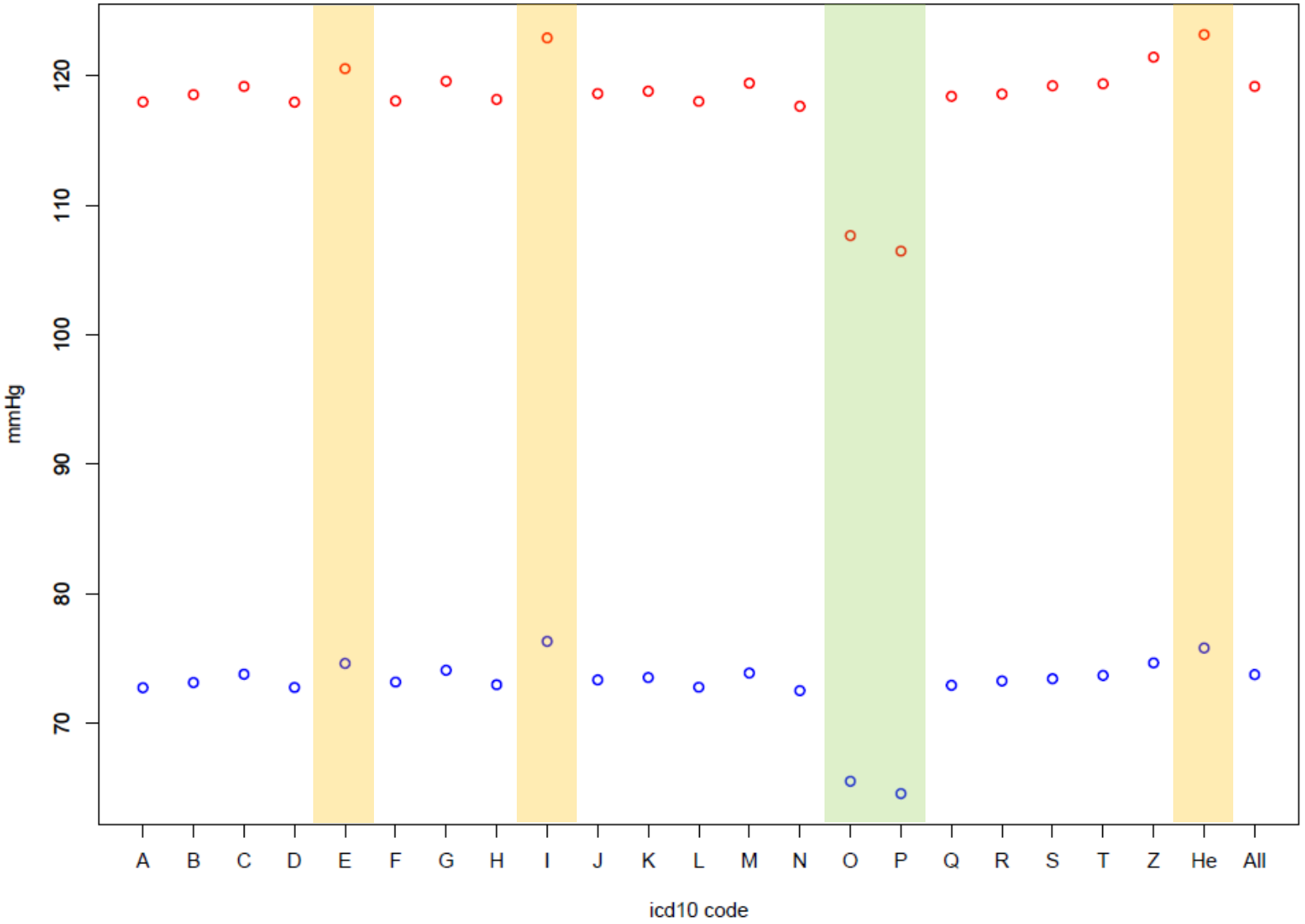


健康分布の割合はグループ間で差が見られた

- **傷病O,P(妊娠に関わるもの)の患者は肥満の割合が低い**
- **傷病E(内分泌,栄養および代謝疾患)や傷病Ⅰ(循環器系の疾患)の患者は肥満の割合が高い**

傷病Eの例：糖尿病，傷病Ⅰの例：脳梗塞

傷病グループごとの違い(収縮期/拡張期血圧)



赤点：収縮期血圧(上の血圧)
青点：拡張期血圧(下の血圧)

収縮期/拡張期血圧の平均値は
グループ間で差が見られた

- 傷病O,P(妊娠に関わるもの)の
  患者は血圧の平均値が低い
- 傷病E(内分泌,栄養および代謝疾患)や
  傷病Ⅰ(循環器系の疾患)の患者は
  血圧の平均値が高い
- 健康集団(He)の人たちも
  血圧の平均値が高い
- 健康集団の人たちは健康診断の
  結果が悪い傾向にあった
  (問診の答えは健康的)

相対リスクによる評価(高血圧)

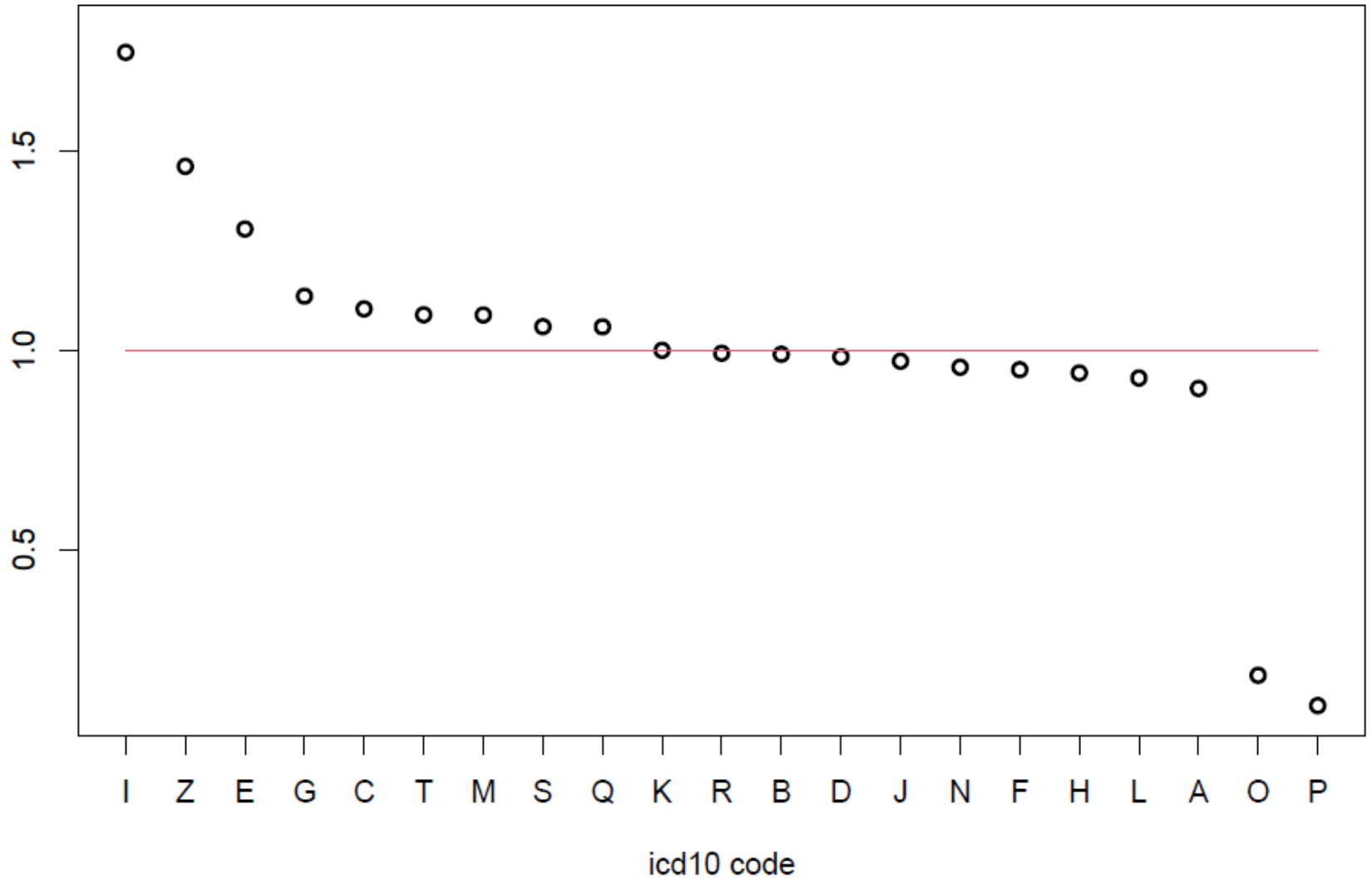
高血圧：
上の血圧が140以上，又は
下の血圧が90以上である病気

高血圧を危険因子としたときの
各傷病の相対リスク$RR_{高血圧}$

$$RR_{高血圧} = \frac{Pr[A|高血圧]}{Pr[A|正常域血圧]}$$

傷病Ⅰの$RR_{高血圧} = 1.77$
→高血圧の患者は，そうでない患者の
1.77倍傷病Ⅰにかかりやすい

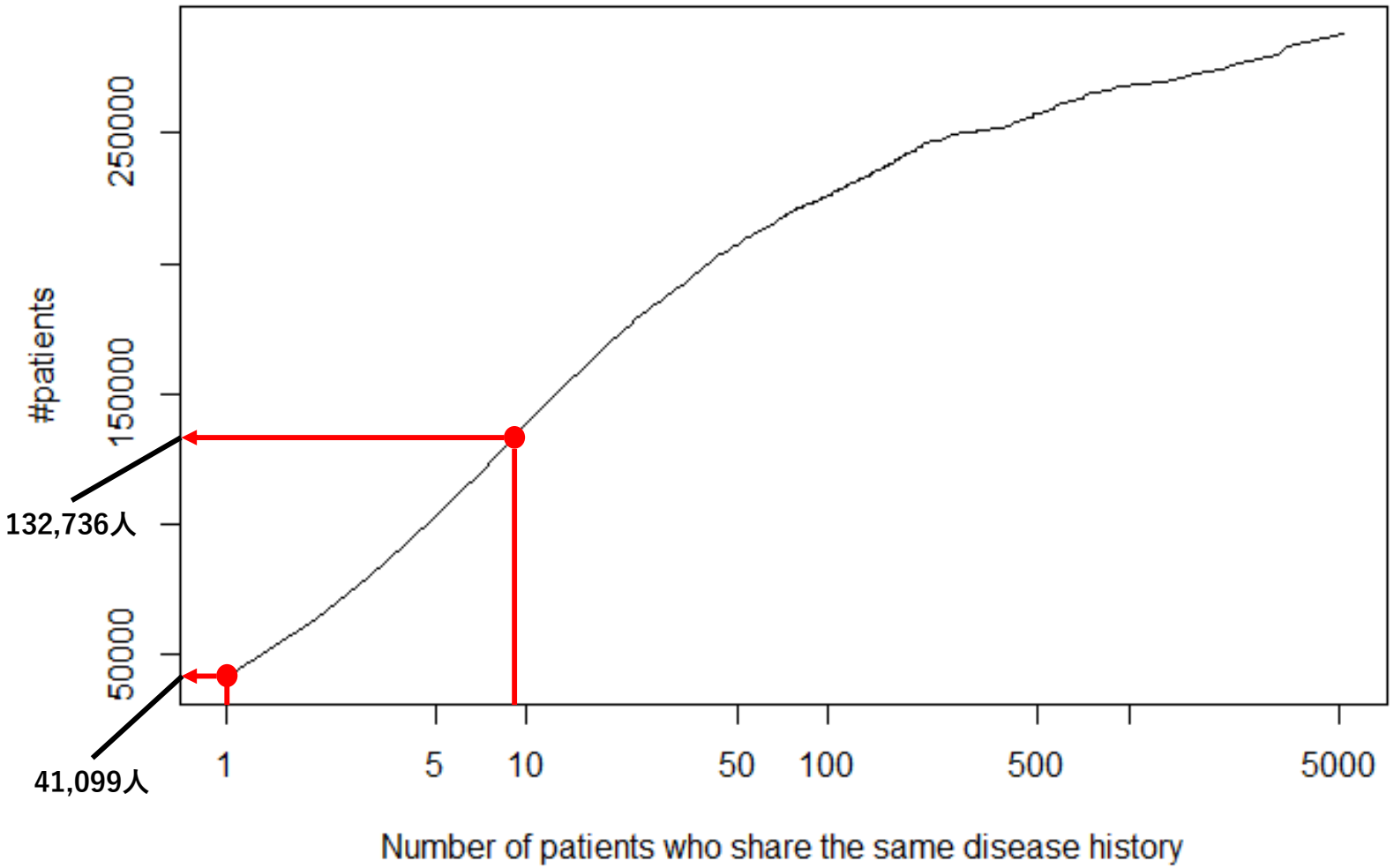
傷病Ⅰ：循環器系の疾患（脳梗塞など）



病歴の一意性

RQ3: 患者の病歴/処方歴や健康診断結果はどれくらい一意であるか？

各病歴に当てはまる患者数の累積分布



病歴の例

患者ID	傷病A	傷病B	傷病C
100	1	0	1
200	1	1	0
300	1	1	0

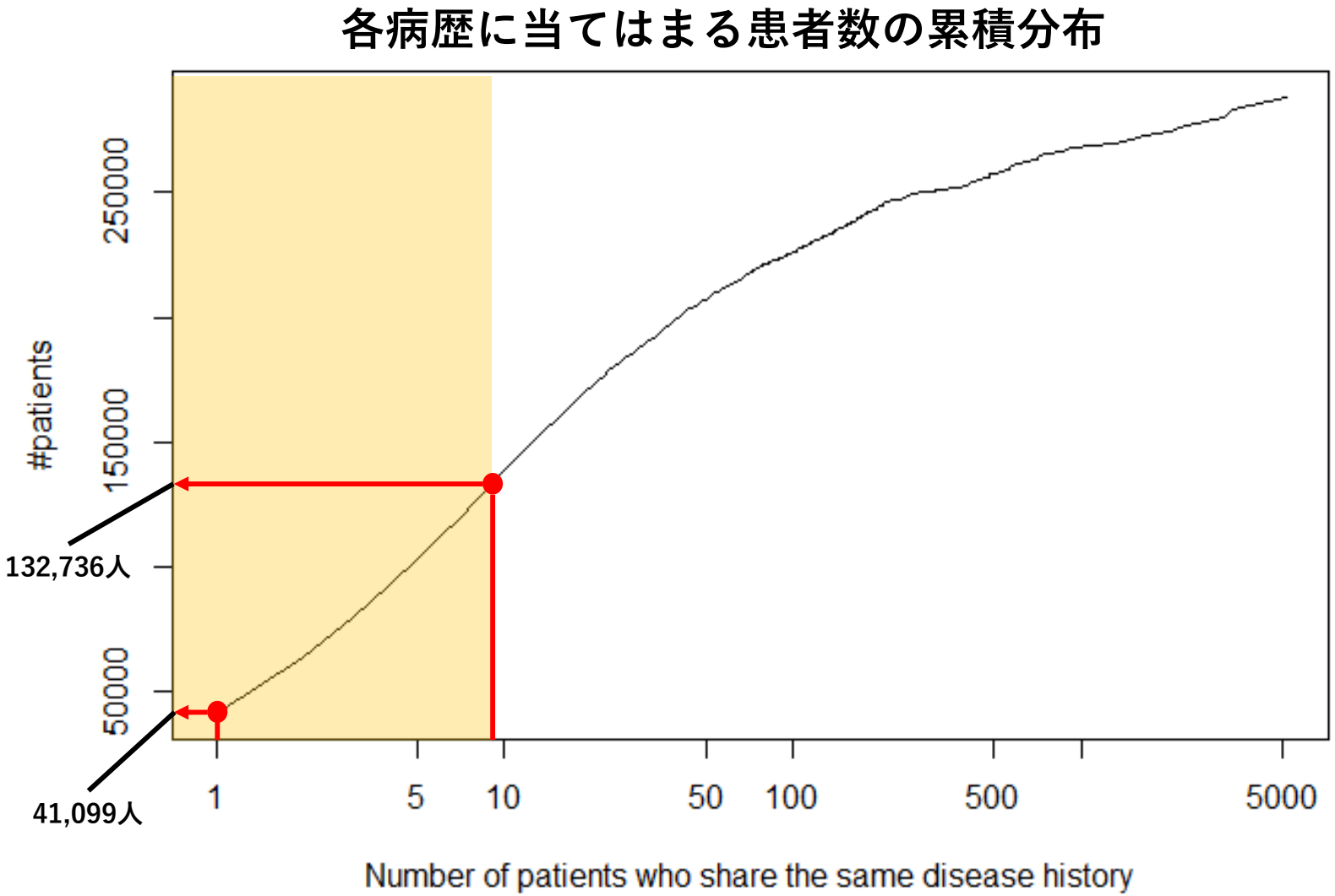
41,099人の病歴が一意である

当てはまる患者が10人未満である
病歴を持つ患者が132,736人いる

最大知識攻撃者によって
無加工の病歴から識別される
患者の人数の期待値は
71,864人である(全体の24.9%)

病歴のk-匿名化

RQ4: 病歴をk-匿名化することによって，
データの安全性・有用性はどう変化するか？

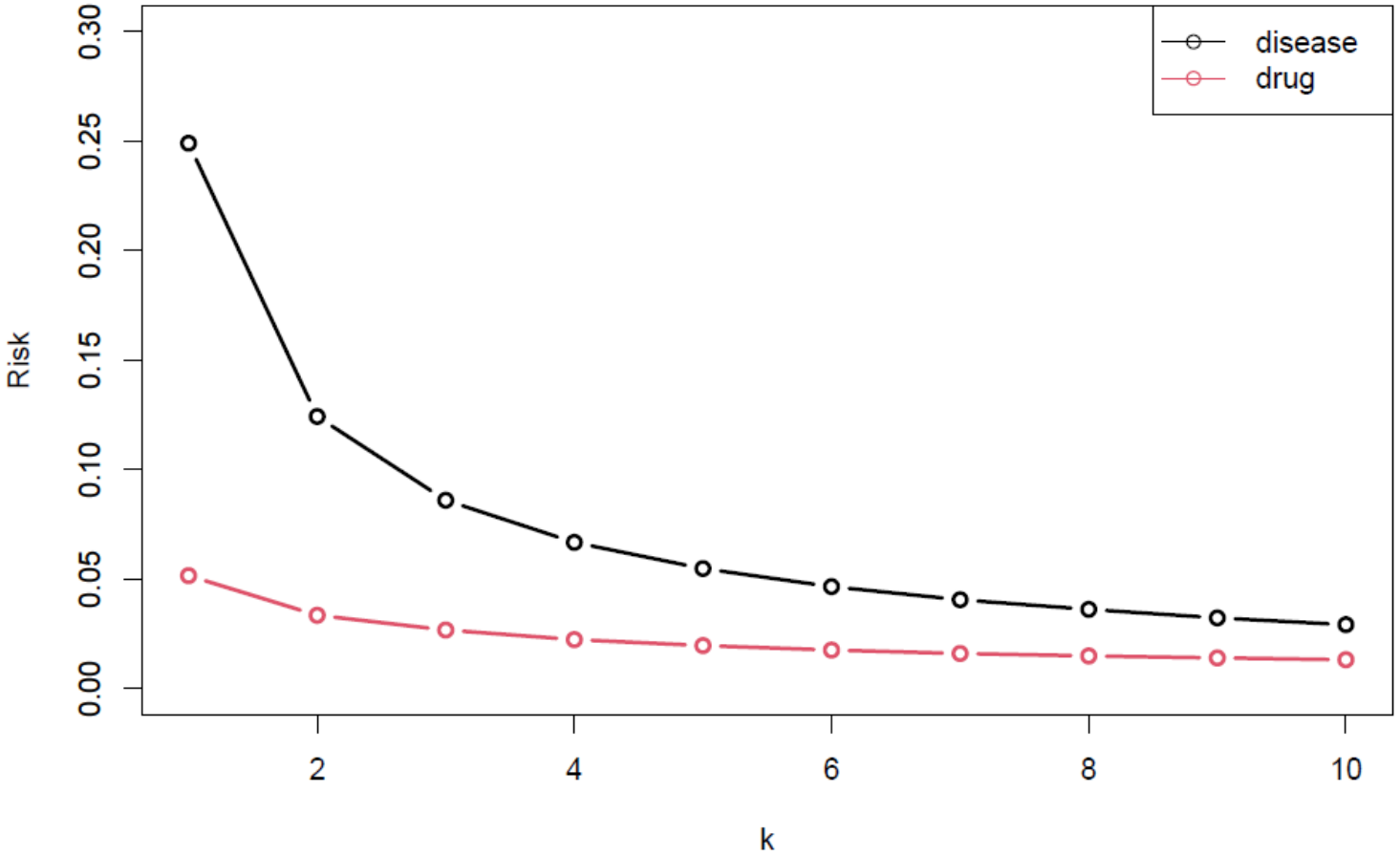


k-匿名化：
データ中の少なくともk人が同じ値を持つように加工すること

当てはまる患者がk人未満である病歴を持つ患者を全て削除すれば病歴をk-匿名化できる

当てはまる患者が10人未満である病歴を持つ患者132,736人を削除すれば，10-匿名化ができる

病歴のk-匿名化による安全性(識別割合)の変化



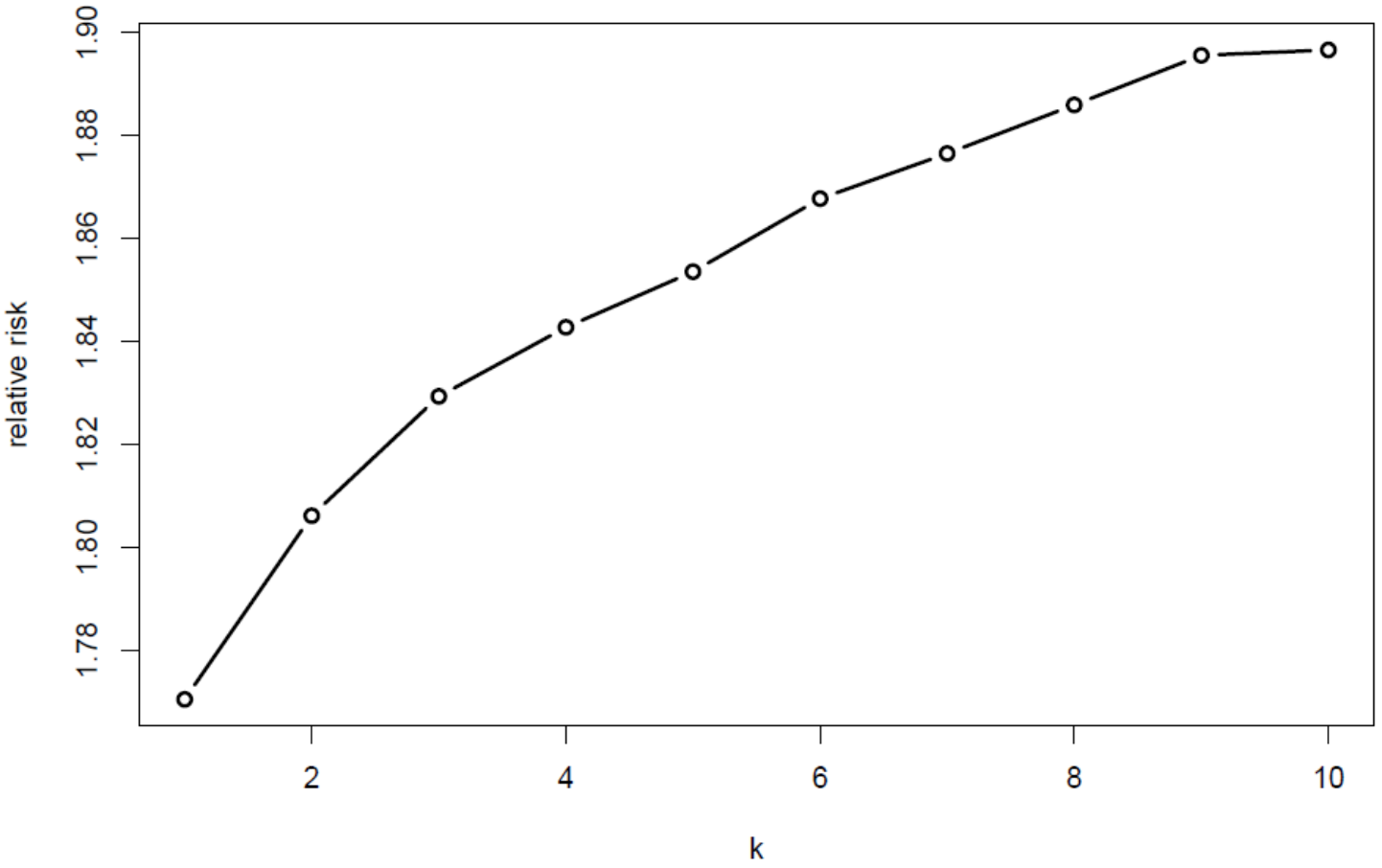
病歴/処方歴を1~10-匿名化した際の識別される患者割合の変化を求めた

黒線：病歴，赤線：処方歴

無加工の病歴からは24.9%の患者が識別されるが，10-匿名化することで識別割合を2.9%まで下げられる

（処方歴は 5.2% → 1.3% と下がる）

病歴のk-匿名化による有用性($RR_{高血圧}$)の変化



病歴を1~10-匿名化した際の
傷病Ⅰの有用性($RR_{高血圧}$)の変化を求めた

無加工の病歴では傷病Ⅰの$RR_{高血圧}$が
1.77であったが，10-匿名化された病歴では
傷病Ⅰの$RR_{高血圧}$が1.90に変化する

相対誤差は(1.90-1.77)/1.77=0.073

リサーチクエスチョンの答え

1. データ中で特異なふるまいをしている記録はあるか？

   →レセプトデータでレコード数が飛びぬけて多い患者9名や，
   健康診断データで一意な値（最大262種類，平均22種類）を持つ患者が存在した

2. 健康診断結果でこれから罹患する病気を予測できるか？

   →傷病グループごとに血圧や健康分布の値に差があった
   高血圧を危険因子とした傷病Ⅰの相対リスクは1.77であった

3. 患者の病歴/処方歴や健康診断結果はどれくらい一意であるか？

   →41,099人の病歴が一意であり，最大知識攻撃者によって無加工の病歴から識別される患者の
   人数の期待値は71,864人（全体の24.9%）であった

4. 病歴をk-匿名化することによって，データの安全性・有用性はどう変化するか？

   →10-匿名化することによって識別率を2.9%まで下げることができた
   傷病Ⅰの相対リスクは7.3%変化した

まとめ

- 健康診断は非常に有用なデータであり，分析することによって
  個人がこれから罹患する病気を予測できる可能性がある

- 個人情報保護法の改正により，健康診断結果や病歴などは要配慮情報に
  分類され，利活用の際に特別な措置が必要になった

- 本稿では，匿名加工された健康診断データとレセプトデータを分析し，
  以下の4つのリサーチクエスチョンに答えた

1. データ中で特異なふるまいをしている記録はあるか？
2. 健康診断結果でこれから罹患する病気を予測できるか？
3. 患者の病歴/処方歴や健康診断結果はどれくらい一意であるか？
4. 病歴をk-匿名化することによって，データの安全性・有用性は
   どう変化するか？