NDB・DPCデータベース研究人材育成Webinar

NDB・DPC等の公的データベースの動向

東京大学大学院医学系研究科臨床疫学・経済学
牧戸 香詠子

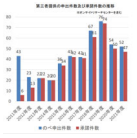
NDB

- 第三者提供は2011年から開始
- 2020年10月より高齢者の医療の確保に関する法律に第三者提供等を規定
- 提供するデータは特別抽出、サンプリングデータセット、集計表情報

https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/reseputo/index.html

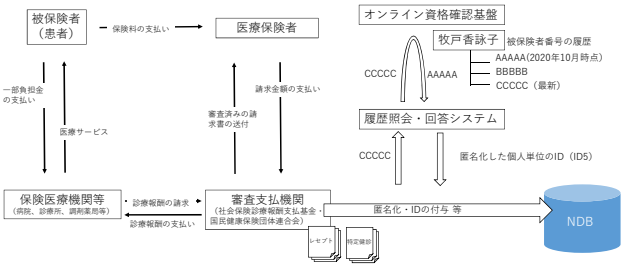
特徴

- 全国の病院・診療所のレセプトデータおよび特定健診情報
- レセプトデータは月1回（2009年4月診療分～最新：3~6ヶ月前）、
  特定健診情報は年1回（2008年度～最新：2年前） DBへ格納
- 第三者提供の実績は多い
- オンサイトリサーチセンターの設置がある
- オープンデータを公表（年1回）
（2013年度～特定健診情報・2014年4月診療分～レセプト情報）
- データ構造が複雑



https://www.mhlw.go.jp/content/12400000/000947953.pdf

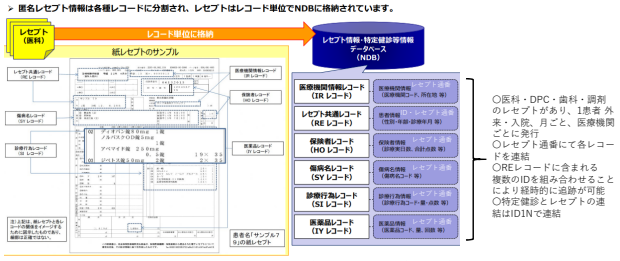
レセプト等がNDBに格納されるまでの流れ



格納されているデータ

- 被用者保険（協会けんぽ、健康保険組合、共済組合等）
- 国民健康保険
- 後期高齢者医療制度
- 医療扶助
- 特定健診情報（40歳～75歳）
  - ×紙レセプト
  - ×自費診療
  - ×労災保険、自賠責保険
  - ×妊娠・出産（保険診療外）
  - ×事業主健診　等
- その他（2022年4月～所得階層情報・住所情報が格納）

格納されているデータ



https://www.mhlw.go.jp/content/12401000/000557476.pdf

格納されているデータ（抽出テンプレート）

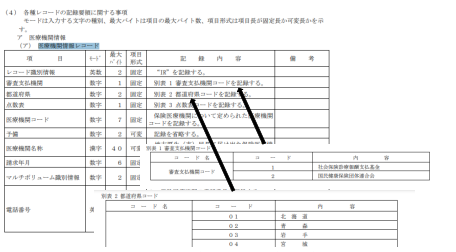
【別添8：抽出テンプレート】

・共通の項目（キー情報）について
以下の項目を使用することで、1つのレセプトを特定する、患者を特定する事が可能です。

項目名	内容
レコード順序	1レセプト内での並び順、名行をソートする際に使用
レセプト通番	全レセプトの中で一意となる通番　（特定健診・保健指導では個票単位）

格納されているデータ（記録条件仕様）

例：医科レセプト
（IRレコード）



https://www.ssk.or.jp/seikyushiharai/rezept/iryokikan/iryokikan_02.html

提供されるデータ

- 特別抽出：個票
- サンプリングデータセット：
  - 医科入院・DPCは10%、医科入院外・調剤は1%抽出
  - 単月（2011年～1・4・7・10月）
  - 探索的研究に対応
  - IDは格納されていないため、経時的な追跡は不能
- 集計表情報：
  - 複雑な集計（IDで追跡した後の集計等）は原則お断り

提供されるデータ(ID)

- 匿名化IDの作成はハッシュ関数を使用

	作成の元となる情報	特徴
ID1	保険者番号、被保険者証の記号・番号、生年月日、性別	保険者の変更や誤記により紐づけができなくなる可能性がある
ID1N		
ID2	漢字氏名・生年月日・性別	氏名の変更や誤記により紐づけができなくなる可能性がある
ID4	カナ氏名・生年月日・性別	
ID5	2020年10月時点の被保険者証の記号・番号・枝番	上記の課題に対応

https://www.mhlw.go.jp/content/12401000/000557476.pdf

オンサイトリサーチセンター

- 厚生労働省・東大・京大に設置
- NDBデータのほぼ全データを閲覧できる
- 利用形態は2種類
  - 研究者がオンサイトリサーチセンターにいき、全解析を完了する
  - 中間生成物の持ち出しを行い、自施設で解析する
- セキュリティ要件はかなり厳格
  - スマホ、PC等電子媒体の持ち込み禁止
  - 監視カメラの設置がある　等
- 外部委託はできない
- PostgreSQL・R・SPSS・STATAが利用可能

死亡情報の収載

- 死亡情報は転帰区分等にしかなく、自殺や事故等で死亡した場合、死亡したかどうかは不明
- 人口動態調査票（ ≒ 死亡診断書（死体検案書） ＋ 死亡届）と同一の収集ルートで市町村に提供を求める
- 下記の内容を検討のうえ、収載
  - 死亡年月日
  - 死亡したところの種別
  - 死亡の原因
  - 死亡の種類　等
- 2023年度分からの死亡情報を2024年以降に収載予定
- DPCDB、介護DBはNDBと連結申出のみ死亡情報が提供される

厚生労働省が第三者提供する公的データベース

- NDB（高齢者の医療の確保に関する法律）
- DPCDB（健康保険法）
- 介護DB（介護保険法）

連結申出が可能

- 難病・小児慢性疾患DB
- がん登録DB（がん登録推進法）
- 人口動態死亡票（統計法）　等

今後、提供もしくは連結を予定しているデータベース

- 難病・小児慢性疾患DB
- がん登録DB（がん登録推進法）
- 感染症DB
- 予防接種DB
- 障害福祉DB
- 次世代DB（次世代医療基盤法）

※次世代DBは認定事業者がDBを保有する民間データベース

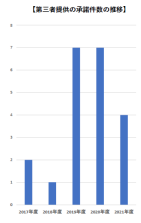
DPCDB

- 第三者提供は2017年から開始（集計表情報のみ）
- 2022年度から個票の提供を開始

https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/dpc/index.html

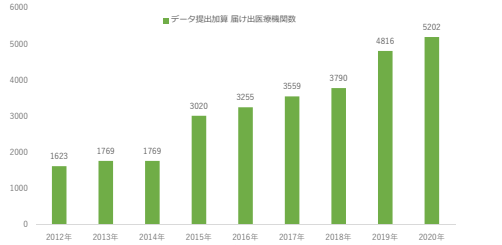
特徴

- DPC病院等から提出されたデータは年１回DBへ格納
  （2011年度実施分～最新のデータ：2年前）
- 第三者提供の実績はNDBと比較すると少ない
- オンサイトリサーチセンターの設置や
  オープンデータの公表はない
- データ構造はNDBほど複雑ではない
- DPC病院等は年々増加傾向



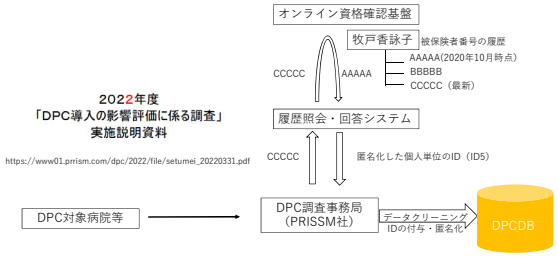
https://www.mhlw.go.jp/content/12301000/000952170.pdf

DPC病院数の推移



https://www.mhlw.go.jp/content/12400000/000833066.pdf

DPCデータ格納の流れ



格納されているデータ

様式名	内容	入力される情報
様式１	患者属性や病態等の情報	性別、生年月日、病名、病期分類など
様式４	医科保険診療以外の診療情報	保険診療以外（公費、先進医療等）の実施状況
Dファイル	診断群分類点数表に基づく診療報酬算定情報	包括レセプトの情報
入院EF統合ファイル	医科点数表に基づく診療報酬算定情報	入院診療患者の医科点数表に基づく出来高情報
外来EF統合ファイル	外来患者の医科点数表に基づく診療報酬算定情報	外来診療患者の医科点数表に基づく出来高情報
Hファイル	日ごとの患者情報	重症度、医療・看護必要度
様式３	施設情報（施設ごとに作成）	入院基本料等の届け出状況
Kファイル	３情報から生成した一次共通IDに関する情報	生年月日、カナ氏名、性別から生成した一次共通ID

詳細な項目の内容は　「DPC導入の影響評価に係る調査」実施説明資料　を参考
https://www.mhlw.go.jp/content/12400000/000863369.pdf

提供されるデータ

- 留意が必要とされるデータ項目は下記のように変換
  例：生年月日→生年月へ変換
  　　患者居住地情報→2次医療圏単位への変換
- 施設コード、保険者番号、医師コードに関しては専門医委員会で特に認める場合を除き、原則として提供対象外

※施設コード ＝ 都道府県番号 ＋ 医療機関コード

DPCデータ内での連結

ファイル名	内容	入力される内容	施設コード	データ識別番号	入院年月日
様式１	患者属性や病態等の情報	性別、生年月日、病名、病期分類など	○	○	○
様式３	施設情報	入院基本料等の届出状況	○	×	×
様式４	医科保険診療以外の診療情報	医科保険診療以外（公費、先進医療等）の実施状況	○	○	○
入院ＥＦ統合ファイル	医科点数表に基づく診療報酬算定情報	入院の出来高レセプト	○	○	○
外来ＥＦ統合ファイル	外来医科点数表に基づく診療報酬算定情報	外来の出来高レセプト	○	○	×
Ｄファイル	診断群分類点数表に基づく診療報酬算定情報	DPCレセプト	○	○	○
Ｈファイル	日ごとの患者情報	重症度、医療・看護必要度	○	○	○
Ｋファイル	３情報から生成した一次共通ＩＤに関する情報	患者の生年月日、カナ氏名及び性別から生成した一次共通ＩＤ及び被保険者番号等	○	○	○

※１「施設コード」＝　都道府県コード（2桁）　＋　保険医療機関コード（7桁）
※２「データ識別番号」＝　同内で利用する患者IDやカルテ番号と連結可能な匿名化番号（10桁）

施設コードとデータ識別番号から各ファイルを連結するキー情報を申出ごとに作成し、提供
もしくは、結合して提供も可能　https://www.mhlw.go.jp/content/12301000/000995129.pdf

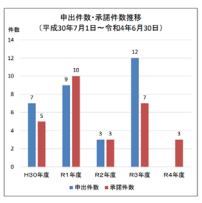
介護DB

- 第三者提供は2018年から開始
- 提供するデータは特別抽出、サンプリングデータセット、集計表情報

https://www.mhlw.go.jp/stf/shingi2/0000198094_00033.html

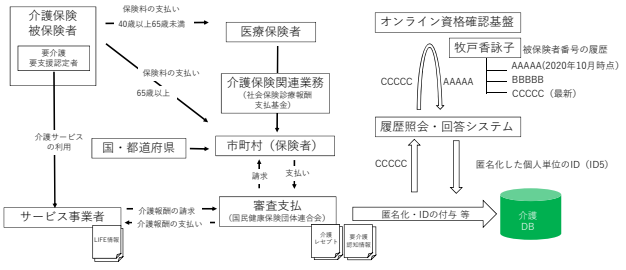
特徴

- 介護事業所から収集された介護レセプト情報・匿名LIFE情報および市町村から収集された要介護認定情報
- 介護保険受給者台帳情報（ID等格納）されているため、経時的に経過を追うことが可能
- 保険者等から提出されたデータは月１回、DBへの格納
  　要介護認定情報：2009年4月分～
  　介護レセプト情報：2012年4月分～
  　LIFE情報：2021年4月～
- 第三者提供の実績はNDBと比較すると少ない
- オンサイトリサーチセンターの設置はない
- 2022年11月からオープンデータの公表を開始
  　（2018年度実施分～年1回秋ごろ）
- テーブル定義書がある



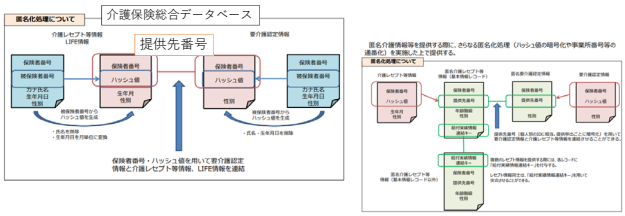
https://www.mhlw.go.jp/content/12301000/000995136.pdf

介護レセプト等が介護DBに格納されるまでの流れ



格納されているデータ

• 介護保険総合データベースは2013年から運用開始



https://www.mhlw.go.jp/content/12301000/000919384.pdf

格納されているデータ（テーブル定義書）

https://www.mhlw.go.jp/content/12301000/000343808.pdf

格納されているデータ（匿名LIFE情報）

〇科学的介護情報システム（Long term care Information system For Evidence：LIFE）の略

〇格納されている主なデータ
利用者の状態・ケアの内容等の情報
・利用者情報
・科学的介護推進情報（アセスメント結果、既往歴情報等 等）

提供されるデータ

• 特別抽出：個票
• サンプリングデータセット：
　介護レセプト等情報のうち、基本情報・明細情報・居宅サービス計画費情報レコード
　2012年度以降の各年 4 月、 7 月、 10 月、 1 月サービス提供分のデータから一定の割合で抽出
　利用の少ないサービス項目コード等について匿名化処理
　IDは格納されていないため、経時的な追跡は不能
• 集計表情報

NDB・DPC・介護DBの共通の特徴

• 申請書類やガイドラインは類似している
• 連結申出の場合は申請書類等は、連結するデータ分必要
• 専門委員会（年4回開催：6・9・12・3月）での審査にて承諾が必要　※連結の場合は合同委員会
• 連結・追跡するための個人単位のIDがある
• 手数料がかかる　※公的機関や厚労科研費等の取得があれば免除
　DPC　4,250円/時間
　NDB　7,700円/時間
　介護DB　5,900円/時間

NDB・DPC・介護DBの共通の特徴

• ガイドライン記載の情報セキュリティ要件を遵守する必要
物理的・技術的安全管理措置

入退室管理	利用場所の施錠、取扱者の名札等の着用、台帳等による入退室管理及び入退室の記録を定期的にチェックし、その妥当性の確認、入退室管理の保管（1年）
データ保護	クリアスクリーン等による窃視防止
	匿名レセプト情報等、中間生成物等が格納された記録媒体の管理
盗難・紛失	専用端末の窃盗防止用ワイヤー等設置による盗難防止
認証・識別	ＭＦＡ（二要素認証）による利用者識別の推奨
ウィルス対策	外部からの情報受領時には、不正なウィルスの混入を防止
消去	利用終了後は専用ソフトウェア等を利用し、復元不可能な形で消去
ログ管理	アクセスの記録及び定期的なログの確認・保管

• 実地監査（対象は主に特別抽出の申出）

https://www.mhlw.go.jp/content/12301000/000952167.pdf

NDB・DPC・介護DBの共通の特徴

• 組織的安全管理措置
a）利用者及び取扱者の権限、責務及び業務を明確にすること。
b）運用管理規程を定めること。

• 人的安全管理措置
a）取扱者は、高齢者の医療の確保法等又はこれらの法律に基づく命令の規定に違反し、罰金以上の刑に処せられ、その執行を終わり、又は執行を受けることがなくなった日から起算して５年を経過しないこと
b）提供申出者（もしくは利用者）は取扱者に対し、匿名レセプト情報等を取り扱う上で必要な教育及び訓練を行うこと。

NDB・DPC・介護DBの共通の特徴

不適切利用の内容
他の情報と照合を行う
利用期間終了以降にレセプト情報等の返却・消去を行わない
情報セキュリティ上の危険にさらした
提供を受けた匿名レセプト情報等を紛失・漏洩した
申出書に記載された目的以外の利用
公表物確認をせずに公表した ※公表物確認：学会発表や論文投稿等をする前の事前の厚生労働省の確認
その他

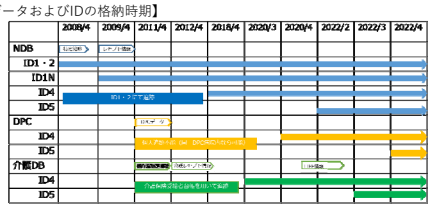
高齢者の医療の確保に関する法律等により、罰則規定があり、
また提供申出者（法人等）に罰則がかかるため、充分注意が必要です！

NDB・DPC・介護DBの追跡・連結

【連結・追跡するためのID】

ID4	3情報（カナ氏名、生年月日、性別）から作成し、匿名化したID
ID5	被保険者番号（2020年10月時点）を匿名化したID

【各データおよびIDの格納時期】



専門委員会での審査ポイント

• 相当の公益性があることが必要
• 実現可能性があるかどうか
• 研究を実施する上で「必要最小限の範囲」となっているか
　探索的研究となっていないか
　複数の研究が1申出に盛り込まれていないか
　研究内容を鑑みて、不必要な項目を抽出していないか
• 個人の識別可能性が上がる場合は慎重な審査を行う
• 研究者や所属施設、研究施設が複数にまたがる申出は、セキュリティ対策実践の難易度が上がるため、その対応方法について慎重な審査を行う

NDBの今後の動向

• 医療介護データ等の解析基盤：HIC（Healthcare Intelligence Cloud）
　研究者がクラウド上でNDB等を解析できる基盤の開発

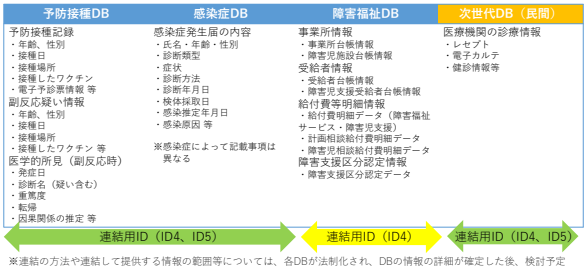
HICの機能		想定される研究者
ポータル機能	NDB・介護DB等の提供申出、利用及び終了に至る一連の手続き 教育・啓発のためのコンテンツの形成や各種マスターの共有 等	全研究者
探索的利用環境	ダミーデータを用いて探索・試行的に分析するための環境	これからNDB研究を始めようと考えている申出者
HIC解析環境	専門委員会の審査にて承認された提供申出ごとに、利用者に対して解析環境	承認を受けた申出者

• 施行的利用後、本格運用予定

他の公的データベース

難病・小児慢性疾患DB	がん登録DB
2019年度～患者の同意に基づく提供	2018年度～
• 臨床調査個人票（難病）・医療意見書（小慢）の情報（2015年1月～審査会の申出締切日前月末まで） • 研究用ID（確率的ID）の提供により経時的に追跡が可能 • 連結用ID（ID4、ID5）の取得予定	• 原発癌の初回の診断・診療および死亡情報 • 利用形態はリンケージ利用（調査研究対象者への説明と同意の取得が必要）および集計統計利用
• 指定難病患者データ及び小児慢性特定疾病児童等データの提供に関するワーキンググループ（審査会）にて個別審査	• リンケージ利用は厚生科学審議会にて審査 • 集計登録利用は国立がん研究センター全国がん登録情報提供等審議委員会にて審査 ※都道府県がん情報は都道府県が開催する審議会等にて審査
• 提供依頼申出者の範囲は、 厚労省・文科省が補助を行う研究事業 都道府県・指定都市・中核市 その他	提供依頼申出者の範囲は、 研究者 国・都道府県の関係者 ※民間事業者は除外されていない
• 申請書、ガイドラインはNDB（DPCDB・介護DB）をもとに作成されており、類似点が多い	• 遵守すべき安全管理措置等はガイドラインにて規定
手数料の徴収はない	手数料の徴収あり

他の公的・民間データベース



予防接種DB	感染症DB	障害福祉DB	次世代DB（民間）
予防接種記録 ・年齢、性別 ・接種日 ・接種場所 ・接種したワクチン ・電子予診票情報 等 副反応疑い情報 ・年齢、性別 ・接種日 ・接種場所 ・接種したワクチン 等 医学的所見（副反応時） ・発症日 ・診断名（疑い含む） ・重篤度 ・転帰 ・因果関係の推定 等	感染症発生届の内容 ・氏名・年齢・性別 ・診断類型 ・症状 ・診断方法 ・診断年月日 ・検体採取日 ・感染推定年月日 ・感染原因 等 ※感染症によって記載事項は異なる	事業所情報 ・事業所台帳情報 ・障害児施設台帳情報 受給者情報 ・受給者台帳情報 ・障害児支援受給者台帳情報 給付費等明細情報 ・給付費明細データ（障害福祉サービス・障害児支援） ・計画相談給付費明細データ ・障害児相談給付費明細データ 障害支援区分認定情報 ・障害支援区分認定データ	医療機関の診療情報 ・レセプト ・電子カルテ ・健診情報等

連結用ID（ID4、ID5）　連結用ID（ID4）　連結用ID（ID4、ID5）

※連結の方法や連結して提供する情報の範囲等については、各DBが法制化され、DBの情報の詳細が確定した後、検討予定

NDB eLearning 教材 1

NDBガイドラインの理解

次の各問に当てはまる選択肢を 1 つ選べ。

１．NDB 活用の本来目的を以下から 1 つ選べ。
① 医薬品・医療機器の市販後調査
② 臨床研究・疫学研究の推進
③ 医療安全管理
④ 医療費適正化
⑤ 医療サービスの情報公開

２．匿名レセプト情報・匿名特定健診等情報の提供は、以下のどの法律に基づいているか？
① 医療法
② 医師法
③ 地域における医療及び介護の総合的な確保を推進するための関係法律の整備に関する法律
④ 高齢者の医療の確保に関する法律
⑤ 医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律

３．匿名レセプト情報・匿名特定健診等情報のデータ保有者はだれか？
① 総務大臣
② 厚生労働大臣
③ 都道府県知事
④ 保険者
⑤ 支払基金

４．有識者から構成される「匿名医療情報等の提供に関する専門委員会」は以下のいずれの下に設けられているか？
① 社会保障審議会

② 厚生科学審議会
③ 社会保険審査会
④ 中央社会保険医療協議会
⑤ 医薬品等行政評価・監視委員会

５．第三者申出件数は 2018 年度以降、年間 X 件を超えている。X に入る適当な数値を以下から 1 つ選べ。
① ５０
② １００
③ ３００
④ ５００
⑤ １０００

６．サンプリングデータセットを除き、提供申出が承諾された日から NDB データが提供されるまでに要した平均所要日数は、2018-2019 年度においては X 日以内であった。X に入る適当な数値を以下から 1 つ選べ。
① ３０
② ９０
③ １８０
④ ３００

７．①―④誤っているものを 1 つ選べ。すべて正しい場合は⑤を選べ。
① 「中間生成物」とは、匿名レセプト情報等を提供したのち取扱者が生成したものであって、最終生成物と成果物以外のものをいう。
② 「中間生成物」については、厚生労働省による公表前の事前の承認を得て成果物となったものを除き、取扱者以外に公表することは禁じれられている。
③ 本ガイドラインにおいて「最終生成物」とは、匿名レセプト情報等を提供したのち取扱者が最終的に生成したものであって、厚生労働省による公表前の事前の承認を得ていないものをいう。
④ 「最終生成物」については、厚生労働省による公表前の事前の承認を得て成果物となったものも含めて、取扱者以外に公表することができる。

８．①―④誤っているものを 1 つ選べ。すべて正しい場合は⑤を選べ。

① 提供申出者及び取扱者は、提供された匿名レセプト情報等について、個人情報の保護に関する法律に規定する個人情報に準じた取扱いを行うこととする。
② 提供申出者及び取扱者は、医療情報システムの安全管理に関するガイドラインに定められた措置に準じた措置を匿名レセプト情報等の利用形態を勘案した上で適切に講ずるものとする。
③ 提供申出者は、外部委託を行う必要性が、研究の目的及び内容に照らして合理的である場合、匿名レセプト情報等を用いた研究を外部委託することができる。
④ オンサイトリサーチセンター内での作業については外部委託することは認められない。

９．厚生労働省は、匿名レセプト情報等の提供により、提供申出者、取扱者及び第三者に患者等の情報が特定されることがないよう、各提供申出書の内容に応じて、専門委員会における議論及び技術的な問題等を勘案し、提供する匿名レセプト情報等に適切な処理を施すものとし、処理を講じた場合には、その内容を提供申出者及び取扱者に明示するものとする、とされている。「適切な処理」に該当しないものを以下から１つ選べ。
① 第三者に情報が特定される可能性のある特定個人の全データ削除
② データの再ソート（配列順の並べ替え）
③ 特定個人又は特定機関の識別情報のトップ（ボトム）・コーディング
④ 特定個人又は特定機関の識別情報のグルーピング（リコーディング）
⑤ リサンプリング

１０．①―④誤っているものを１つ選べ。すべて正しい場合は⑤を選べ。
① 医療機関・薬局コードについては、専門委員会が特に認める場合を除き、原則として提供しないこととされている。
② 保険者番号については、専門委員会が特に認める場合を除き、原則として提供しないこととされている。
③ 技術的な問題等により適切な処理が行い難い場合には、専門委員会の議論を経て、匿名レセプト情報等の提供を行わない場合もあり得る。
④ 厚生労働省は、提供する匿名レセプト情報等について利用方法や情報の範囲等を勘案し、専門委員会の意見を聴取した上で適切な処理を行うこととする。

１１．①―④誤っているものを１つ選べ。すべて正しい場合は⑤を選べ。
① 提供申出書は、匿名レセプト情報等の提供の判断要件となる「利用目的」ごとに作成するものとする。

② 提供申出者が実施する複数の研究に用いる匿名レセプト情報等について併せて提供申出を行って差し支えない。
③ 提供された匿名レセプト情報等１ファイルについて、当該ファイルを別の記憶装置に複写・保存する行為は２回に限定する。
④ 当該記憶装置の保存・複製ファイルが消去されない限り、別の記憶装置への保存・複写は原則として認めない。

１２．匿名レセプト情報等の提供申出者の範囲について、①―④のうち該当しないものはどれか？すべて該当する場合は⑤を選べ。
① 公的機関（国の行政機関、都道府県、市区町村）
② 大学その他の研究機関
③ 研究開発独立行政法人等
④ 民間事業者等

１３．①―④誤っているものを1つ選べ。すべて正しい場合は⑤を選べ。
① 法人（公的機関を除く法人その他の団体で代表者又は管理人の定めがあるもの）は、原則として登記された法人等を単位として提供申出を行う。
② 公的機関が開設する医療機関の場合、当該医療機関を開設する公的機関を単位として提供申出を行う。
③ 国立病院機構及び労働者健康安全機構が開設する医療機関の場合、当該医療機関を単位として提供申出を行う。
④ 大学病院の場合、当該大学病院を単位として提供申出を行う。

１４．①―③のうち、誤っているものを１つ選べ。すべて正しい場合は④を選べ。
① 提供申出者が公的機関の場合、担当者の身分証明書等の写しを提出する。
② 提供申出者が法人等の場合、担当者の身分証明書等の写しを提出する。
③ 提供申出者が個人の場合、提供申出者の身分証明書等の写しを提出する。

１５．①－④のうち、誤っているものを１つ選べ。すべて正しい場合は⑤を選べ。
① 利用目的が特定の商品又は役務の広告又は宣伝に直接利用する又は利用されると推測される場合、提供は認められない。
② 匿名レセプト情報等の直接的な利用目的が、企業等の組織内部に企業等の組織内部における業務上の資料として利用される場合、提供は認められない。
③ 匿名レセプト情報等の直接的な利用目的が、企業等の特定の顧客に対するレポート作成の基礎資料とされるような場合、提供は認められない。

④　他の研究や政策利用等を阻害するような場合でも、特許の取得は認められる。

１６．①－④のうち、ガイドラインが定める研究の内容に該当するものを１つ選べ。すべて該当場合は⑤を選べ。

①　医療分野の研究開発に資する分析
②　適正な保健医療サービスの提供に資する施策の企画及び立案に関する調査
③　疾病の原因並びに疾病の予防、診断及び治療の方法に関する研究
④　保健医療の経済性、効率性及び有効性に関する研究

１７．①－⑤のうち、誤っているものを１つ選べ。すべて正しい場合は⑥を選べ。

①　匿名レセプト情報等を他の情報と照合してはならない。
②　提供申出を行う匿名レセプト情報等が研究内容に鑑みて最小限であるとする根拠を記入する。
③　匿名レセプト情報等を実際に利用する場所は日本国内に限る。
④　匿名レセプト情報等の利用期間の上限は、原則として、１年間とする。
⑤　匿名レセプト情報等の提供に必要な媒体（CD-R、DVD、外付けハードディスク等）は、匿名レセプト情報等の情報量等を勘案し、厚生労働省において用意する。

１８．①―④誤っているものを1つ選べ。すべて正しい場合は⑤を選べ。

①　提供申出に係る手数料は、人件費等を踏まえた時間単位の金額（１時間までごとに6100円）に、作業に要した時間を乗じて得た額とする。
②　作業に要した時間のうち申出処理業務には、申出書類確認・専門委員会への諮問手続・データの抽出条件の精査等が含まれる。
③　作業に要した時間のうちデータ抽出業務とは、SQL作成・テスト実施・結果の検証等が含まれる。
④　補助金は免除されない。

１９．①―④のうち誤っているものを1つ選べ。すべて正しい場合は⑤を選べ。

①　提供申出書等の受付窓口は、厚生労働省保険局医療介護連携政策課保険データ企画室である。
②　事務処理を円滑に行うため受付窓口を外部委託する場合がある。
③　厚生労働省は、担当者及び代理人に対して、氏名、生年月日及び住所を確認できる書類のコピーを求める。

④　受付窓口へ郵送により提出する書類は、原則として直筆の必要がある書類のみとし、その他についてはEメールでの送付を可とする。

２０．　匿名レセプト情報等を利用する必要性等の基準のうち、該当しないものはどれか？①―④から1つ選べ。すべて正しい場合は⑤を選べ。

①　利用する匿名レセプト情報等の範囲及び匿名レセプト情報等から分析する事項が研究内容から判断して必要最小限であること。

②　匿名レセプト情報等の性格に鑑みて、その利用に合理性があり、他の情報では研究目的が達成できないこと。

③　匿名レセプト情報等の利用期間と研究の計画・公表時期が整合的であること。

④　匿名レセプト情報等の利用について、申し出られている研究内容を現時点で行うことについて合理的な理由があること。

２１．①―④のうち誤っているものを1つ選べ。すべて正しい場合は⑤を選べ。

①　匿名レセプト情報等を複写した情報システムの利用場所、保管場所及び管理方法は、あらかじめ申し出られた施錠可能な物理的なスペースに限定されており、原則として持ち出してはならない。

②　匿名レセプト情報等を複写した情報システムは、インターネット等の外部ネットワークに接続してはならない。

③　提供された匿名レセプト情報等は、あらかじめ申し出られた取扱者のみが利用することとし、その他の者へ譲渡、貸与又は他の情報との交換等を行ってはならない。

④　現に匿名レセプト情報等の提供を承諾された提供申出における担当者が、当該匿名レセプト情報等の利用を終了していない場合でも、新たな提供申出を行うことができる。

２２．①―④のうち誤っているものを1つ選べ。すべて正しい場合は⑤を選べ。

①　情報システム運用責任者の設置及び担当者（システム管理者を含む。）の限定を行わなければならい。所属機関が小規模な場合においても明確な規程を定めなければならない。

②　個人情報が参照可能な場所においては、来訪者の記録・識別、入退を制限する等の入退管理を定めること。

③　情報システムへのアクセス制限、記録、点検等を定めたアクセス管理規程を作成すること。

④　個人情報の取扱いを委託する場合、委託契約において安全管理に関する条項を含めること。

２３．①―④のうち誤っているものを1つ選べ。すべて正しい場合は⑤を選べ。

①　匿名レセプト情報等が保存されている機器の設置場所及び記録媒体の保存場所には施錠しなければならない。

②　匿名レセプト情報等を物理的に保存している区画への入退管理を実施しなければならない。

③　匿名レセプト情報等の消去にあたっては、専用ソフトウェア等を用い、復元不可能な形で行わなければならない。

④　匿名レセプト情報等を利用する情報システムへのアクセスにおける取扱者の識別と認証を行わなければならない。

２４．結果の公表基準について、①―④のうち誤っているものを1つ選べ。すべて正しい場合は⑤を選べ。

①　原則として、公表される研究の成果物において患者等の数が 10 未満になる集計単位が含まれていないこと（ただし患者等の数が「０」の場合を除く。）

②　集計単位が市区町村の場合、人口 2,000 人未満の市区町村では、患者等の数が 20 未満になる集計単位が含まれないこと。

③　原則として、公表される研究の成果物において年齢区分が、5 歳毎にグルーピングして集計されていること。

④　原則として、特定健診等情報にかかる受診者の住所地については、公表される研究の成果物における最も狭い地域区分の集計単位は２次医療圏又は市区町村とすること。

＜解答と解説＞

１．正解④
NDB 活用の本来目的は医療費適用であり、それ以外は目的外利用という位置づけである。

２．正解④
匿名レセプト情報・匿名特定健診等情報の提供は「高齢者の医療の確保に関する法律」に基づいている。

「高齢者の医療の確保に関する法律」
（医療費適正化計画の作成等のための調査及び分析等）
第十六条　厚生労働大臣は、全国医療費適正化計画及び都道府県医療費適正化計画の作成、実施及び評価に資するため、次に掲げる事項に関する情報（以下「医療保険等関連情報」という。）について調査及び分析を行い、その結果を公表するものとする。
一　医療に要する費用に関する地域別、年齢別又は疾病別の状況その他の厚生労働省令で定める事項
二　医療の提供に関する地域別の病床数の推移の状況その他の厚生労働省令で定める事項
２　保険者及び後期高齢者医療広域連合は、厚生労働大臣に対し、医療保険等関連情報を、厚生労働省令で定める方法により提供しなければならない。
３　厚生労働大臣は、必要があると認めるときは、都道府県及び市町村に対し、医療保険等関連情報を、厚生労働省令で定める方法により提供するよう求めることができる。
（国民保健の向上のための匿名医療保険等関連情報の利用又は提供）
第十六条の二　厚生労働大臣は、国民保健の向上に資するため、匿名医療保険等関連情報（医療保険等関連情報に係る特定の被保険者その他の厚生労働省令で定める者（次条において「本人」という。）を識別すること及びその作成に用いる医療保険等関連情報を復元することができないようにするために厚生労働省令で定める基準に従い加工した医療保険等関連情報をいう。以下同じ。）を利用し、又は厚生労働省令で定めるところにより、次の各号に掲げる者であつて、匿名医療保険等関連情報の提供を受けて行うことについて相当の公益性を有すると認められる業務としてそれぞれ当該各号に定めるものを行うものに提供することができる。
一　国の他の行政機関及び地方公共団体　適正な保健医療サービスの提供に資する施策の企画及び立案に関する調査
二　大学その他の研究機関　疾病の原因並びに疾病の予防、診断及び治療の方法に関する研究その他の公衆衛生の向上及び増進に関する研究
三　民間事業者その他の厚生労働省令で定める者　医療分野の研究開発に資する分析その他の厚生労働省令で定める業務（特定の商品又は役務の広告又は宣伝に利用するために行

うものを除く。）

２　厚生労働大臣は、前項の規定による利用又は提供を行う場合には、当該匿名医療保険等関連情報を健康保険法第百五十条の二第一項に規定する匿名診療等関連情報及び介護保険法第百十八条の三第一項に規定する匿名介護保険等関連情報その他の厚生労働省令で定めるものと連結して利用し、又は連結して利用することができる状態で提供することができる。

３　厚生労働大臣は、第一項の規定により匿名医療保険等関連情報を提供しようとする場合には、あらかじめ、社会保障審議会の意見を聴かなければならない。

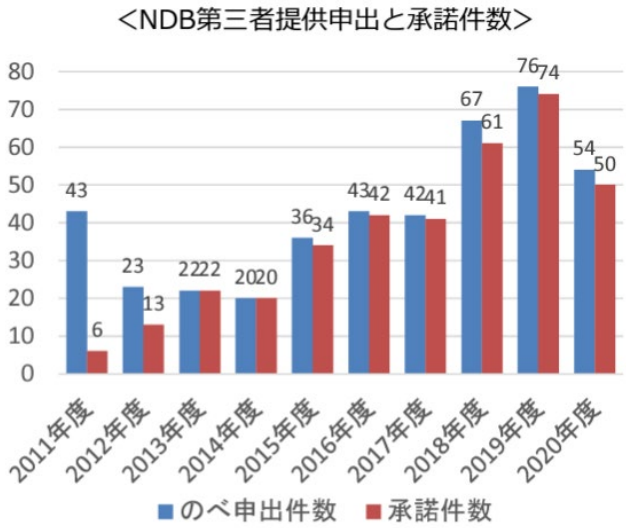
３．正解②

匿名レセプト情報・匿名特定健診等情報のデータ保有者は厚生労働大臣である。

４．正解①

「匿名医療情報等の提供に関する専門委員会」は社会保障審議会の下に設けられている。

５．正解①



６．正解④

平均所要日数は、2018 年度は 281.2 日、2019 年度は 203.1 日であった。

７．正解④

「最終生成物」については、厚生労働省による公表前の事前の承認を得て成果物となったものを除き、取扱者以外に公表することを禁ずる。

８．正解⑤

すべて正しい。

９．正解①
特定個人又は特定機関の識別情報の削除

１０．正解⑤
すべて正しい。

１１．正解③
提供された匿名レセプト情報等１ファイルについて、当該ファイルを別の記憶装置に複写・保存する行為は１回に限定する。

１２．正解⑤
NDBは2009年に稼働開始し、2011年から「有識者会議」（現在は「専門家委員会」）の審査を経て、第三者の研究者等への提供が開始された。2020年10月の法改正により、民間事業者を含めた幅広い主体への提供が可能となった。

１３．正解④
大学の学長

１４．正解②
提供申出者が法人等の場合、提供申出書の提出日前６ヶ月以内に作成された登記事項証明書等を提出する。

１５．正解④
推測されるものは認めない。また、匿名レセプト情報等の提供の制度 趣旨 は国民保健の向上に資す るといった相当の公益性を有することを求めるものであることを考慮し、他の研究や政策利用等を阻害するような特許の取得を禁止する。

１６．正解⑤
すべて正しい。

１７．正解④
匿名レセプト情報等の利用期間の上限は、原則として、２年間とする。

１８．正解④

ガイドラインに掲げる補助金等を充てて匿名レセプト情報等を用いて研究又は業務を行う者は手数料を免除される。手数料の免除を希望する場合は、補助金等の交付決定通知の写し及び研究計画書又は交付申請書を添付する。免除申請は、提供申出時から、厚生労働省が提供申出者に手数料額を通知する時までとする。

１９．正解⑤
すべて正しい。

２０.正解⑤
すべて正しい。

【審査基準】
①利用目的
レセプト情報等の利用目的は、医療サービスの質の向上等を目指した施策の推進や、学術の発展に資する研究に資するものであるか
②利用の必要性
利用するレセプト情報の範囲が利用目的に照らして必要最小限であるか、レセプト情報の性格に鑑みて情報の利用が合理的か
③研究内容の実行可能性
研究計画の内容は、申出者の過去の研究実績や人的体制に照らして実行可能であるか
④セキュリティ
適切な措置（レセプト情報等を複写した情報システムを外部ネットワークに接続しない、個人情報保護に関する方針の策定・公表、外部委託契約における安全管理条項の有無等）を講じているか
⑤結果公表等
学術論文等の形で研究成果が公表される予定か、施策の推進に適切に反映されるか等

２１．正解④
現に匿名レセプト情報等の提供を承諾された提供申出における担当者が、当該匿名レセプト情報等の利用を終了していない場合については、新たな提供申出を行うことは原則認められない。

２２．正解①
所属機関が小規模な場合において役割が自明の場合は、明確な規程を定めなくとも良い。

２３．正解⑤
すべて正しい。

２４．正解②
人口2,000人未満の市区町村では、患者等の数を表示しないこと。
人口2,000人以上25,000人未満の市区町村では、患者等の数が20未満になる集計単位が含まれないこと。
人口 25,000 人以上の市区町村では、患者等の数が 10 未満になる集計単位が含まれないこと。

NDB･DPCデータベース研究人材育成Webinar

SQL入門

東京大学大学院医学系研究科糖尿病･生活習慣予防講座
岡田 啓

SQL概論

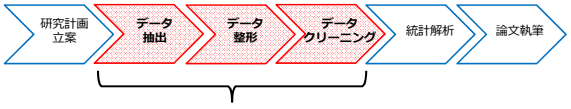
SQLをなぜ使うか？①

- データテーブルを自分で構築する
  - ・テーブルが複数あるようなデータの場合、自分で解析用のテーブルを作る必要がある
  - ・商用データベース（DeSCデータベースなど）、ナショナルレセプトデータベース(NDB)などで研究を行う際には自分で、解析のためのテーブルを作成する必要がある。
- 統計ソフトではなぜ不十分か？

SQLをなぜ使うか？②

- データベースサイズの拡大
  - ・ビッグデータが更にhugeに変遷：MB単位→GB単位のデータベースへ
- 統計ソフトではデータクリーニングが難しい
  - ・統計ソフトはメモリ上にデータを載せるので数十GBのデータだと載せるのが困難
  - ・統計ソフトでは、結合などに時間がかかり、そもそもhugeデータには不向き

データ抽出・クリーニングでSQLは有用

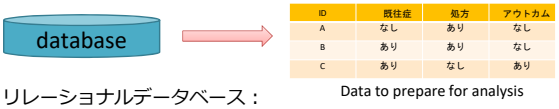


SQLが必須の場合もそうではない場合もSQLを使った方が
圧倒的に処理が早い

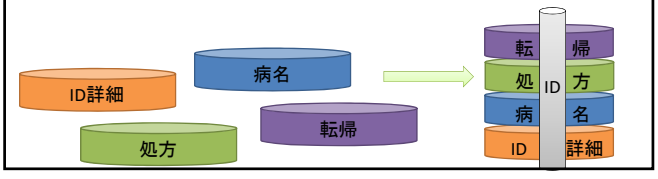
← どちらにしてもSQLを身につけるメリットは大きい

リレーショナルデータベース(Relational DataBase)

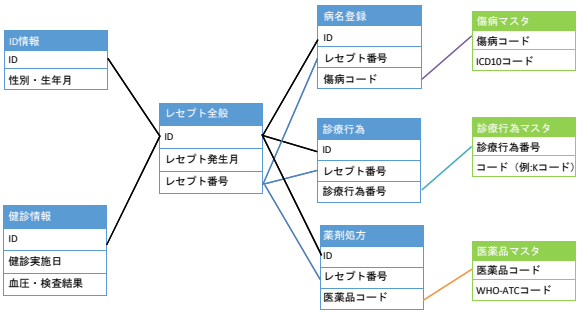
・データベース：Data + Base なので実際扱うデータの源



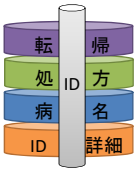
・リレーショナルデータベース：
ID/コードに関連（relational）させて複数のデータベースに情報を格納



リレーショナルデータベースの構造



リレーショナルデータベースの活用法



解析用テーブル作成

クエリにより抽出

ID	既往症	処方	アウトカム
A	なし	あり	なし
B	あり	あり	なし
C	あり	なし	あり

データの要求文[問い合わせ]
← 今回学ぶこと

統計ソフトでの解析

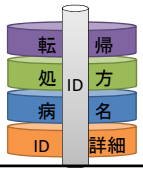
SQLのソフトウェアで格納
（リレーショナルデータベース管理システム）
- PostgreSQL
- Microsoft SQL server
- My SQL
- Oracle SQL

など

ここまでのまとめ：SQL導入

①SQLをなぜ使うか？
→ データサイズの大きいデータベースは統計ソフトでは扱うのが難しい

②リレーショナルデータベース
データテーブル同士が、キーIDや病名番号などで結合されたデータベース

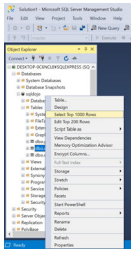


SQL入門①

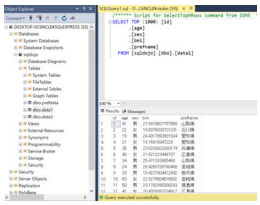
データ外観、SELECT文、条件設定まで

注意：
・これから挙げるクエリ例は、すべてMicrosoft SQL Serverの記述方法であり、ほかのソフトウェアでは方言が異なることにご注意ください。
・また、同じ結果を得るためのクエリの書き方は、必ずしも1通りではなく、わかりやすさを優先して例を載せているので、各自でもっとも適したと思われるクエリを探してみてください。

それぞれのデータテーブルを眺める



データベース内のテーブル名を右クリック
「上位1000行の選択」をクリック



出現したクエリ
SELECT TOP 1000 *
FROM [データシート名]

→ 右下に結果が表示される

SELECT文①基本、一般

基本クエリ

SELECT [変数名1], [変数名2], ..., [変数名n]
FROM [データシート名]

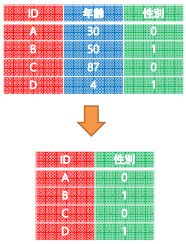
ポイント
＊ SELECTの後の変数名は、半角のカンマ「,」で繋ぐ
＊ SELECT A FROM Bという語順に慣れる

参考：
①上から数えてN行だけ表示したいときはTOP 1000などと変数名の前に置く
②全ての変数を表現したいときは半角のアスタリスク「*」で可能

SELECT文②：基本、具体

(1) IDと性別だけのテーブルが欲しい

SELECT [変数名1], [変数名2], ..., [変数名n]
FROM [データシート名]

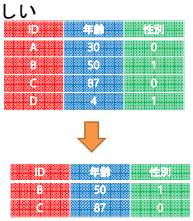


SELECT [ID], [性別]
FROM [TEST_TABLE]

SELECT文③：条件

(1) 条件を満たす（例:40歳以上）テーブルが欲しい

SELECT [変数名1], [変数名2], ..., [変数名n]
FROM [データシート名]



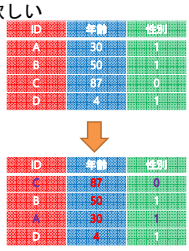
SELECT [ID], [性別]
FROM [TEST_TABLE]
WHERE [年齢]>=40

ポイント
＊ 条件はWHEREでFROMよりも後ろに記述

SELECT文④：並べ替え

(2) 並べ替えた（例: 年齢で降順）テーブルが欲しい

SELECT [変数名1], [変数名2], ..., [変数名n]
FROM [データシート名]

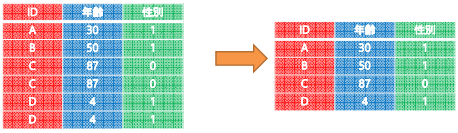


SELECT [ID] , [年齢], [性別]
FROM [TEST_TABLE]
ORDER BY [年齢] DESC

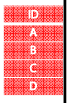
ポイント
＊ 並べ替え指定はORDER BYでFROMよりも後ろに記述
＊ 降順はDESC、昇順はASCで変数の後ろに記載

SELECT文⑤：重複の削除

同じ情報を削除して表示する
SELECT DISTINCT * FROM [データシート名]



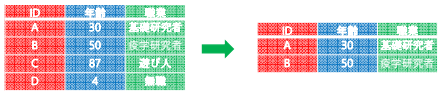
SELECT DISTINCT ID FROM [データシート名]だとどうなる？



SELECT DISTINCT 性別 FROM [データシート名]だとどうなる？

条件の指定①LIKE/IN

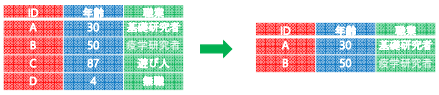
☆データ抽出条件に用いる
SELECT * FROM [tablex]
WHERE [職業] LIKE '基礎研究者' OR [職業] LIKE '疫学研究者' /*☆*/



ポイント
・ 複数の条件はきちんと変数名から書く（同じものでも省略不可）
☆ LIKEの具体例を並べるときは下記のようにINを使っても書ける
WHERE [職業] IN ('基礎研究者', '疫学研究者')
☆ LIKEの後に部分一致を使いたいときは%を用いて'%研究者%'とすればよい（後方一致なら2回目の%は不要）

条件の指定②BETWEEN

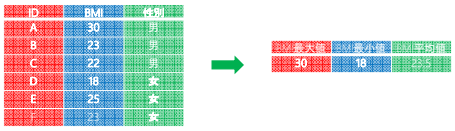
☆データ抽出条件に用いる
SELECT * FROM [tablex]
WHERE [年齢] BETWEEN 30 AND 50 /*☆*/



ポイント
・ BETWEEN [下限値] AND [上限値]で2つの数字の間という条件を作れる
← 上限と下限を逆にすると誤った条件と考えられ不可
（クエリは流れ、赤字のエラーはでないが、うまくいかない）
☆ 本表記は下記のように2条件の組み合わせでも書くことが出来る
WHERE [年齢] >=30 AND [年齢] <=50

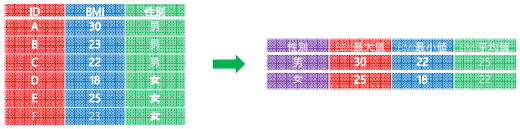
集計値表示①MAX/MIN/AVG/SUM

☆データの集計値を出す（ほかに標準偏差, STDEV）
SELECT MAX(BMI) as BMI**最大値**,
MIN(BMI) as BMI**最小値**,
AVG(BMI) as BMI**平均値**
SUM(BMI) as BMI**合計値**
FROM [tablex]



集計値表示②GROUP BY

☆層ごと（ここでは性別）のデータの集計値を出す
SELECT MAX(BMI) as BMI**最大値**,
MIN(BMI) as BMI**最小値**,
AVG(BMI) as BMI**平均値**
FROM [tablex]
GROUP BY **性別**



テーブルの保存と削除

・保存
SELECT *
INTO [**保存したいテーブル名**] FROM [**既存の**Table]
注）基本的には絶対パスで記述
注）テーブル名を#で始めると一時的なテーブルとして保存される
　　一時的なテーブルとはタブを閉じると消去されるテーブル

・削除
DROP TABLE　[**削除したいテーブル名**]
注）一旦落とすと戻せないから注意が必要

ここまでのまとめ：SQL入門①

①基本構文を理解
SELECT [**変数名**1], [**変数名**2], [**変数名**3], ..., [**変数名**N]
INTO [**保存したいテーブル名**] FROM [**既存の**Table]

②条件や並べ替えの書き方の理解
条件はWHEREを使って記述
並べ替えはORDER BYを使って記述
共に FROM [テーブル名]の後ろに記述

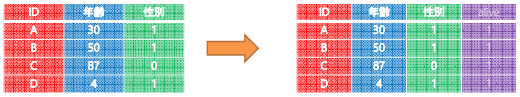
②集計値記述
GROUP BYを使った記述方法

SQL**入門②**

変数作成と日付関数

変数作成①：新変数の作成(1)

新変数（alive: 全て1）を追加する
SELECT *, 1 as alive
INTO #1 FROM [**データシート名**]

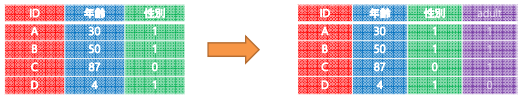


STATA:
generate alive =1
R:
tablex %>% mutate(alive = 1)

変数作成②：新変数の作成(2)

新変数（adult: 20歳以上が1、それ以外は0）を追加する

SELECT *, case when 年齢 >= 20 then 1 else 0 end as adult
INTO #1 FROM [データシート名]



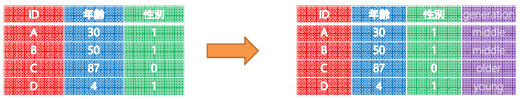
STATA:
generate adult =(年齢>=20)
もしくは
generate adult =1
replace adult = 0 if 年齢<20

R:
tablex %>% mutate(adult= (年齢>=20)) %>%
as.integer()
もしくは
tablex %>% mutate(adult=ifelse((年齢
>=20)|(年齢<20),1|1)

変数作成③：新変数の作成(3)

新変数（generation: 18未満をyoung、18-64をmiddle、65以上をolder）を追加する

SELECT *, case when 年齢 < 18 then 'young' when 年齢 >= 18 and年齢 <65
then 'middle' when 年齢 >= 65 then 'older' else null end as generation
INTO #1 FROM [データシート名]

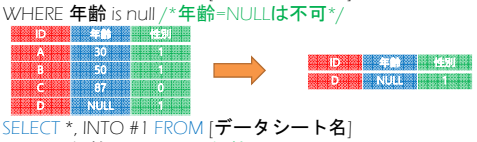


ポイント
＊ 複数条件はwhenを繰り返す
＊ 最後のendを忘れないように！

変数作成④：nullの扱い

☆欠測はNULLで表される（数字変数でも文字列扱いだが”は付けない）

SELECT *, INTO #1 FROM [データシート名]
WHERE 年齢 is null /*年齢=NULLは不可*/

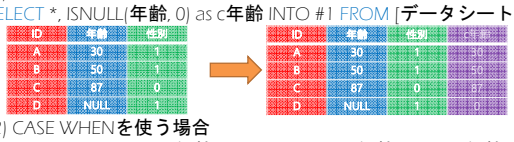


SELECT *, INTO #1 FROM [データシート名]
WHERE 年齢 is not null /*年齢!=NULLは不可*/

変数作成⑤：nullを置き換える

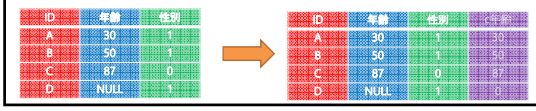
(1) ISNULLを使う場合

SELECT *, ISNULL(年齢, 0) as c年齢 INTO #1 FROM [データシート名]



(2) CASE WHENを使う場合

SELECT *, CASE WHEN 年齢 is null then 0 else 年齢 end as c年齢 INTO #1 FROM [データシート名]



変数作成⑥：DATEDIFF関数を用いた年齢作成

☆DATE関数を用いて日付と認識させるところから（StataやRと同様）

ID	誕生年月日	入院年月日
A	19750105	20191001
B	19611221	20200502
C	19871029	20220819

ID	誕生年月日	入院年月日	年齢
A	19750105	20191001	
B	19611221	20200502	
C	19871029	20220819	

☆SQL自体が日付と変数を認識している場合
SELECT *, DATEDIFF(YEAR, [誕生年月日],[入院年月日]) as 年齢 into #tableX1
FROM [tablex]
注）月齢ならYEAR → MONTH、日齢ならYEAR → DAYにすればよい

☆ [年月日]が日付認定されずエラーが生じる場合
日付認識されていない場合は[誕生年月日]のところに
CONVERT(date, [誕生年月日], 112)
とすると日付に一旦変換される

この「112」はYYYYMMDDで表記された数字を日付として読み込むための番号
ちなみに、YYYY/MM/DDを読み込むなら111（詳細は成書参照）

ここまでのまとめ：SQL入門②

①変数作成を理解
(1) 新変数の場合
SELECT 変数の値 as [新変数名] FROM [既存のTable]
(2) 既存値を用いた変数の場合
SELECT [旧変数名] as [新変数名] FROM [既存のTable]

②変数作成に役立つ関数
case when [条件] then [値1] else [値2] end as [新変数名]

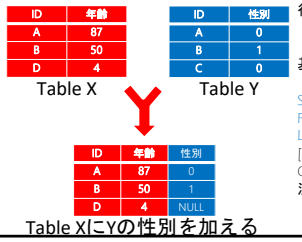
②日付計算
DATEDIFFを使った計算方法

SQL入門③

データ結合

テーブル結合①：導入

☆SQLを用いる一番のメリット → データ結合の容易さ
例：年齢しかないテーブルに性別という情報を追加したい



行うこと
IDをキーにして情報を結合

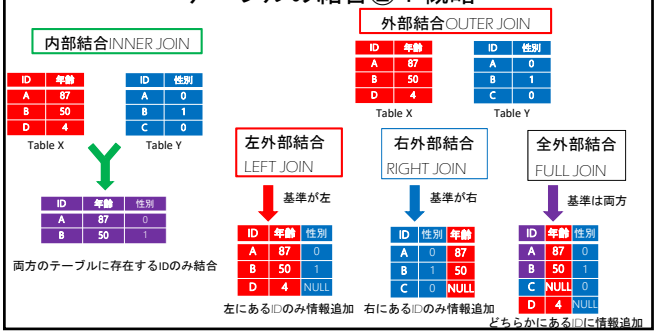
基本クエリ

```
SELECT X.*, Y.性別
FROM [データシート1] X
LEFT JOIN
[データシート2] Y
ON X.ID=Y.ID
```

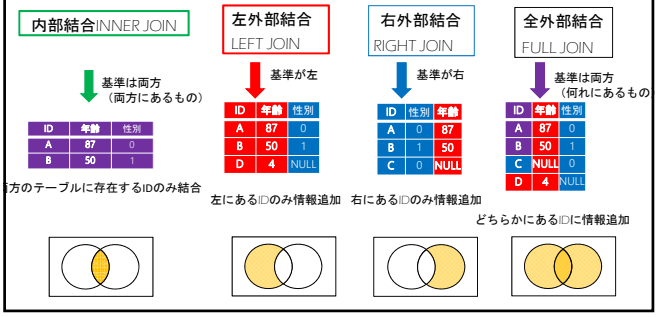
注）元々対応する

Table XにYの性別を加える

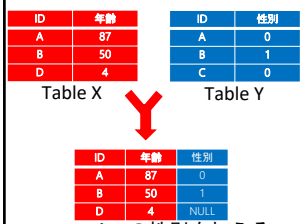
テーブルの結合②：概略



テーブルの結合③：概念



テーブル結合④：具体例



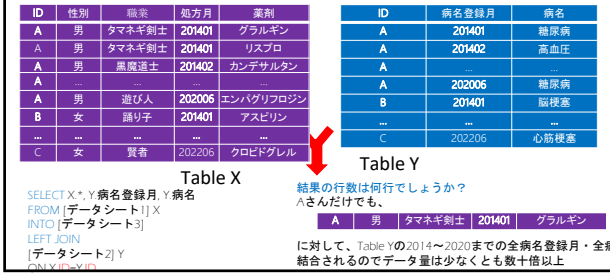
実際のクエリ

```
SELECT X.*, Y.性別 /*Xの全てを、Yから性別だけを取る①*/
FROM [データシート1] X /*シートをXと名付ける②*/
INTO [データシート3] /*into の位置はここ③*/
LEFT JOIN
[データシート2] Y /*Yと名付ける②*/
ON X.ID=Y.ID /*XとYのIDをキー変数と宣言④*/
```

注）
①XとYの両方にある変数を両方選ぶとエラーが出る
(1) どちらから選んでくるのか不明
(2) 同じ変数名は同じテーブルには存在できない
②名付けるのはX/Yではくても A/Bなどとしてもよい
③INTOの位置に注意
④キー変数はIDのみならず、二つ以上も可能
その場合はON X.ID=Y.ID AND X.AGE=Y.AGE
などとする

テーブル結合⑤：注意点！

☆結合されるキーを意識しないと、データ量が膨大になる




```
SELECT X.*, Y.病名登録月, Y.病名
FROM [データシート1] X
INTO [データシート3]
LEFT JOIN
[データシート2] Y
ON X.ID=Y.ID
```

結果の行数は何行でしょうか？
Aさんだけでも、

A	男	タマネギ剣士	201401	グラルギン

に対して、Table Yの2014～2020までの全病名登録月・全病名が結合されるのでデータ量は少なくとも数十倍以上

テーブル結合⑥：問題解決へ(1)

☆結合されるキーを意識してクエリを作る！

ID	性別	職業	処方月	薬剤
A	男	タマネギ剣士	201401	グラルギン
A	男	タマネギ剣士	201401	リスプロ
A	男	黒魔道士	201402	カンデサルタン
A	…	…	…	…
A	男	遊び人	202006	エンパグリフロジン
B	女	踊り子	201401	アスピリン
…	…	…	…	…
C	女	賢者	202206	クロピドグレル

Table X

ID	病名登録月	病名
A	201401	糖尿病
A	201402	高血圧
A	…	…
A	202006	糖尿病
B	201401	脳梗塞
…	…	…
C	202206	心筋梗塞

Table Y

☆解決策①
Table X, Yともに繋げる前にシンプルにしておく
例：select distinct ID, 性別 into #X1 from #X
select distinct ID, 病名 into #Y1 from #Y
それぞれ作ってから、結合する

☆解決策②
結合されるキーを複数にする

```
SELECT distinct X.ID, X.性別, X.処方月,
Y.病名登録月, Y.病名
FROM #X X
LEFT JOIN
#Y Y
ON X.ID=Y.ID AND X.処方月=Y.病名登録月
```

テーブル結合⑦：問題解決へ(2)

☆よく生じる問題 → 結合で同一のものが複数行生成

ID	性別	処方月	薬剤	病名
A	男	201401	グラルギン	糖尿病
A	男	201401	グラルギン	高血圧
A	男	201401	グラルギン	心筋梗塞
A	男	201401	リスプロ	糖尿病
A	男	201401	リスプロ	高血圧
A	男	201401	リスプロ	心筋梗塞
…	…	…	…	…

☆事前の解決策
① 結合するテーブルをそれぞれ、繋げる前にシンプルにしておく（既出）
② 上位XX行（例えば1000行）だけで実行してみる

☆事後の解決策：統計ソフトと同じく結果は必ず確認すること
① 結果の「XXXX行 処理されました」の「XXXX」が欲しい行数か確認
② 同じIDが複数あるかを確認, distinctなどを使い重複のIDの有無を確認

```
SELECT count([ID]) from TableZ
SELECT count(distinct [ID]) from TableZ
```

テーブル結合⑧：データのappend（参考）

☆データのappendはUNIONを使う

ID	病名登録月	病名
A	201401	糖尿病
A	201402	高血圧
A	…	…
A	202006	糖尿病
B	201401	脳梗塞
…	…	…
C	202206	心筋梗塞

Table X

ID	病名登録月	病名
D	201401	関節リウマチ
D	201402	ステロイド糖尿病
D	…	…
D	202006	IgA腎症
E	201401	真性多血症
…	…	…
F	202206	多発性嚢胞腎

Table Y

```
SELECT ID, 病名登録月, 病名
FROM TableX
INTO TableZ
UNION
SELECT ID, 病名登録月, 病名
FROM TableY
```

ここまでのまとめ：SQL入門③

☆データ結合、特に左外部結合LEFT (OUTER) JOINを理解

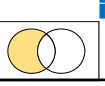
```
SELECT X.*, Y.性別 /*Xの全てを、Yから性別だけを取る①*/
FROM [データシート1] X /*シート1をXと名付ける②*/
INTO [データシート3] /*into の位置はここ③*/
LEFT JOIN
[データシート2] Y /*Yと名付ける②*/
ON X.ID=Y.ID /*XとYのIDをキー変数と宣言④*/
```

ID	年齢
A	87
B	50
D	4

Table X

ID	性別
A	0
B	1
C	0

Table Y



ID	年齢	性別
A	87	0
B	50	1
D	4	NULL

Table XにYの性別を加える

本日のまとめ：SQL入門

①ビッグデータ時代におけるSQL活用の必要性を説明出来る

②SELECT文を用いて、データテーブルの内容や要約量計算を表示できる

③外部結合を用いて、データテーブルの結合が出来る

ご清聴ありがとうございました

NDB・DPCデータベース研究人材育成Webinar

SQLを用いた
レセプトデータのハンドリング

東京大学大学院医学系研究科臨床疫学・経済学
松居宏樹

この授業では

我が国における代表的かつ重要な大規模医療リアルワールドデータベースである、**レセプト情報・特定健診等情報データベース(NDB)**の成り立ち、構造、利活用事例について学ぶ。
そのうえで、データ（特にレセプト情報部分）について、その**データハンドリング手法を実際にデータを触りながら学ぶ。**

学習サイト

学習を円滑に進めるため、一時的なWebサイトを公開しました。

https://sites.google.com/m.u-tokyo.ac.jp/ndb-study/

はじめに

NDB とは何か

- 正式名称：レセプト情報・特定健診等情報 データベース
- 利用目的は全国医療費適正化計画及び都道府県医療費適正化計画の作成、実施及び評価 に資するため（高齢者の医療の確保に関する法律 第16条）
- 保有：厚生労働大臣
- 収集元：審査支払機関
- 内容
  - レセプトデータ 年間約18億7000万件（H28年度）
  - 特定健診・保健指導データ 年間約2,730万件（H28年度）
  - Administrative Claims Database の一種
- 窓口など：
  https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/reseputo/index.html

ここ最近の研究成果

1. Ishimaru M, Matsui H, Ono S, Hagiwara Y, Morita K, Yasunaga H. Preoperative oral care and effect on postoperative complications after major cancer surgery. British Journal of Surgery. 2018 Oct 12;105(12):1688–96.
2. Takeuchi Y, Kumamaru H, Hagiwara Y, Matsui H, Yasunaga H, Miyata H, et al. Sodium-glucose cotransporter-2 inhibitors and the risk of urinary tract infection among diabetic patients in Japan: Target trial emulation using a nationwide administrative claims database. Diabetes Obes Metab. 2021 Jun;23(6):1379–88.
3. Ishimaru M, Ono S, Morita K, Matsui H, Hagiwara Y, Yasunaga H. Prevalence, Incidence Rate, and Risk Factors of Medication-Related Osteonecrosis of the Jaw in Patients With Osteoporosis and Cancer: A Nationwide Population-Based Study in Japan. Journal of Oral and Maxillofacial Surgery. 2022 Apr 1;80(4):714–27.
4. Hashimoto Y, Matsui H, Michihata N, Ishimaru M, Yasunaga H, Aihara M, et al. Incidence of sympathetic ophthalmia after inciting events: a national database study in Japan. Ophthalmology. 2021 Sep 21;S0161-6420(21)00719-3.
5. Kasajima M, Eggleston K, Kusaka S, Matsui H, Tanaka T, Son BK, et al. Projecting prevalence of frailty and dementia and the economic cost of care in Japan from 2016 to 2043: a microsimulation modelling study. The Lancet Public Health. 2022;7(5):e458–68.

NDBデータ構造

NDBデータの大まかな構造

レセプト部分と健診情報部分がある。また、介護情報との接続も可能に

- 医科(MED)
- DPC(DPC)
- 調剤(PHA)
- 歯科(DEN)

→ レセプト情報

- 特定健診データ
- 保健指導データ

→ 健診情報

- 介護レセデータ

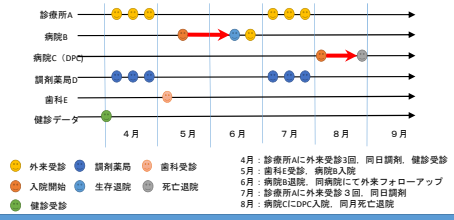
→ 介護情報

今回の演習では主として医療レセプトのデータハンドリングを学ぶ

NDBデータ構造

NDBデータ の大まかな構造

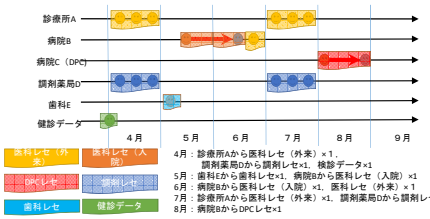
患者のエピソードで考えると以下の様になる。



4月：診療所Aに外来受診3回，同日調剤，健診受診
5月：歯科E受診，病院B入院
6月：病院B退院，同病院にて外来フォローアップ
7月：診療所Aに外来受診３回，同日調剤
8月：病院CにDPC入院，同月死亡退院

NDBデータ構造

NDBデータ の大まかな構造

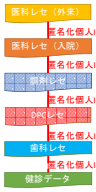


4月：診療所Aから医科レセ（外来）×１，
　　調剤薬局Dから調剤レセ×1，検診データ×1
5月：歯科Eから歯科レセ×1，病院Bから医科レセ（入院）×1
6月：病院Bから医科レセ（入院）×1，医科レセ（外来）×１
7月：診療所Aから医科レセ（外来）×1，調剤薬局Dから調剤レセ×1
8月：病院BからDPCレセ×1

NDBデータ構造

NDBデータ の大まかな構造

匿名化された個人IDで紐付けを行う事が出来る



- NDB には2つのID が存在する。
- 匿名化ID1
  保険者番号、被保険者証 の記号・番号、生年月日、性別を基に作成
- 匿名化ID2
  氏名、生年月日、性別を基に作成

ハンドリング

ID構造と縦断データの作成

匿名化個人IDリストから、追跡可能な全ての匿名化個人IDリストを抽出する

匿名化ID1	匿名化ID2	ID3_A	ID3_B	ID3_C	ID3_D
A	1	B_1	A_1	C_2	A_1
B	1	B_1	B_2	C_2	B_1
B	2	C_2	B_2	C_2	B_2
C	2	C_2	C_2	C_2	C_2
D	3	D_3	D_3	D_3	D_3
E	4	E_4	E_4	E_4	E_4

ID3という個人IDを作成するとして、
ID3_AはID2を基準に作成（例：名前が変更されると追跡不能だが、保険者変更を追跡可能）
ID3_BはID1を基準に作成（例：保険者変更を追跡不能だが、名前の変更を追跡可能）
ID3_CはID1 or ID2 を基準に作成（例：保険者の変更・名前の変更を追跡可能だが、同姓同名問題がある。）
ID3_Dは両者を基準に作成（例：性別が同じ双子も判別可能だが、追跡率は落ちる。）

NDBデータ構造

NDBデータ の大まかな構造

- レセプトの記載内容がそのまま登録され

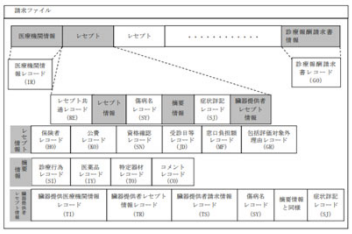
11/

レセプトデータの構造化

共通の仕様が公開されている。

- レセプト仕様書（診療報酬情報提供サービス）
  - http://shinryohoshu.mhlw.go.jp/shinryohoshu/receMenu/doReceInfo
- レセプト仕様書（社会保険診療報酬支払基金）
  - https://www.ssk.or.jp/seikyushiharai/rezept/iryokikan/iryokikan_02.html

レセプトデータの構造化



レセプトデータの構造化

演習：データを検索してみる

社会保険診療報酬支払基金よりサンプル医科レセをダウンロード

https://www.ssk.or.jp/jigyonaiyo/jigyonaiyo_03/index.html

./medsample/サンプル１/0_COMMON001_MED/11_RECODEINFO_MED.CSV

レセプトデータの構造化

演習：データを検索してみる

- CSV形式のレセプトをエクセルで開く
  "./medsample/0_COMMON001_MED/11_RECODEINFO_MED.CSV"
- サンプルデータをレセプト単位に切り分ける
  MNレコードが切れ目
- サンプルデータから、医科レセの医薬品処方履歴を集める。
- ”サンプル　ー“さんの”尿素クリーム１０％「フジナガ」”の処方日は？

レセプトデータの構造化

- レセプトデータは生のままでは扱いにくい。
- 行ごとに書いてある情報が違う

- 行の種類ごとに分けてデータベース化しよう。

実際の研究とデータ構造

構造化されたデータの研究利用

NDBのテーブル構造はあまり大っぴらにされていない。

NDBの提供について（厚労省）のサイト内にある、抽出依頼テンプレートに詳細がある。

- 申出依頼テンプレート：当方と相談の上、必要に応じて作成してください
  - 申出依頼テンプレートの注意点［XLS形式：1,210KB］
  - 申出依頼テンプレート（抽出）［XLS形式：1,720KB］
  - 申出依頼テンプレート（集計）［XLS形式：1,530KB］

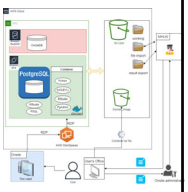
DB仕様は機械判読可能な形で

エクセル内に提供ファイルのカラム型情報あり

特別抽出仕様NDB データベース

postgresql 上にサンプルレセプトを基にレセプトデータベースを作成

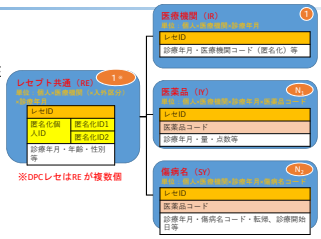
- 特別抽出のデータ仕様に合わせたデータベースを構築
  - テーブル名・カラム名は先述の仕様を基に決定
  - サンプルレセプトをデータベースに格納
- データベースソフトウェアはpostgresql ver.14.6を用いた。
  - オンサイトセンターでの採用
  - Redshift などの種々のクラウドマネージドデータベースと構文が似ている。
  - オープンソースソフトウェアで導入しやすい。



NDBデータ構造

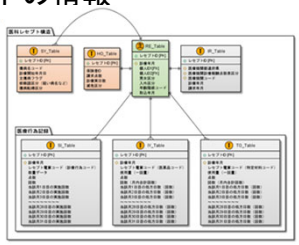
NDBデータ の大まかな構造

- 複数のテーブル（右図）で1レセの情報となる。
- 匿名化個人ID は各REに含まれる、匿名化ID1と匿名化ID2 の2つ
- 他のレセプト（MED, DPC, PHA, DEN）情報（や健診情報）とは匿名化個人IDで接続可能
- 月単位で登録されるため、月またぎ入院や月内外来受診の処理が煩雑



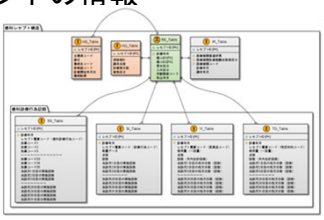
医科レセプトの情報

- 医療機関毎・診療年月毎・入外毎にRE（レセプトヘッダ）が発生
- RE に紐づく形で全18テーブル
- 研究に主として用いるテーブルは概ね７種類



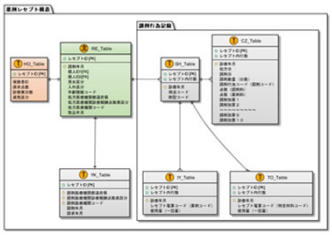
歯科レセプトの情報

- 医療機関毎・診療年月毎・入外毎にRE（レセプトヘッダ）が発生
- RE に紐づく形で全15テーブル
- 研究に主として用いるテーブルは概ね8種類



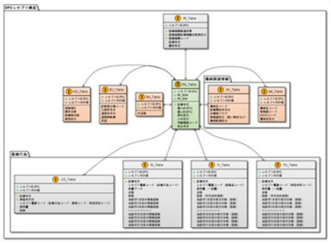
調剤レセプトの情報

- 調剤機関毎・調剤年月毎にREが発生
- RE に紐づく形で全14テーブル
- 研究に主として用いるテーブルは概ね7種類
- 薬剤の処方状況を確認するために、レセ中の出現行数を基にSHに接続が必要



DPCレセプトの情報

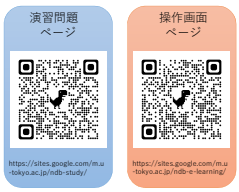
- 医療機関毎・診療年月毎に複数のREが発生
- REとその出現行数 に紐づく形で全26テーブル
- 研究に主として用いるテーブルは概ね11種類



演習環境の説明

以降は構築したデータベースを用いた演習を行う

- この演習では、オープンソースPostgresql Webassembly である、postgres-wasm(https://github.com/snaplet/postgres-wasm) を用いて、参加者のブラウザ上に小規模なpostgresql サーバーを立ち上げています。
- ブラウザを閉じると、構築された環境が消去されます。
- 演習問題ページとデータベース操作画面を並べて表示してください。



二つのページを並べて表示

演習終了後のまとめ

今回触れなかった内容

- 環境の整備
  - postgresql サーバーのインストール（ https://www.postgresql.jp/ ）
  - DB構築用DDL
  - CSVデータのデータベースへの取り込み
  - データの圧縮やインデックスを用いた効率的なクエリに書き方
- レセプトデータを整理した後の疫学研究

- アンケートにご協力ください。

NDB・DPCデータベース研究人材育成Webinar

NDBを用いた研究

東京大学大学院医学系研究科臨床疫学・経済学
佐藤壮

目次

1. リアルワールドデータ研究のレビュー
2. 研究紹介

目次

1. リアルワールドデータ研究のレビュー
2. 研究紹介

背景

- リアルワールドデータを二次利用した研究は近年増えているが、その具体的な程度については不明である。
- リアルワールドデータを使用した研究を包括的にレビューした論文はHiroseのNDB研究(ACE 2020;2:13-26)に関するもののみである。
- リアルワールドデータを使用した研究の全体としての傾向や、それぞれのデータベース研究での研究分野等を知ることは今後の研究を行う上で有用である。

【参考】先行NDB研究レビュー[1]の要約

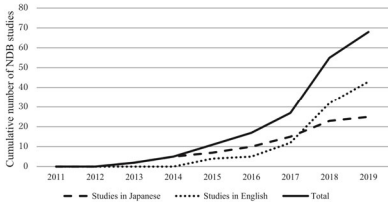
【方法】

- 情報源：PubMed, 医中誌Web, 厚生労働省社会保障審議会（医療保険部会 匿名医療情報等の提供に関する専門委員会)資料
- 検索期間：2011/1 – 2019/6
- 検索対象：NDB, NDB Open Dataを使用した研究の原著論文
  1. 研究デザイン
  2. 研究分野
  3. 設定、サンプル数
  4. 研究結果
  5. 研究のstrength, limitation

【参考】先行NDB研究レビューの要約

【結果】

- 最終的なレビュー対象：英語論文43件、日本語論文25件→合計**68件**
- 研究デザイン：記述研究42件、横断研究14件、後ろ向きコホート研究12件
- 研究分野：臨床医学(18件)、薬理学・薬学(8件)、感染症学(7件)を含む36分野



今回レビュー対象としたリアルワールドデータ

- NDB (National Database of Health Insurance Claims and Specific Health Checkups of Japan)
- NDB Open Data
- JMDC Claims Database
- DPC (Diagnosis Procedure Combination) database

6

レビュー方法（NDB, NDB Open Data）

- 検索エンジン：PubMed
- 検索期間：2013/1/1 – 2022/7/31
- 検索ワード：(((claims and NDB) OR (National Database of Health Insurance Claims) OR (National Database of Health Insurance Claim) OR (National Database of Japanese Health Insurance Claims) OR (nationwide administrative claims database)) AND (Japan OR Japanese)) AND (2013:2022[pdat]) NOT (Korea) NOT (Taiwan)

7

レビュー方法（JMDC Claims Database）

- 検索エンジン：PubMed
- 検索期間：2010/1/1 – 2022/7/31
- 検索ワード：(("JMDC") OR ("jmdc database") OR ("Japan Medical Data Center")) AND (Japan OR Japanese)

8

レビュー方法（DPC database）

- 検索エンジン：PubMed
- 検索期間：2010/1/1 – 2022/7/31
- 検索ワード：

(("Diagnosis Procedure Combination database"[Title/Abstract]) OR ("nationwide database"[Title/Abstract]) OR ("administrative database"[Title/Abstract]) OR ("inpatient database"[Title/Abstract]) OR ("discharge database"[Title/Abstract])) AND (Japan OR Japanese)

9

レビュー方法（全体）

- 組み入れ基準
  1. 原著論文かつNDB, NDB Open Data, the JMDC Claims Database, the DPC databaseのいずれかを使用した研究であること
  2. 英語論文であること
- 除外基準
  1. letters, notes等のoriginal article以外のもの
  2. PubMedからFull textにアクセスできない場合

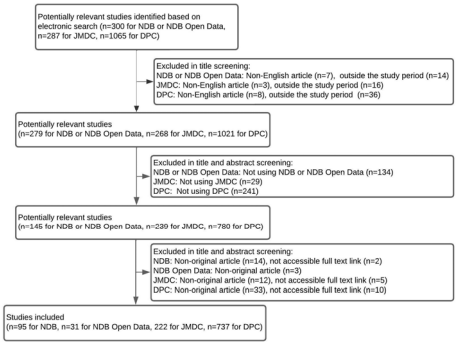
10

レビュー方法（全体）

- 抽出したデータ
  1. 主に使用されたデータベース
  2. 研究論文が雑誌に掲載された時期（年）
  3. 研究分野 (注1)

11

結果

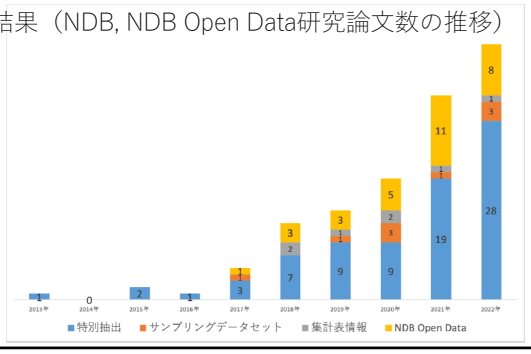


結果（前ページ要約）

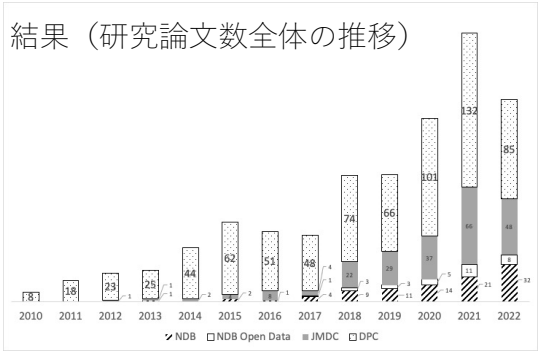
- NDB研究：95件
  - 特別抽出79件、サンプリングデータセット9件、集計表情報7件
- NDB Open Data研究：31件
- the JMDC Claims Database研究：222件
- the DPC database研究：737件

→合計1085件

結果（NDB, NDB Open Data研究論文数の推移）



結果（研究論文数全体の推移）



結果（研究分野）

	NDB	NDB Open Data	JMDC Claims Database	DPC Database
1位	内科	内科	内科	内科
2位	精神科	整形外科	精神科	外科
3位	整形外科	その他	眼科	救急科
4位	産婦人科	産婦人科 リハビリテーション科	小児科	整形外科
5位	リハビリテーション科 総合診療科	-	産婦人科	小児科

研究分野	NDB, n(%)	NDB Open Data, n(%)	JMDC, n(%)	DPC, n(%)	All, n(%)
合計	95	31	222	737	1085
内科	49 (51.6)	12 (38.7)	144 (64.9)	216 (29.3)	421 (38.8)
小児科	3 (3.2)	1 (3.2)	12 (5.4)	44 (6.0)	60 (5.5)
皮膚科	1 (1.1)	1 (3.2)	4 (1.8)	3 (0.4)	9 (0.8)
精神科	12 (12.6)	1 (3.2)	17 (7.7)	13 (1.8)	43 (4.0)
外科	3 (3.2)	1 (3.2)	0 0.0	139 (18.9)	143 (13.2)
整形外科	7 (7.4)	5 (16.1)	6 (2.7)	57 (7.7)	75 (6.9)
産婦人科	5 (5.3)	2 (6.5)	7 (3.2)	14 (1.9)	28 (2.6)
眼科	3 (3.2)	1 (3.2)	14 (6.3)	4 (0.5)	22 (2.0)
耳鼻咽喉科	0 0.0	0 0.0	3 (1.4)	15 (2.0)	18 (1.7)
泌尿器科	0 0.0	1 (3.2)	2 (0.9)	22 (3.0)	25 (2.3)
脳神経外科	1 (1.1)	0 0.0	2 (0.9)	43 (5.8)	46 (4.2)
放射線科	0 0.0	1 (3.2)	0 0.0	3 (0.4)	4 (0.4)
麻酔科	0 0.0	0 0.0	0 0.0	15 (2.0)	15 (1.4)
病理	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
臨床検査	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0
救急科	0 0.0	0 0.0	2 (0.9)	100 (13.6)	102 (9.4)
形成外科	0 0.0	0 0.0	0 0.0	3 (0.4)	3 (0.3)
リハビリテーション科	4 (4.2)	2 (6.5)	2 (0.9)	24 (3.3)	32 (2.9)
総合診療科	4 (4.2)	0 0.0	2 (0.9)	15 (2.0)	21 (1.9)
その他	3 (3.2)	3 (9.7)	5 (2.3)	7 (0.9)	18 (1.7)

考察

- (NDBを含む)リアルワールドデータを使用した研究は全体かつ個別で増加傾向にある。
  - 海外のデータベース研究数と比較しても遜色ない。
- JMDC Claims Database, DPC databaseを使用した研究が多い。
- NDB, NDB Open Data, JMDC Claims Databaseでは記述研究が多く、DPC databaseでは比較研究が多い。
- NDB, NDB Open Data, JMDC Claims Databaseは健診情報が含まれることから生活習慣病関連の研究も多い。

18

補足（Limitation等）

- NDBを使用した研究に関しては、日本語論文も数多く出版されている。
- 研究分野の選択は、議論の余地がある可能性。
- 今回調査したデータベース以外にも日本のリアルワールドデータが存在する。

19

【小括】リアルワールドデータ研究レビュー

- リアルワールドデータを使用した研究は増加傾向（平均増加率約41%）。
- それぞれのデータベースごとの特色があるため、研究目的に適したデータベース選択を行う必要がある。

20

目次

21

当研究室の先行NDB研究紹介

1. Ishimaru M, Matsui H, Ono S, Hagiwara Y, Morita K, Yasunaga H. Preoperative oral care and effect on postoperative complications after major cancer surgery. Br J Surg. 2018 Nov;105(12):1688-1696.
2. Takeuchi Y, Kumamaru H, Hagiwara Y, Matsui H, Yasunaga H, Miyata H, Matsuyama Y. Sodium-glucose cotransporter-2 inhibitors and the risk of urinary tract infection among diabetic patients in Japan: Target trial emulation using a nationwide administrative claims database. Diabetes Obes Metab. 2021 Jun;23(6):1379-1388.
3. Hashimoto Y, Matsui H, Michihata N, Ishimaru M, Yasunaga H, Aihara M, Kaburaki T. Incidence of Sympathetic Ophthalmia after Inciting Events: A National Database Study in Japan. Ophthalmology. 2022 Mar;129(3):344-352.
4. Ishimaru M, Ono S, Morita K, Matsui H, Hagiwara Y, Yasunaga H. Prevalence, Incidence Rate, and Risk Factors of Medication-Related Osteonecrosis of the Jaw in Patients With Osteoporosis and Cancer: A Nationwide Population-Based Study in Japan. J Oral Maxillofac Surg. 2022 Apr;80(4):714-727.

22

当研究室の先行NDB研究紹介①

歯科医による術前口腔ケア介入はがん手術後合併症を減少させるか

Preoperative oral care and effect on postoperative complications after major cancer surgery. ***Br J Surg.*** 2018;105:1688-96

- 主ながん手術の術後合併症として誤嚥性肺炎は2.6－3.5%発生すると言われており、入院日数の延長や死亡率上昇と関連している。
- 口腔ケアによる、口腔環境改善等を通じて術後肺炎を予防できる可能性が示唆されている。
- 先行研究は、サンプルサイズや参加施設数が小さいというLimitationがあり、術前口腔ケアと術後肺炎の関連は不明である。

23

対象者	頭頸部がん、食道がん、肺がん、肝がん、消化器系がん手術を行なった患者
曝露・対照	周術期口腔機能管理料の算定の有無
アウトカム	術後30日以内（入院中）の肺炎発症、全死亡率
デザイン	コホート研究、傾向スコアによる重みづけ

- 509,179名中、81,632名(16%)が術前口腔ケアを受けた
- 傾向スコアによる重み付け解析で、術前口腔ケア群は有意にアウトカム発生割合が少なかった
  （30日肺炎発症 3.28% vs. 3.76%; 30日全死亡 0.30% vs. 0.42%）

24

当研究室の先行NDB研究紹介②

糖尿病患者におけるSGLT2阻害薬と尿路感染症の関連

Sodium-glucose cotransporter-2 inhibitors and the risk of urinary tract infection among diabetic patients in Japan: Target trial emulation using a nationwide administrative claims database. ***Diabetes Obes Metab.*** 2021;23:1379-88.

- ２型糖尿病治療の第一選択薬としてSGLT2阻害薬はDPP-4阻害薬等と同様に使用されているが、尿路感染症（UTI）との関連が示唆されている。
- リアルワールドでは、治療中に治療薬の変更が起こるが先行研究はその影響を考慮しておらず、妥当性が低い可能性がある。

25

対象者	40歳以上の2型糖尿病患者で新規に薬物療法を開始した者
曝露・対照	SGLT2阻害薬 or DPP-4阻害薬 vs. ビグアナイド薬
アウトカム	新規薬物治療開始後1080日以内の尿路感染症発症
デザイン	コホート研究、傾向スコアによる重み付けCox回帰

- 11,364名のSGLT2阻害薬群、9,035名のDPP-4阻害薬群、10,359名のビグアナイド薬群を解析。
- 傾向スコアによる重み付け解析によるハザード比は以下のとおりで、UTIのリスク増加を認めなかった。

SGLT2阻害薬 vs. ビグアナイド

ITT: 0.94 (95% CI 0.86-1.03), PP: 0.90 (95% CI 0.78-1.03)

DPP-4阻害薬 vs. ビグアナイド

ITT: 0.85 (95% CI 0.77-0.94), PP: 0.83 (95% CI 0.71-0.95)

(ITT, intention-to-treat; PP, per-protocol)

26

当研究室の先行NDB研究紹介③

外傷・手術後の交感性眼炎の発生率

Incidence of Sympathetic Ophthalmia after Inciting Events: A National Database Study in Japan. ***Ophthalmology.*** 2022;129:344-52

- 交感性眼炎(Sympathetic ophthalmia)は、一方の目への外傷や手術の曝露で他方の目に生じる汎ぶどう膜炎であり、発症率0.01%程度と稀な疾患である。
- 先行研究では、外傷後の交感性眼炎発症率の報告や、特に硝子体切除術(vitrectomy)でリスク増加の可能性等が示唆されているものの、単施設研究であったり、現在の医療技術水準にはそぐわない時代のものである等のLimitationが挙げられる。

27

対象者	眼への外傷、特定の眼手術を受けた患者
曝露・対照	眼への外傷、特定の眼手術
アウトカム	交感性眼炎、Vogt-Koyanagi-Harada病、（汎ぶどう膜炎）の診断
デザイン	コホート、記述研究

- 888,041の曝露発生(704,717名)が解析され、60ヶ月での累積交感性眼炎発生率は0.044%であった。
- 交感性眼炎は硝子体切除術群 0.016% vs. 外傷群 0.073%であり、外傷の方が4 - 5倍誘引することが示唆された。

28

当研究室の先行NDB研究紹介④

骨粗鬆症およびがん患者における薬剤関連顎骨壊死の有病率・発生率・リスク因子

Prevalence, Incidence Rate, and Risk Factors of Medication-Related Osteonecrosis of the Jaw in Patients With Osteoporosis and Cancer: A Nationwide Population-Based Study in Japan. ***J Oral Maxillofac Surg***. 2022;80:714-27.

- 薬剤関連顎骨壊死（Medication-Related Osteonecrosis of the Jaw, MRONJ）は、骨粗鬆症や、担がん患者への治療薬であるビスホスホネート薬（BP）やRANKL阻害薬に関連する発症率0.001 – 0.1%の稀な疾患である。
- 先行研究は、症例対照研究デザインである、発症イベントは自己報告に基づく等のLimitationもあり、MRONJの有病率、発症率、リスク因子は不明である。

29

対象者	骨粗鬆症、または（かつ）がんで骨吸収抑制薬を使用している患者
曝露・対照	新規の骨吸収抑制薬の開始（BP薬、RANKL阻害薬）
アウトカム	MRONJの発症
デザイン	コホート研究、Cox回帰

- 2,819,310名が新規骨吸収抑制薬を開始されており、2,664,104 (94.5%)が骨粗鬆症、155,206が担がん患者だった。
  - 骨粗鬆症：1,603名 (0.06%)がMRONJを発症、100,000人年あたり22.9の発症率
  - がん：2,274名 (1.47%)がMRONJを発症、100,000人年あたり1,231.7の発症率
- MRONJの発症は、口腔環境の悪さ（骨粗鬆症患者の歯周病でHR 1.30 (95%CI 1.10 – 1.52)）等の因子と関連していることが判明した。

30

【総括】NDBを用いた研究

- NDBを用いた研究は増加傾向
- 国民全体（約98%）のレセプトデータを含む悉皆性
  - 稀なアウトカムの研究
  - 1入院中のアウトカムに限らず追跡可能
- 歯科や健診のデータも含んでいる
  - 特に歯科関連の研究を行うのに有用な可能性

31

ご清聴ありがとうございました

32

NDB・DPCデータベース研究人材育成Webinar

DPCデータを用いた研究１
–概要と研究紹介–

東京大学大学院医学系研究科ヘルスサービスリサーチ講座　特任准教授
城　大祐

DPCデータベース

DPCデータは、一次的には各DPC病院が厚生労働省に提出する目的で作成される。様々な団体がDPCデータを二次的に収集し、臨床研究などに活用している。

（１）厚生労働省が提供するDPCデータ
2022年から個票データの提供を開始
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/dpc/index.html

（２）学会や学術団体等が独自に収集
日本循環器学会JROAD-DPC、厚生労働科学研究DPCデータ調査研究班、など

（３）商用のDPCデータ
Medical Data Vision、など

COI

本発表に関連するCOIはございません

DPCデータベース

様式１の項目
病院属性等　施設コード、診療科コード
データ属性等　データ識別番号、性別、年齢

入退院情報　予定・救急入院、救急車による搬送、
退院時転帰、在院日数

診断情報　主傷病名、入院の契機となった傷病名、
医療資源を最も投入した傷病名、
二番目に医療資源を投入した傷病名
入院時併存症名（10）、入院後発症疾患名（10）
手術情報　手術術式、麻酔

5

- **DPCデータベースとは**
- DPC研究の紹介
- まとめ

DPCデータベース

様式１の項目
診療情報
身長・体重、喫煙指数、入院時・退院時JCS、
入院時・退院時ADL スコア、
がんUICC 病期分類・Stage分類、
入院時・退院時mRS、脳卒中の発症時期、
Hugh-Jones 分類、NYHA 心機能分類、狭心症CCS 分類、急性心筋梗塞Killip 分類、肺炎の重症度、肝硬変Child-Pugh 分類、急性膵炎の重症度、
精神保健福祉法における入院形態・隔離日数・身体拘束日数、
入院時GAF 尺度、SOFAスコア

6

DPCデータベース

ＥＦファイルから得られるデータ

薬剤・特定保険医療材料の
　名称・使用日・使用量

検査・処置の実施

医療費

など

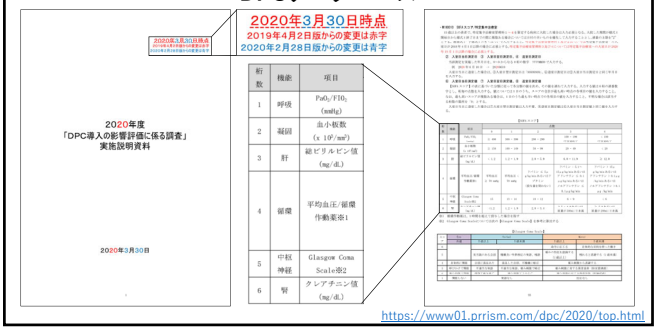
7

DPCデータベース研究の限界

- 患者が病院を変えると追跡できない
  ⇒長期生存の追跡には向いていない
  （5年生存率など）
- 検査結果データが無いので、リスク調整がいつも十分にできるわけではない
  ⇒未測定の交絡がありうる

10

DPCデータベース



https://www01.prrism.com/dpc/2020/top.html

- DPCデータベースとは
- **DPC研究の紹介**
- まとめ

DPCデータベース研究

前向き研究が行い難い研究テーマ
　対象症例（疾患や病態）が少ない
　対象が小児、高齢者や合併症のある症例
　すでに広く普及している介入
　重篤な病態への介入

前向き研究の制約があるテーマ
　外的妥当性（選抜された症例のため、結果が外挿しにくい）
　時間がかかる（サロゲートアウトカムのことが少なくない）
　費用がかかる

9

DPC研究の疾患分野

（2014-2022　城共著）

呼吸器内科
　COPD
　肺がん
　喘息
　びまん性肺疾患
　希少疾患
　肺炎・感染症
　診断および治療の手技
呼吸器外科
リハビリ
薬剤
看護

循環器内科
血液内科
感染症内科
集中治療
救急
麻酔・周術期
小児科
消化器外科
整形外科
眼科

DPC研究の分類
（2014-2022　城共著）

記述的研究　descriptive study

予測モデル prediction model
ノモグラム　nomogram
機械学習　machine learning

分析的研究 analytic study

傾向スコア分析 propensity score analysis

生存分析 survival analysis

目的

薬剤カテゴリー毎に薬剤性肺障害発症のリスクを評価する。

研究事例

Risk of drug-induced interstitial lung disease in hospitalised patients: a nested case–control study

Jo T, et al. *Thorax*. 2021

方法

研究デザイン：コホート内**症例対照研究**
データベース：DPCデータベース
期間：2010年7月から2016年3月

対象：全入院患者
症例：入院中に薬剤性肺障害と診断されステロイド治療を要した症例
対照：それ以外の入院症例からマッチングにより選択

薬剤：
- 添付文書の有害事象に間質性肺炎の記載がある、約6000の薬剤を薬効により75のカテゴリーに分類
- そのうち、過去の大規模コホート研究やシステマティックレビューなど英文誌に記載のある42の薬剤カテゴリーを評価

背景

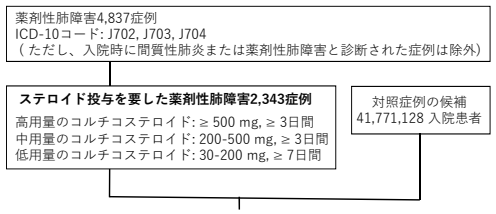
- 薬剤性肺障害は、発症頻度が低く研究が難しい
- 症例報告やケースシリーズでは、因果推論が困難
- 小規模研究の結果は、一般化が難しい
- リンパ球刺激試験などの臨床検査の診断における有効性は高くない
- 原因薬剤の推定が難しい場合がある

方法

統計解析：
- 対照の選択
  症例と対照を、
  性別
  年齢（10歳以内）、
  施設
  前後1年の会計年度
  で1対4マッチング
- 条件付き単変量ロジスティック回帰分析
- 多重代入法による欠損値補完後に
  条件付き多変量ロジスティック回帰分析

結果

対照の選択

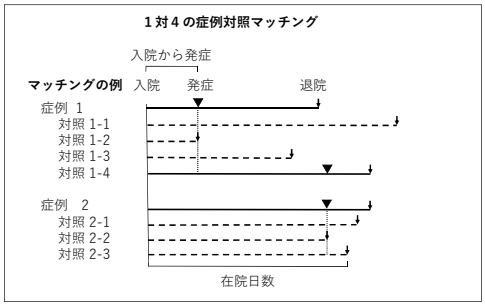


結果

薬剤の選択

75薬剤カテゴリー
　6つの大規模コホート研究やシステマティックレビューに含まれるものを選択
　⇒33薬剤カテゴリーを除外
42薬剤カテゴリー
　カテゴリーあたり5症例未満を除外
　⇒23薬剤カテゴリーを除外
19薬剤カテゴリー
　退院3日前に処方された薬剤を除外後にカテゴリーあたり５症例未満を除外
　⇒2薬剤カテゴリーを除外
17薬剤カテゴリー
　⇒単変量の条件付きロジスティック回帰分析

結果

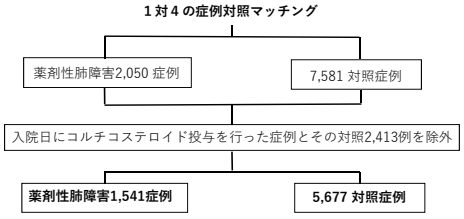


結果

単変量ロジスティック回帰分析による症例と対照のステロイド投与量および在院死亡の比較

	症例		対照				
	N=1541		N=5677				
	n	(%)	n	(%)	オッズ比	95%信頼区間	P値
高用量ステロイド	1246	(80.9)	156	(2.8)	185.92	123.0–280.9	<0.001
中用量ステロイド	119	(7.7)	28	(0.5)	19.30	12.10–30.80	<0.001
低用量ステロイド	671	(43.5)	263	(4.6)	18.47	15.17–22.48	<0.001
在院死亡	235	(34.7)	610	(10.8)	5.24	4.49–6.10	<0.001

結果



結果

17薬剤カテゴリーと薬剤性肺障害発症の条件付き単変量ロジスティック回帰分析

	オッズ比	95%信頼区間	P値
Antiplatelets	1.01	0.69–1.46	0.977
Anticoagulants	3.11	1.16–8.35	0.024
Statins	0.68	0.50–0.92	0.012
Sodium channel blockers	1.28	0.61–2.67	0.511
Class III antiarrhythmic drugs	6.80	3.96–11.66	<0.001
Angiotensin/converting enzyme inhibitor	1.83	0.80–4.20	0.156
Thiazides	0.59	0.31–1.14	0.116
NSAIDs	2.43	2.04–2.88	<0.001
Antiepileptics	1.07	0.59–1.94	0.836
Sulfamethoxazole/trimethoprim	2.89	1.36–6.13	0.006
Quinolones	2.93	2.35–3.65	<0.001
Tetracyclines	2.30	1.50–3.52	<0.001
Beta-lactams	1.62	1.39–1.89	<0.001
Anti-tuberculosis drugs	2.14	0.94–4.89	0.071
EGFR inhibitors	16.48	9.57–28.39	<0.001
Pyrimidine	1.18	0.63–2.21	0.599
Anthracyclines	3.09	1.24–7.70	0.015

結果

多重代入法後に単変量で有意差のあった10薬剤カテゴリーを含めた条件付き多変量ロジスティック回帰分析

	オッズ比	95%信頼区間	P値
Anticoagulants	2.58	0.76–8.81	0.13
Statins	0.53	0.37–0.75	<0.001
Class III antiarrhythmic drugs	7.01	3.86–12.73	<0.001
NSAIDs	1.9	1.56–2.31	<0.001
Sulfamethoxazole/trimethoprim	2.54	1.04–6.24	0.042
Quinolones	3.1	2.41–3.99	<0.001
Tetracyclines	1.6	0.97–2.66	0.067
Beta-lactams	1.54	1.29–1.84	<0.001
EGFR inhibitors	16.84	9.32–30.41	<0.001
Anthracyclines	1.89	0.68–5.23	0.223

調整項目：バーセルインデックス、ブリンクマンインデックス、チャールソン併存疾患インデックス、肺癌、その他の癌、入院２日以内のICU入室、入院２日以内の挿管人工呼吸器管理

結語

コホート内症例対照研究により、薬剤カテゴリーを薬剤性肺障害発症の関連を評価し、薬剤カテゴリー毎の薬剤性肺障害の発症リスクを定量化して示した。

考察

- 本研究では、初めて大規模データベースを用いて薬剤性肺障害の発症リスクを評価した。
- 薬剤性肺障害のリスクが高いと考えられる42カテゴリーの薬剤を評価し、そのうち６つの薬剤が薬剤性肺障害の高リスクであることが示された。
- 薬剤以外では既報と同様に、男性、60歳以上、多い併存疾患、痩せ、肺癌が薬剤性肺障害の発症リスクであることが示された。
- 薬剤性肺障害発症のリスクの高い薬剤のなかでも、EGFR阻害剤とⅢ群の抗不整脈薬が、とりわけ高いオッズ比を示し、これは小規模な既報と合致した。

- DPCデータベースとは
- DPC研究の紹介
- **まとめ**

限界

- 薬剤性肺障害の正確な発症日は不明。
- シスプラチンのようにステロイドと一緒に投与される薬剤の評価が出来ていない。
- 薬剤間の交互作用の探索的評価をしていない。
- 画像や採血などの検査結果は評価が出来ていない。
- 処方が少ない薬剤は評価が出来ていない。
- 未測定の交絡が存在する可能性を排除できない。

まとめ

- DPCデータベースについて概説した
- DPCデータベースを用いた薬剤性肺障害の解析事例を紹介した
- 入院診療にかかわる様々な職種・診療科の方々が、種々の解析法を駆使した研究を行っている
- 柔軟な発想により、前向き研究を補完するような研究が可能である
- データの蓄積により今後も多くの研究が行われていくと考える

NDB・DPCデータベース研究人材育成Webinar

DPCデータを用いた研究2

東京大学大学院医学系研究科臨床疫学・経済学
東京大学大学院医学系研究科乳腺・内分泌外科学
小西孝明

DPCデータに適した研究

➢DPCデータは以下の研究に適している
- 症例数が膨大な点を活かす
  - ✓希少な病態・介入の記述疫学研究
  - ✓（倫理的・経済的・実務的に）RCTが困難な比較研究
- 入院データである点を活かす
  - ✓入院を要する病態・介入の研究
  - ✓入院中に重要なアウトカムが発生する研究

➢救急領域や外科領域などの急性疾患を扱いやすい

臨床疫学研究の内容による分類

1. 治療・予防などの効果判定
2. 診断法や患者評価法
3. 疾病のリスク因子の同定
4. 疾病の予後予測
5. 希少疾患の予後予測
6. 診療実態の分析（practice pattern analysis）
7. 医療の質研究、ヘルス・サービス・リサーチ

などなど

ランダム化比較試験がゴールドスタンダードだが、
後ろ向きデータベース研究でも実施可能

外科領域のDPCデータ研究例

1. 統合失調症患者の乳癌手術成績
2. Body mass indexが食道癌手術に及ぼす影響
3. 尾側膵切除における開腹 vs. 鏡視下手術

外科領域のDPCデータ研究例

1. 統合失調症患者の乳癌手術成績
2. Body mass indexが食道癌手術に及ぼす影響
3. 尾側膵切除における開腹 vs. 鏡視下手術

統合失調症患者の乳癌手術成績

- “Breast cancer surgery in patients with schizophrenia: short-term outcomes from a nationwide cohort”
- Konishi T, Fujiogi M, Michihata N, Tanaka-Mizutani H, Morita K, Matsui H, Fushimi K, Tanabe M, Seto Y, Yasunaga H
- *British Journal of Surgery* 2021;108:168–173

Introduction

➢統合失調症は人口の1%が罹患する慢性精神疾患である
- 乳癌の罹患リスクが高い
- 治療後の長期予後も不良である
- 診断の遅れや不十分な治療の影響があると考えられている

➢乳癌術後短期成績への影響はいまだ不明である
- 股関節手術50万例では、統合失調症は高リスクとの報告あり
- 術後ICUに入室した約9千例では、統合失調症は有意なリスクでないとの報告あり

➢外科医と精神科医のリエゾンが世界的に重要視されつつある

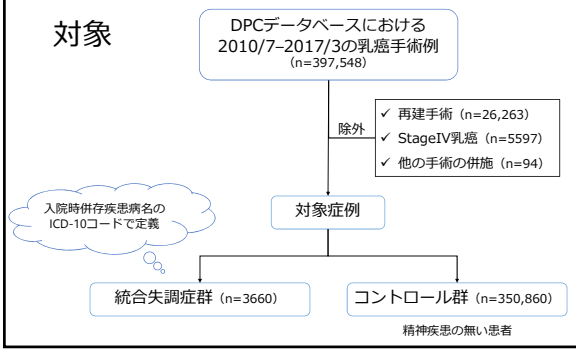
PECO

P：乳癌手術を受けた患者
E：統合失調症あり
C：精神疾患なし
O：全合併症（術後出血・創部感染など）、入院費用

入院後発症病名の
ICD-10コードで定義

➢主解析　：多変量解析（マルチレベル分析を併用）
➢感度分析：1:4 matched-pair cohort解析 など

対象



多変量解析

DPCデータ様式1に
含まれる情報

➢背景因子として下記で調整
- 患者要因：性別、年齢、Body mass index (BMI)、喫煙歴、Charlson併存疾患指数
- 治療要因：癌病期、術式
- 施設要因：臨床研修指定の有無、年間手術件数

➢マルチレベル分析を併用
- 一般化推定方程式
- 患者・治療背景の病院ごとの違いを考慮する

マルチレベル分析について
は別講義も参照ください

感度分析

病名と処方の
組み合わせで
精度向上

① 1:4 matched-pair cohort解析
✓両群から年齢・施設・治療年が同一の患者を1:4の割合で抽出する
✓この解析でも施設間差異の影響を除去できる

② 統合失調症群を抗精神病薬の処方がある患者のみに限定した解析
✓統合失調症群に統合失調感情障害の患者が含まれているかもしれないため
✓真の統合失調症であれば継続的な抗精神病薬の内服が必要

③ 統合失調症群から強制入院の患者を除外した解析
✓自傷他害のおそれがある重症患者と考えられるため

患者要因	統合失調症群 n=3 660	コントロール群 n=350,860	ASD*
女性	3636 (99)	348,818 (99)	0.9
年齢カテゴリー, 歳			
18–55	403 (11)	46,835 (13)	7.2
55–65	753 (21)	78,997 (23)	4.7
65–75	860 (24)	80,726 (23)	1.2
75–85	917 (25)	85,228 (24)	1.8
≥85	727 (20)	59,074 (17)	7.8
Body mass index, kg/m²			
<18.5	399 (11)	31,684 (9.0)	6.3
18.5–25	1792 (49)	222,319 (63)	29.3
25–30	956 (26)	73,101 (21)	12.5
≥30	455 (12)	20,616 (5.9)	22.9
欠損値	58 (1.6)	3140 (0.9)	6.2
喫煙歴あり	852 (23)	70,791 (20)	7.5
Charlson併存疾患指数			
0	2268 (62)	268,132 (76)	31.7
1	683 (19)	38,840 (11)	21.5
≥2	709 (19)	43,888 (13)	18.8

治療・施設要因	統合失調症群 n=3 660	コントロール群 n=350,860	ASD*
癌病期			
0	230 (6.3)	38,168 (11)	16.5
I	929 (25)	143,511 (41)	33.4
II	1550 (42)	119,995 (34)	16.8
III	570 (16)	26,016 (7.4)	25.8
欠損値	611 (17)	61,338 (18)	2.1
乳房全切除	2591 (71)	162,955 (46)	51.0
脳高卵清	1711 (47)	118,025 (34)	27.0
臨床研修指定病院	3290 (90)	311,846 (89)	3.3
Hospital volume			
Low (<70)	1832 (50)	116,216 (33)	34.9
Medium (70–152)	1186 (32)	117,043 (33)	2.0
High (≥152)	642 (18)	117,601 (34)	37.3

*ASD(Absolute standardized difference)＞10%は
差が大きいとみなす

主解析の結果

	統合失調症群		コントロール群		多変量解析	
	n=3660		n=350,860			
	n	(%)	n	(%)	OR	95%CI
在院中合併症	288	(7.9)	16,356	(4.7)	1.37	(1.21–1.55)
術後出血	85	(2.3)	4882	(1.4)	1.34	(1.05–1.71)
創部感染	107	(2.9)	7511	(2.1)	1.22	(1.04–1.43)
	中央値	四分位範囲	中央値	四分位範囲	係数	95%CI
総入院費用, Euro	7204	5941–8970	6133	5444–7259	743	(680–806)

OR, オッズ比; CI, 信頼区間

統合失調症患者で有意に合併症が多く、入院費用も高い

13

感度分析の結果

	1:4 matched-pair cohort解析		統合失調症群を抗精神病薬処方ありに限定		統合失調症群から強制入院を除外	
	OR	95%CI	OR	95%CI	OR	95%CI
在院中合併症	1.25	(1.09–1.43)	1.54	(1.33–1.79)	1.33	(1.17–1.51)
術後出血	1.16	(0.90–1.51)	1.40	(1.04–1.88)	1.31	(1.02–1.69)
創部感染	1.08	(0.88–1.32)	1.35	(1.12–1.62)	1.23	(1.05–1.44)
	係数	95%CI	係数	95%CI	係数	95%CI
総入院費用, Euro	1233	(1047–1418)	1180	(1101–1259)	549	(487–611)

OR, オッズ比; CI, 信頼区間

感度分析でも同様の傾向

14

考察

➢統合失調症患者では合併症が多い
- ✓ 多数の薬剤内服による薬剤相互作用、意思疎通の困難さ
- ✓ 通常通りの医療が提供されていない可能性がある
- ✓ 外科医と精神科医の協働が必要だろう

➢Limitationとして以下の情報が不明
- ✓ 術前化学療法の有無
- ✓ 乳癌の罹患歴
- ✓ 統合失調症の重症度

こうしたデータベースに含まれない情報が重大な交絡でないかは注意が必要

15

外科領域のDPCデータ研究例

1. 統合失調症患者の乳癌手術成績
2. Body mass indexが食道癌手術に及ぼす影響
3. 尾側膵切除における開腹 vs. 鏡視下手術

16

BMIが食道癌手術に及ぼす影響

- "Impact of body mass index on major complications, multiple complications, in-hospital mortality, and failure to rescue following esophagectomy for esophageal cancer: A nationwide inpatient database study in Japan"
- Hirano Y, Kaneko H, Konishi T, Itoh H, Matsuda S, Kawakubo H, Uda K, Hiroki M, Fushimi K, Itano O, Hideo Y, Kitagawa Y
- *Annals of Surgery* 2021 online ahead of print

17

Introduction

➢食道癌手術は高侵襲である
- 合併症や術後死亡が比較的多い
- 合併症後の死亡（failure to rescue）も注目されている

➢BMIと術後短期成績との関連は未だ明らかでない
- 肥満は縫合不全や呼吸器合併症のリスクとの既報あり
- やせは術後死亡のリスクとの既報あり
- Failure to rescueなどとの関連は報告されていない

18

PECO

病名と処置の組み合わせで定義

P：食道癌手術を受けた患者

E：やせ・肥満

C：標準BMI

O：院内死亡、合併症、Failure to rescue、術後在院日数、入院費用

➢主解析：Restricted cubic spline(RCS)を併用した多変量解析

19

対象

様式1の
手術情報で定義

DPCデータベースにおける
2010/7–2019/3の食道癌手術例
（n=40,245）

除外
- ✓ 18歳未満（n=1）
- ✓ 咽喉頭癌手術の併施（n=466）
- ✓ BMIの欠損値/外れ値（n=1318）

対象症例
（n=39,406）

20

RCSを併用した多変量解析

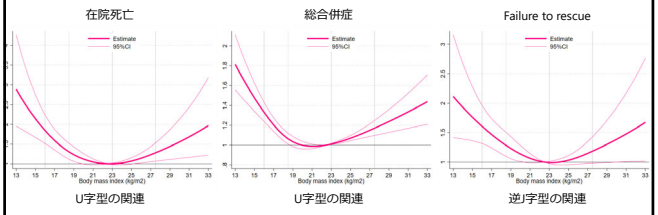
➢背景因子として下記で調整
- 患者要因：性別、年齢、喫煙歴、Charlson併存疾患指数
- 治療要因：癌病期、術前治療、術式
- 施設要因：臨床研修指定の有無、年間手術件数

➢マルチレベル分析を併用
- 患者・治療背景の病院ごとの違いを考慮する

➢RCS解析を併用
- 連続変数（BMI）をカテゴリー化せずに非線形の関連を図示できる

21

結果



U字型の関連　U字型の関連　逆J字型の関連

低BMIでは、合併症が多いだけでなく、
合併症が発生した場合の死亡も多かった

22

考察

➢BMIは食道癌の重要な術前評価項目であるだろう
- ✓ 低BMIは低い呼吸機能・サルコペニアなどの指標とされる
- ✓ 注意深い術後管理によるFailure to rescueの改善が望まれる
- ✓ 術前リハビリや栄養管理が有用かもしれない

癌病期はvalidateされているが、
組織型まではわからない

➢Limitationとして以下の情報が不明
- ✓ 食道癌の組織型
- ✓ 鏡視下手術の保険償還開始（2014年）前の、鏡視下での実施の有無
- ✓ 術前体重減少の有無

保険償還の開始時期の
知識が必要

23

外科領域のDPCデータ研究例

1. 統合失調症患者の乳癌手術成績
2. Body mass indexが食道癌手術に及ぼす影響
3. 尾側膵切除における開腹 vs. 鏡視下手術

24

尾側膵切除における開腹 vs. 鏡視下手術

- “Laparoscopic versus open distal pancreatectomy with or without splenectomy: A propensity score analysis in Japan”
- Konishi T, Takamoto T, Fujiogi M, Hashimoto Y, Hiroki M, Fushimi K, Tanabe M, Seto Y, Hideo Y
- *International Journal of Surgery* 2022;104:106765

25

Introduction

➢膵体尾部腫瘍に対する腹腔鏡下切除が普及しつつある
- 本邦では2012年4月に保険適応になった

➢術後合併症や入院費用への影響はいまだ不明である
- 術後合併症に関しては、既報内で結果が不一致
- 高額な腹腔鏡手術の、入院費用への影響も不明
- 施設因子などの背景因子も調整されていない

26

PICO

P：尾側膵切除を受けた患者

I ：鏡視下手術

C：開腹手術

O：合併症、死亡、再手術、30日再入院、
麻酔時間、ドレーン留置期間、術後在院日数、入院費用

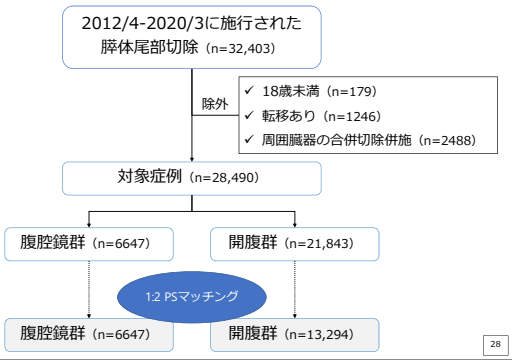
費用はDPCデータベースならではのアウトカム

➢主解析：1:2傾向スコア(PS)マッチング

➢感度分析：PSを用いたオーバーラップ重み付けなど

27

対象



28

1:2 PSマッチング

➢背景因子として下記でPSを作成
- 患者要因：性別、年齢、BMI、喫煙歴、Charlson併存疾患指数
- 治療要因：術直前の血糖降下薬の使用、良性/悪性、癌TN分類、脾臓温存
- 施設要因：臨床研修指定の有無、年間手術件数

➢腹腔鏡群と開腹群が1:2になるようマッチング
- PSを用いた最近傍マッチング

傾向スコア解析については別講義も参照ください

29

感度分析

① PSを用いたオーバーラップ重み付け

操作変数法については別講義も参照ください

② 操作変数法
- ✓ 未測定の交絡（詳細な術式など）を調整できる手法
- ✓ 操作変数として、施設における腹腔鏡手術の施行割合（選好率）を用いた

◆サブグループ解析も実施
- ✓ Low-volume施設を除いた解析 （熟練した施設に限定するため）
- ✓ 良性腫瘍に限った解析 （リンパ郭清の影響を除くため）

30

腹腔鏡群は若い

腹腔鏡群は併存疾患が少ない

Table 1. 背景 患者因子

		全患者（マッチング前） 腹腔鏡 n = 6647	開腹 n = 21,843	標準偏差 (%)	1:2PSマッチ後の患者 腹腔鏡 n = 6647	開腹 n = 13,294	標準偏差 (%)
性別	男性	2669 (40)	12,014 (55)	**30.1**	2669 (40)	5200 (39)	2.1
年齢	<50	1613 (24)	1807 (8.3)	**44.4**	1613 (24)	3487 (26)	4.5
	50–59	1008 (15)	2429 (11)	**12.0**	1008 (15)	1826 (14)	4.1
	60–69	1608 (24)	6164 (28)	9.2	1608 (24)	3260 (25)	0.8
	70–79	1863 (28)	8600 (39)	**24.2**	1863 (28)	3692 (28)	0.6
	≥80	555 (8.3)	2843 (13)	**15.1**	555 (8.3)	1029 (7.7)	2.2
Body mass index, kg/m²	<18.5	647 (9.7)	2623 (12)	7.3	647 (9.7)	1261 (9.5)	0.8
	18.5–21.9	2339 (35)	7955 (36)	2.6	2339 (35)	4807 (36)	2.0
	22.0–24.9	2006 (30)	6707 (31)	1.1	2006 (30)	4078 (31)	1.1
	25.0–29.9	1351 (20)	3851 (18)	6.9	1351 (20)	2645 (20)	1.1
	≥30.0	282 (4.2)	571 (2.6)	9.0	282 (4.2)	469 (3.5)	3.7
	欠損値	22 (0.3)	136 (0.6)	4.2	22 (0.3)	34 (0.3)	1.4
喫煙症	あり	2636 (40)	9919 (45)	**11.6**	2638 (40)	5301 (40)	0.4
Charlson Comorbidity index	2	5284 (79)	10,651 (49)	**67.6**	5284 (79)	10,754 (81)	3.5
	3	820 (12)	6617 (30)	**44.9**	820 (12)	1559 (12)	1.9
	4	347 (5.2)	3182 (15)	**31.7**	347 (5.2)	691 (5.2)	0.1
	≥5	196 (2.9)	1393 (6.4)	**16.3**			0.9

Data are presented as n and (%).

背景が調整できている

31

腹腔鏡群は
- 良性が多い
- 悪性でも初期

腹腔鏡群は
- High-volume施設で多い
- 近年増加している

Table 1. 背景 治療・施設因子

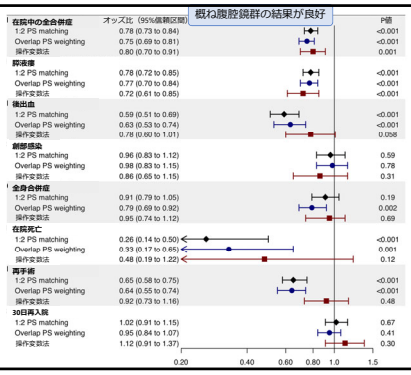
		全患者（マッチング前） 腹腔鏡 n = 6647	開腹 n = 21,843	標準偏差 (%)	1:2PSマッチ後の患者 腹腔鏡 n = 6647	開腹 n = 13,294	標準偏差 (%)
血糖降下薬の術直前使用	インスリン	638 (10)	3894 (18)	**24.1**	638 (9.6)	1070 (8.0)	5.5
	その他	482 (7.3)	2294 (11)	**11.4**	482 (7.3)	737 (5.5)	7.0
術前診断	悪性	1969 (30)	16,069 (74)	**97.9**	1969 (30)	3771 (28)	2.8
	T2以上	882 (13)	10,742 (49)	**84.1**	882 (13)	1783 (13)	0.4
	N1以上	168 (2.5)	3232 (15)	**44.7**	168 (2.5)	325 (2.4)	0.5
術式	脾温存	924 (14)	1052 (4.8)	**31.6**	924 (14)	1417 (11)	9.9
臨床研修指定	あり	4692 (71)	14,413 (66)	9.9	4692 (71)	9491 (71)	1.8
Hospital volume, 件/年	<7	1001 (15)	7956 (36)	**50.4**	1001 (15)	2037 (15)	0.7
	7–13	2459 (37)	7134 (33)	9.2	2459 (37)	5038 (38)	1.9
	≥14	3187 (48)	6763 (31)	**35.3**	3187 (48)	6219 (47)	2.3
手術年度	2012–2015	2322 (35)	10,630 (49)	**28.1**	2322 (35)	5028 (38)	6.0
	2015–2019	4325 (65)	11,213 (51)	**28.1**	4325 (65)	8266 (62)	6.0

背景が調整できている

Data are presented as n and (%).

32

概ね腹腔鏡群の結果が良好



サブグループ解析でも同様の結果であった

33

連続変数のアウトカム	PS解析 1:2マッチング 差	95%信頼区間	P値	Overlap weighting 差	95%信頼区間	P値	操作変数法 差	95%信頼区間	P値
麻酔時間, 分	59	(56–63)	<0.001	59	(55–63)	<0.001	77	(70–83)	<0.001
ドレナージ期間, 日	-4.0	(-4.5to -3.6)	<0.001	-4.6	(-5.1 to -4.0)	<0.001	-2.8	(-3.9 to -1.8)	<0.001
術後在院期間, 日	-4.4	(-4.9 to -3.9)	<0.001	-5	(-5.6 to -4.5)	<0.001	-2.5	(-3.6 to -1.4)	<0.001
入院費用, 円	-138,394	(-166,750 to -110,037)	<0.001	-195,294	(-227,002 to -163,586)	<0.001	-49,744	(-115,191 to 15,702)	0.14

腹腔鏡群で
- 麻酔時間は約1時間長い
- ドレナージ期間・在院期間は短い
- 入院費用も安い

サブグループ解析でも同様の結果であった

34

考察

➢腹腔鏡では膵液瘻や後出血が少ない
- ✓ 死亡や入院期間も減少
- ✓ 高額な術式のはずだが、入院費用も減少していた
- ✓ 麻酔時間は1時間長いが、それは許容可能だろう

入手不能な因子はあるが、代替変数で評価したり、疫学的手法で対処したりする

➢Limitationとして以下の情報が不明
- ✓ 胃排泄遅延　（代わりに入院期間で評価可能か）
- ✓ 術者情報　（代わりに施設因子で調整した）
- ✓ 腹腔動脈合併切除　（操作変数法で調整し、サブグループ解析も行った）

35

小括

➢治療などの効果判定に関する外科領域のDPCデータ研究例を紹介した

➢多変量解析に種々の臨床疫学的手法を併用した
- マルチレベル分析、RCS分析、傾向スコア、操作変数法など

➢信頼性を高める様々な工夫を行った
- 病名と処置・薬剤処方との組み合わせ
- 代替変数の活用
- 感度分析の実施

36

おわりに

- DPCデータを用いた救急・外科領域の研究例を提示した
- 病名・処置・処方データを活用した様々な研究が行えた
- DPCに含まれるデータの内容に注意して解析を行った

37

NDB・DPCデータベース研究人材育成Webinar

DPCデータを用いた研究2

東京大学大学院医学系研究科臨床疫学・経済学
大邉寛幸

DPCデータの利点

- 大規模、N数は桁外れ
  - –症例数やアウトカムが稀な研究が実施可能
- 母集団代表性に優れる
  - –日本の急性期病院の50%、ICU・HCUの70-80%、３次救急病院の90%をカバー
- 実臨床を反映している
  - –ランダム化が方法論や倫理的に不可能な臨床課題を解決

DPCデータの欠点

- 患者が病院を変えるともはや追跡できない
  - –長期生存の追跡には向いていない
- バイタルや、検査結果データが無い
  - –Risk adjustmentが いつも十分にできるわけではない
- 診断病名の妥当性が中程度
  - –主要な病名（心不全など）の感度78.9%、特異度93.2%

Yamana H, et al. J Epidemiol. 2017;27:476-482

DPCデータに適した研究

➢DPCデータは以下の研究に適している
- 症例数が膨大な点を活かす
  - ✓希少な病態・介入の記述疫学研究
  - ✓（倫理的・経済的・実務的に）RCTが困難な比較研究
- 入院データである点を活かす
  - ✓入院を要する病態・介入の研究
  - ✓入院中に重要なアウトカムが発生する研究

➢救急領域や外科領域などの急性疾患を扱いやすい

4

臨床疫学研究の内容による分類

1．治療・予防などの効果判定
　⇒RCTはこれに含まれるが、RCTだけでない
2．診断法や患者評価法
3．疾病のリスク因子の同定
4．疾病の予後予測
5．稀少疾患の記述疫学
6．診療実態の分析(practice pattern analysis)
7．医療の質研究、ヘルス・サービス・リサーチ
　などなど

救急・集中治療領域のDPCデータ研究例

1．稀少疾患の記述疫学
　コリン作動性クリーゼの記述疫学

2．診療実態の分析(practice pattern analysis)
　日本のICU病床稼働率の実態

3．医療の質研究、ヘルス・サービス・リサーチ
　「重症度、医療・看護必要度」ICU入室基準の導入が与えた影響

救急・集中治療領域のDPCデータ研究例

1．稀少疾患の記述疫学
　コリン作動性クリーゼの記述疫学

2．診療実態の分析(practice pattern analysis)
　日本のICU病床稼働率の実態

3．医療の質研究、ヘルス・サービス・リサーチ
　「重症度、医療・看護必要度」ICU入室基準の導入が与えた影響

論文掲載報告

ORIGINAL ARTICLE

Cholinergic Crisis Caused by Cholinesterase Inhibitors: a Retrospective Nationwide Database Study

Hiroyuki Ohbe[1] · Taisuke Jo[1] · Hiroki Matsui[1] · Kiyohide Fushimi[2] · Hideo Yasunaga[1]

Received: 7 January 2018 / Revised: 3 May 2018 / Accepted: 31 May 2018 / Published online: 15 June 2018

1. *Journal of Medical Toxicology (2018) 14:237–241*
https://doi.org/10.1007/s13181-018-0669-1

研究例1：稀少疾患の記述疫学

Clinical Question

“コリン作動性クリーゼ”という希少疾患についてコリンエステラーゼ阻害薬による重篤な副作用で呼吸不全を来たすとされている。
しかし既存の報告は10数例程度のケースシリーズが報告されているのみであり予後などは不明。

患者抽出条件

- 病名のいずれかに以下のICD10コードを含む
  T44.0（主として自律神経系に作用する薬物による中毒，コリンエステラーゼ阻害薬）
- その患者のうち“コリン作動性クリーゼ”と日本語で傷病名に記載がある

病名くん(ICD10コード)

http://www.byomei.org/byomei-kun.2.0/

病名くん(ICD10コード)

ICD10 国際疾病分類第10版（2013年版）

大分類（章）一覧

章	ICDコード	分類見出し
17	Q00-Q99	先天奇形,変形および染色体異常
18	R00-R99	症状,徴候および異常臨床所見・異常検査所見で他に分類されないもの
19	S00-T98	損傷,中毒およびその他の外因の影響
20	V01-Y98	傷病および死亡の外因
21	Z00-Z99	健康状態に影響をおよぼす要因および保健サービスの利用
22	U00-U99	特殊目的用コード

http://www.byomei.org/byomei-kun.2.0/

病名くん(ICD10コード)

T44　主として自律神経系に作用する薬物による中毒

- T44.0　主として自律神経系に作用する薬物による中毒，コリンエステラーゼ阻害薬
- T44.1　主として自律神経系に作用する薬物による中毒，その他の副交感神経興奮薬[コリン作動薬]
- T44.2　主として自律神経系に作用する薬物による中毒，神経節遮断薬，他に分類されないもの

http://www.byomei.org/byomei-kun.

結果

- DPCデータにおいて2010年7月から2016年3月の期間中、235人のコリン作動性クリーゼの患者が同定された。
- 20%の患者が人工呼吸器を必要とし、院内死亡率は6.4%であった。
- 入院患者の半数以上がカテコラミン投与を必要とした。

Ohbe H, et al. J Med Toxicol. 2018;14:237-241

限界

- 患者抽出条件の感度・特異度が不明。
- 原因となるコリンエステラーゼ阻害薬をいつどのくらい内服したのか不明。
- コリン作動性クリーゼの診断基準に含まれる症状、バイタル、身体所見、血液検査値のデータが無い。

救急・集中治療領域のDPCデータ研究例

1．稀少疾患の記述疫学
　コリン作動性クリーゼの記述疫学

2．診療実態の分析(practice pattern analysis)
　日本のICU病床稼働率の実態

3．医療の質研究、ヘルス・サービス・リサーチ
　「重症度、医療・看護必要度」ICU入室基準の導入が与えた影響

論文掲載報告

Journal of Epidemiology

Original Article　J Epidemiol 2022

Intensive Care Unit Occupancy in Japan, 2015–2018: A Nationwide Inpatient Database Study

Hiroyuki Ohbe[1], Yusuke Sasabuchi[2], Ryosuke Kumazawa[1], Hiroki Matsui[1], and Hideo Yasunaga[1]

[1]Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo, Tokyo, Japan
[2]Data Science Center, Jichi Medical University, Tochigi, Japan

1. *J Epidemiol 2021 Apr 10. doi: 10.2188/jea.JE20210016.*

研究例2：診療実態の分析

- **人口10万人当たりICUベッド数**は既存によく使われている供給能力を示す指標の一つであり、日本は5.6床と他先進国に比べ少ないため供給が足りないと議論されている[2]。
- **ICUの病床稼働率**もその地域におけるCritical Care Systemの供給能力を示す指標の一つ[1]。
- 過去に日本の**ICU病床稼働率**は検証されたことは無い。

1. *Intensive Care Med. 2019;45:1231-1240.*
2. *J Japanese Soc Intensive Care Med. 2010;17:227-232.*

方法

- 記述疫学
- 厚生労働科学研究調査研究班のDPCデータ + 病床機能報告2017
- 観察期間：2015年1月1日～2018年12月31日
- 対象患者：ICU入室した全ての患者
  - ICU定義: 1:2看護患者比率の病床

主要評価項目

- ICU稼働率を以下のように算出

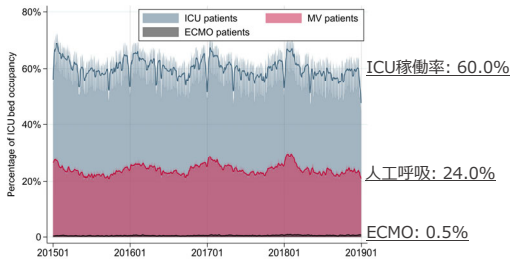
$$\text{ある日のA病院のICU稼働率(\%)} = \frac{\text{ある日のA病院のICU管理料取得人数の合計(DPC)}}{\text{A病院の2017年許可 ICU 病床数(病床機能報告)}} \times 100$$

- 人工呼吸器,ECMOの患者の稼働率も算出

結果

- 137万9618人のICU患者(495病院, 5341のICU病床) が同定された。
- これは2017年度の日本の全ICU病床の75%(N=5341/7109) を含んでいた。

結果 –ICU稼働率–



考察 –先行研究との比較-

	日本	米国
ICU稼働率	60%	70%
人工呼吸	24%	30%
10万人対ICU病床数	5.6床	20床

- 10万人対ICU病床数が4倍多い米国と比較しても、ICU稼働率並びに人工呼吸稼働率は日本は低い傾向にあった[1]。
- これまでの欧米諸国の声明ではICU稼働率は70-75%が最適であろうと推奨されている。

1. *Crit Care Med. 2013;41:2712-2719.*
2. *Crit Care Med. 2009;37:2753-2758.*
3. *Aust Crit Care. 2014;27:77-84.*

考察 -限界-

- ICU算定を受けれないICU長期滞在患者（14日以降）を無視したため、稼働率が低く見積もられている

救急・集中治療領域のDPCデータ研究例

1．稀少疾患の記述疫学
　コリン作動性クリーゼの記述疫学

2．診療実態の分析(practice pattern analysis)
　日本のICU病床稼働率の実態

3．医療の質研究、ヘルス・サービス・リサーチ
　「重症度、医療・看護必要度」ICU入室基準の導入が与えた影響

論文掲載報告

Associations of Government-issued Intensive Care Unit Admission Criteria with Clinical Practices, Outcomes, and Intensive Care Unit Bed Occupancy

Hiroyuki Ohbe[1], Tadahiro Goto[1,2], Hiroki Matsui[1], Kiyohide Fushimi[3], and Hideo Yasunaga[1]

[1]Department of Clinical Epidemiology and Health Economics, School of Public Health, University of Tokyo, Tokyo, Japan; [2]TXP Medical Co. Ltd., Tokyo, Japan; and [3]Department of Health Policy and Informatics, Graduate School of Medicine, Tokyo Medical and Dental University, Tokyo, Japan

ORCID IDs: 0000-0001-8544-2569 (H.O.); 0000-0002-5880-2968 (T.G.); 0000-0003-0004-4743 (H.M.); 0000-0002-1894-0290 (K.F.); 0000-0002-6017-469X (H.Y.).

1. *Ann Am Thorac Soc. 2021 Nov 23. doi: 10.1513/AnnalsATS.202107-844OC.*

3. 医療の質研究、ヘルス・サービス・リサーチ

- 日本政府は2014年4月から「重症度・医療・看護必要度」という新たなICU入室基準を医療政策として導入[2]
- 各ICUは表の基準を満たす患者を**70-90%**以上受け入れなくてはならないと定められた。

モニタリング及び処置	2014-15年	2016-17年
1 心電図モニター	1点	1点
2 動脈圧測定	1点	2点
3 中心静脈圧測定	1点	2点
4 肺動脈圧測定	1点	2点
5 輸液ポンプ	1点	1点
6 シリンジポンプ	1点	1点
7 人工呼吸器	1点	2点
8 輸血・血液製剤	1点	2点
9 特殊な治療法	1点	2点
基準を満たす合計点数	≥3点以上	≥4点以上

3. 医療の質研究、ヘルス・サービス・リサーチ

- 「重症度・医療・看護必要度」は**不必要なICU入室を減らす**目的で策定された[1]。
- 本基準は医療者の診療行為に基づく指標であり、他国の患者重症度に基づく指標(SOFAなどの臓器障害スコア) と大きく異なっていた[2,3]。

1. 日本医療・病院管理学会誌　*2008;45:37-48*
2. *Ann Intensive Care 2013;3:37.*
3. *Curr Opin Crit Care 2009;15:591–596.*

方法

- 分割時系列デザイン
- 厚生労働科学研究調査研究班のDPCデータ + 病床機能報告2017
- 観察期間：2012年4月～2018年3月

PECO

P　全ICU入室患者（1:2看護患者比率）

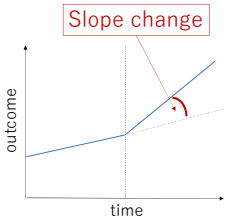
E　政策導入後4年間（2014年4月から2018年3月）

C　政策導入前2年間（2012年4月から2014年3月）

O　1. ICUベッド稼働率
　　2. モニタリングおよび処置
　　3. 臨床アウトカム(院内死亡, 合併症, 入院期間, 費用)

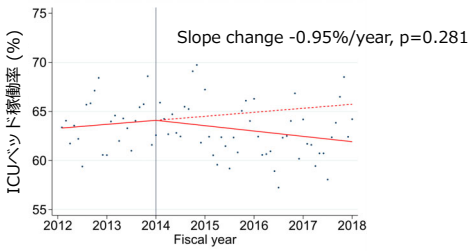
統計解析

- 分割時系列デザイン[1]
- Impact modelは傾き変化のSlope changeを仮定

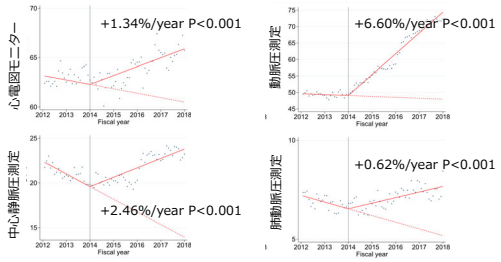


1. *Int J Epidemiol 2017;46:348–355.*

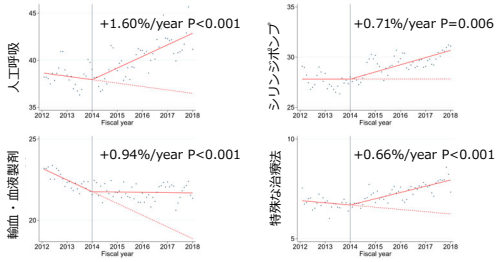
結果 –ICUベッド稼働率–



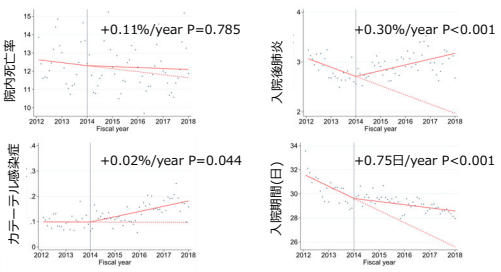
結果 –モニタリング及び処置①–



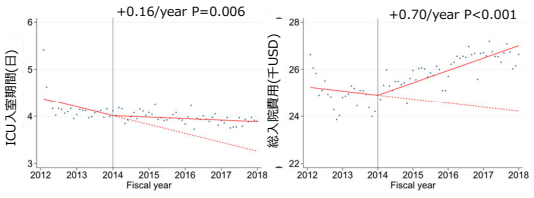
結果 –モニタリング及び処置②–



結果 –臨床アウトカム①–



結果 –臨床アウトカム②–



考察 –Study implication-

- 分割時系列デザインによる医療政策の評価を行った。
- 医療者の診療行為に基づくICU入室基準は、ICU病床稼働率の減少とは関連せず、ICUでの診療行為・合併症・入院期間・入院費の増加と関連していた。

考察 -限界-

- 同時期に発生したイベントの影響。
- 重症度スコアなどの未測定時間依存性交絡因子。
- 他国への外的妥当性。

小括

以下の救急・集中治療領域のDPCデータ研究例を紹介した

1. 稀少疾患の記述疫学
2. 診療実態の分析
3. 医療の質研究、ヘルス・サービス・リサーチ

急性期医療に関する母集団代表性は高く、DPCデータを用いた類似の研究立案の参考になれば幸いです。

NDB・DPCデータベース研究人材育成Webinar

DeSCデータベースを用いた研究

東京大学大学院医学系研究科臨床疫学・経済学
康永秀生

本講義のトピック

1．はじめに
2．DeSCデータベースの概要
3．DeSCデータベースの母集団代表性
4．DeSCデータベースを用いた保険者種類別の有病率推計
5．DeSCデータを用いた研究例
6．まとめ

1．はじめに

リアルワールドデータ(real-world data, RWD)の類型

- 患者登録(patient registry)
- 保険データベース(administrative claims database)
- 電子カルテ情報(electronic medical records)

など

RWD研究が隆盛するであろう理由

(i)莫大な資金をかけて多数のRCTをやれるほどの財力が産・官・学のいずれにもない。
(ii)RCTの対象から外される高齢患者が増加してきた。
(iii)治療の選択肢も患者のニーズも多様化し、すべてをRCTで検証することは不可能。
(iv)限られた答えしか得られないRCTでは、「エビデンスの隙間」を埋められない。

RWDの利点

- 疾患の疫学データなどがわかる。
- 薬だけでなく、手術・処置などあらゆる治療について検討可能。
- 臨床試験では分からない、実臨床における治療効果を明らかにできることもある。
- 薬剤の費用効果分析などにも応用可能。

RWDの欠点

- 適応による交絡(confounding by indication)を十分に調整しきれない。
- 入力されているデータの妥当性にやや難がある。

2. DeSCデータベースの概要

- DeSCデータベース：DeSCヘルスケア株式会社が、データ利活用事業の一環として、いくつかの健保・国保・高齢者医療広域連合より提供されたレセプト情報・特定健診情報の匿名加工情報 。
- データベースに含まれる総人数は、2022年10月時点で約1110万人。

レセプト情報

被保険者台帳データ

匿名化ID、生年月、性別、保険者種別、本人または家族の別、
レセプト開始・終了年月、資格取得・喪失年月など

医科・歯科・調剤の各レセプト

傷病名（ICD-10コードおよび日本語テキスト）
診療行為とその算定日
医薬品情報（医薬品名称とその算定日、量）

特定健診情報

身長・体重、腹囲、血圧、血液・尿検査、保健指導レベル、
生活歴・現病歴に関する質問票の情報など

7

3．DeSCデータベースの母集団代表性

Okada, A, Yasunaga H. Prevalence of Non-communicable Diseases in Japan Using a Newly Developed Administrative Claims Database Covering Young, Middle-aged, and Elderly People. *JMA Journal* 2022;5(2):190-198

- 健保の被保険者はサラリーマン等、国保の被保険者は自営業者や退職者等であり、両者は年齢分布、年収を含む社会経済状況、健康状態が異なっていると考えられる。
- 健保、国保および後期高齢者のいずれか単独のデータでは、母集団から乖離した標本になる。

8

DeSCデータと他の統計情報を比較し、DeSCデータの母集団代表性を検証

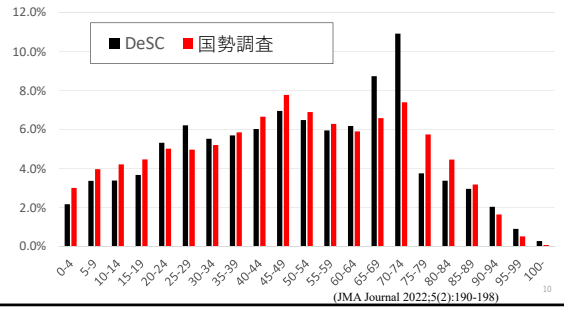
- 集団の年齢分布を、DeSCデータと国勢調査の間で比較
- 糖尿病・高血圧の有病割合を、DeSCデータと国民健康・栄養調査の間で比較
- 胃がん手術の実施割合を、 DeSCデータとNDBオープンデータの間で比較

DeSCデータの中で、健保と国保を比較

- うつの年齢階級別有病割合
- 乳がんの年齢階級別有病割合

9

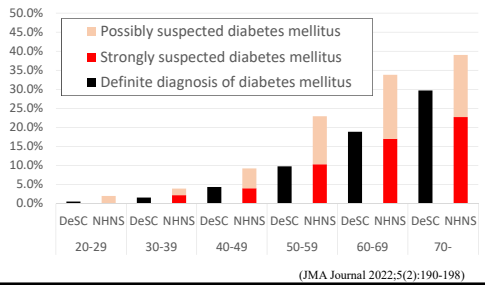
DeSCデータベースと国勢調査（2019年）の年齢分布比較



(JMA Journal 2022;5(2):190-198)

10

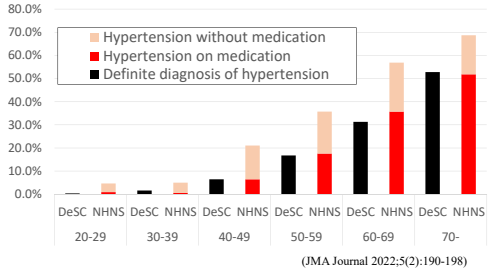
DeSCデータベースと国民健康・栄養調査(NHNS)の比較
～糖尿病の有病割合



(JMA Journal 2022;5(2):190-198)

11

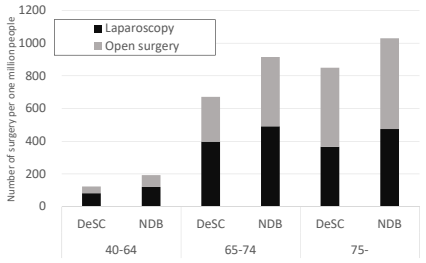
DeSCデータベースと国民健康・栄養調査(NHNS)の比較
～高血圧の有病割合



(JMA Journal 2022;5(2):190-198)

12

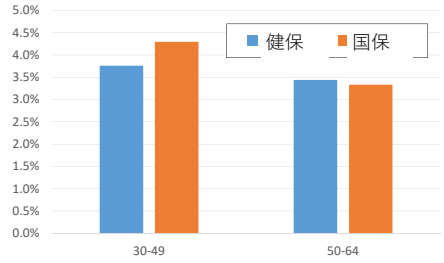
DeSCデータベースとNDBオープンデータの比較
～胃がん手術の100万人当たり実施件数



(JMA Journal 2022;5(2):190-198)

13

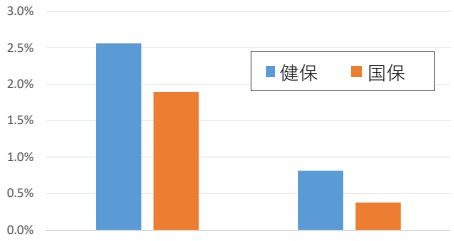
DeSCデータベース内での健保データと国保データの比較
～うつの有病割合



(JMA Journal 2022;5(2):190-198)

14

DeSCデータベース内での健保データと国保データの比較
～乳がんの有病割合



(JMA Journal 2022;5(2):190-198)

15

DeSCデータベースと国勢調査では、年齢分布がほぼ類似。
（65-74歳の年齢層のみ、DeSCデータベースにおける割合がやや高かった）

DeSCデータベースと国民健康・栄養調査では、糖尿病・高血圧の年齢階級別の有病割合がほぼ類似。

DeSCデータベースとNDBオープンデータでは、年齢階級別の胃がん手術実施件数（人口百万人当たり）がほぼ類似。

うつの有病割合は、30-49歳では健保より国保の方が高い傾向。
乳がんの有病割合は、 30-64歳で健保より国保の方が低い傾向。

16

4. DeSCデータベースを用いた保険者種類別の有病率推計

岡田 啓, 康永 秀生. DeSC データベースの概要と臨床疫学・薬剤疫学研究への活用. 薬剤疫学 2022;27(1)11-18

- 健保と国保に含まれる集団は社会経済状況等が異なり、健康状態も異なると考えられる。しかし、両集団における疾病の有病率や健康指標の差異について、単一のデータベースを用いて比較した研究はほとんどない。
- 本稿では、DeSCデータベースを用いて、健保・国保・後期高齢者の保険者種類別に、糖尿病・高血圧・心血管疾患・悪性腫瘍等の有病率を推計し、さらに特定健診で得られるBody Mass Index (BMI)と収縮期血圧の分布を記述することとした。

17

方法

データソース：2021年6月時点でのDeSCデータベース
対象：2019年10月から2020年9月までの20歳以上の男女（N=1,685,438）
年齢層別(20歳代, 30歳代, 40歳代, 50歳代, 60歳代, 70歳以上)、
保険者種類別に、1年有病率を算出。

保険者種類別の構成：
健保669,215人（39.7%）、国保715,750人（42.5%）、後期高齢者300,473人（17.8%）

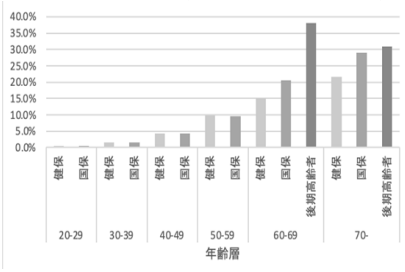
対象とする疾患：
糖尿病（ICD-10コード：E10-14）、高血圧（I10-15）、虚血性心疾患（I20-25）、心不全（I50）、脳卒中（I60-63）、大腸がん（C18-20）

BMIと血圧の値が欠測でない30歳以上の対象者（N=651,435人）について、特定健診情報を用いて、BMIと収縮期血圧の性・年齢層別の分布を保険者種類別に記述。

18

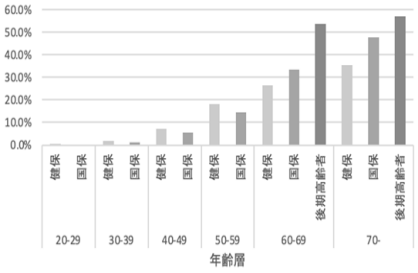
結　果

糖尿病の1年有病率（保険者種別・年齢層別）



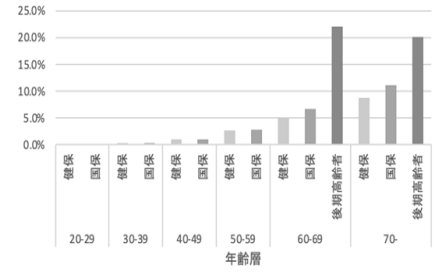
19

高血圧の1年有病率（保険者種別・年齢層別）



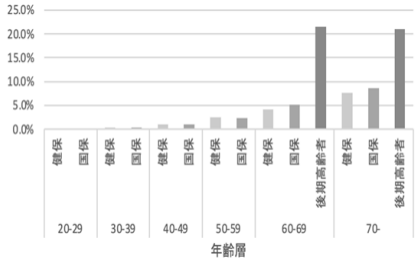
20

虚血性心疾患の1年有病率（保険者種別・年齢層別）



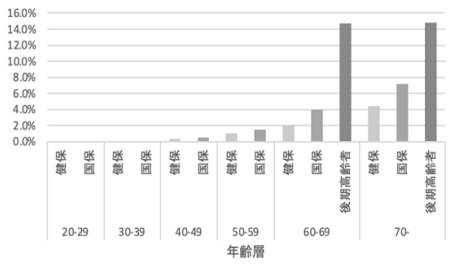
21

心不全の1年有病率（保険者種別・年齢層別）



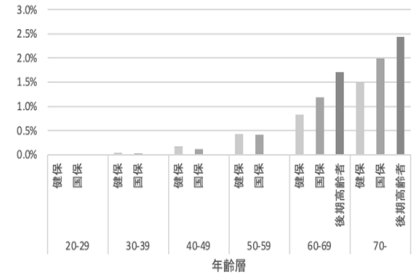
22

脳卒中の1年有病率（保険者種別・年齢層別）



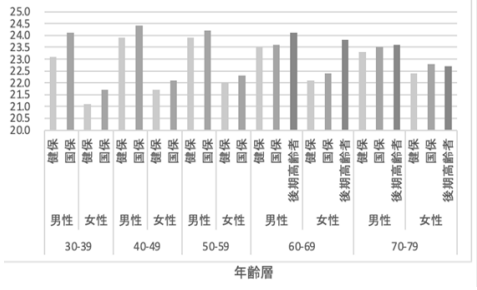
23

大腸がんの1年有病率（保険者種別・年齢層別）



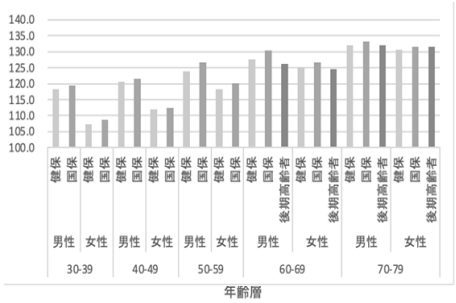
24

性別・年齢層別・保険者種別のBMIの平均値



25

性別・年齢層別・保険者種別の収縮期血圧の平均値



26

考　察

- 糖尿病・高血圧の有病率は、59歳までの年齢層では健保よりも国保の方が低くなっており、60歳以上の層では逆に健保よりも国保の方が高くなっていた。これは、健保の方が特定健診を受診する割合が高く、59歳までの糖尿病・高血圧が受診を契機として発見されやすいことを反映しているかもしれない。
- 虚血性心疾患・心不全の有病率は、59歳までは健保と国保で同程度、60歳以上では、健保よりも国保の方が高い傾向が認められた。
- 脳卒中の有病率は、いずれの年齢層においても、健保よりも国保の方が高かった。

27

- 大腸がんの有病率は、59歳までは健保よりも国保の方がわずかに低く、60歳以上では健保よりも国保の方が高かった。
- 特定健診を受診した集団内の分析において、同じ性・年齢層の中でも、BMI・収縮期血圧がわずかながら健保よりも国保の方が高い傾向が示された。
- 上記のいずれも、国保の被保険者は健保の被保険者よりも健康状態が良くないことを示唆する。

28

5．DeSCデータを用いた研究例

骨盤臓器脱に対する腹腔鏡下およびロボット支援下仙骨膣固定術後の有害事象と再治療

Shigemi D, Okada A, Yasunaga H. Postoperative adverse events and re-treatment among patients who have undergone laparoscopic and robotic sacrocolpopexy for pelvic organ prolapse in Japan. *Int J Gynaecol Obstet*. 2022. PMID: 36200666

骨盤臓器脱（pelvic organ prolapse, POP）は女性に比較的よく見られる疾患であり、近年合併症の少ない腹腔鏡下やロボット支援下仙骨膣固定術が行われている。
この研究の目的は、日本人のPOP患者における腹腔鏡下およびロボット支援下仙骨膣固定術後の有害事象と再発に対する再治療を記述し比較することである。

29

方　法

データソース：DeSCデータベース　2014年4月から2021年5月まで
対象：腹腔鏡下またはロボット支援下仙骨膣固定術を受けた患者

患者背景と術後アウトカムを腹腔鏡下仙骨膣固定術群とロボット支援下仙骨膣固定術群で比較。
アウトカム：
複合有害事象（膣びらん、術後尿失禁、術後排尿困難、尿路傷害、腹部切開ヘルニア）
POP再発に対する再治療（ペッサリーの使用とPOPに対する手術）

30

結　果

POPと診断された患者28,748人
　腹腔鏡下仙骨膣固定術　409人（1.4%）
　ロボット支援下仙骨膣固定術　52人（0.2%）

術後有害事象の割合
　腹腔鏡下仙骨膣固定術群　20.8%
　ロボット支援下仙骨膣固定術群　13.5%（P＝0.27）

ロボット支援下仙骨膣固定術の後に手術を受けた患者は1名（1.9%）であり，両群とも術後にペッサリーを投与された患者はいなかった。

31

結　論

本邦において、腹腔鏡下およびロボット支援下仙骨膣固定術の術後アウトカムは、欧米諸国の報告と同程度であり、遜色はなかった。

32

6．まとめ

- DeSCデータベースは、健保・国保・後期高齢者のすべてのデータを含んでおり、小児から後期高齢者までの全年齢層をカバーし、母集団代表性が一定程度担保されている。
- DeSCデータベースを利用することにより、NDB以外では難しかった、全年齢層の外来・入院を含めたレセプト情報・特定健診情報を用いた臨床疫学・薬剤疫学研究が可能と考えられる。

33

ご清聴ありがとうございました

34

NDB・DPCデータベース研究人材育成Webinar

JMDCデータを用いた研究

東京大学医学部循環器内科（先進循環器病学講座）

金子 英弘

JMDC Claims Databaseとは？

- 株式会社JMDCが提供する保険者（健保）から集められた保険レセプトデータベース
- データ入手先:　保険者（主に大企業）
- 収集期間: 2005-現在（毎月更新）
- 累積母集団数:約1400万人（2022年2月時点）
  (https://www.jmdc.co.jp/jmdc-claims-database/)
- データ種別: 外来、入院、調剤レセプトデータ、（特定）健診データ、被保険者台帳、気象情報(集計データのみ)

Agenda

1. JMDCデータの概要
2. 解析で注意すべき点
3. 研究紹介

JMDC Claims Databaseの特徴

- 20-65 歳のデータが最も多い（男性が女性より若干多い）
- 家族IDがある
- 小児のデータも含まれる
- 高齢者のデータが乏しい（最高齢75歳）
- 数年の経過を追跡可能（平均観察期間は3.5年程度）
- 保険から離脱しない限り複数の医療機関を受診しても追跡可能

Agenda

1. JMDCデータの概要
2. 解析で注意すべき点
3. 研究紹介

JMDC Claims Databaseに含まれる健診データ

- 問診（特定健診の問診票）
  喫煙歴、飲酒歴、運動習慣、睡眠の質など
- BMI、腹囲（特定健診受診者）　※身長・体重データは含まれない
- 血圧　※脈拍は含まれない
- 血液検査
  血糖、HbA1c、脂質プロファイル、尿酸、クレアチニンなど
- 尿検査（尿糖、尿蛋白定性）
- 眼底検査

※項目ごとに欠損値が存在すること、また欠損値が存在する比率は項目毎に異なる点に注意が必要。

JMDC Claims Databaseの利点と限界

利点
- 外来レセプト、外来処方、調剤薬局データが存在する
- 病院を変更しても追跡可能
- 特定健診データが存在する
- 保険者台帳が存在する
- 家族IDが存在する

限界
- DPCの様式１情報がない
- 疾病の重症度指標がない
- 通常診療の検査値がない
- 高齢者の解析には向き
- SESの高い集団に偏っている
- 観察期間が比較的短い

注意すべき点（２）リスク調整

- 処方や処置、手術など保険算定される因子についてはデータが充実している
- 傷病名の妥当性については更なるValidation研究が必要
- 個々の疾病の重症度や進行度（例. がんのステージ）については情報が乏しい（JMDCが提供するDPCデータには一部存在）
- 検査値は健診時のみデータが存在する

Agenda

1. JMDCデータの概要
2. 解析で注意すべき点
3. 研究紹介

傷病名の妥当性

	n	Prevalence	感度(%)	特異度(%)
高血圧	286,139	23.7		
病名コード			80.7	95.3
薬コード			75.0	97.9
病名＋薬			74.5	98.2
糖尿病	52,014	8.3		
病名コード			91.4	92.5
薬コード			78.6	99.5
病名＋薬			78.6	99.6
脂質異常	277,707	38.4		
病名コード			48.9	90.4
薬コード			34.6	97.2
病名＋薬			34.5	97.2

J Clin Epidemiol. 2018 Jul;99:84-95.

注意すべき点（１）データハンドリング

- リレーショナルデータベース
- 様々なデータが複数のテーブルで存在
- 必要な変数を自分で集める
- 統計ソフトだけでなくSQLのスキルが必要

加入者ID	受診日	病名コード	・・・
0001	20151103	J09	
0002	20151010	J10	
0003	20141123	J11	

加入者ID	処方日	薬剤コード	・・・
0001	20151103	10067130	
0005	20151009	10002295	
0007	20141023	10067130	

リスク調整（対応策）

- 処方・処置・手術などの保険点数データの活用
  - 高血圧病名→降圧薬処方
  - ショック病名→カテコラミンの使用量
  - 出血量→輸血料
  - 呼吸不全→気管内挿管、人工呼吸器管理
  - 腎不全病名→透析
  - 手術時間→麻酔時間
  - リハの量→20分毎に点数加算
  - 腹腔ドレーン留置期間→ドレーン法 50点/日

リスク調整（対応策）

- ICD10コードや処置コードを用いた重症度指標
  - Charlson Comorbidity Index
    –ICD-10コードを用いた併存症指数
  - Yamana Severity Index
    –処置データを用いた重症度指数
  - ICD-10-based trauma mortality prediction scoring system
    –ICD-10コードで外傷患者の死亡率を予測するスコア
  - ICD-10-Based Disability Predictive Index
    –ICD-10コードで外傷患者の障害を予測する指標

注意すべき点（５）Dropout

- JMDCは保険離脱によって追跡不能となる
  - 退職（病気、転職、定年、etc.）
  - 75歳
  - 被保険者である配偶者と離婚
  - 被保険者の子が就職
- 病気の重症化などによる退職が解析上は問題になる

→Censoring Weight（打ち切りの重みづけ）などで対応

注意すべき点（３）交絡因子の調整

- 生物統計学的対応
  - 層別化
  - 多変量回帰分析
  - 傾向スコア分析（重症度のバランシング）
  - 操作変数法（疑似ランダム化）
  - 固定効果モデル（時間不変の交絡除去）
  - 自己対象研究デザイン（時間不変の交絡除去）

Agenda

1. JMDCデータの概要
2. 解析で注意すべき点
3. 研究紹介

注意すべき点（４）欠損値の取り扱い

- 欠損値のメカニズムの検討(MCAR, MAR, MNAR)
- Complete case analysis
- 単一代入法
- 多重代入法

JMDCはどんな研究に向いているか？

一般的に

“登録型の疫学コホートやRCTでは解決が困難な課題の検討”

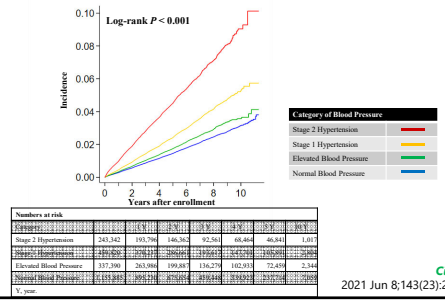
にリアルワールドデータは強みを有します。

- 豊富なサンプル数・イベント数が必要な研究
- 悉皆性のあるデータが必要な研究
- 薬剤疫学研究

JMDCはどんな研究に向いているか？

- 豊富なサンプル数・イベント数が必要な研究
- 悉皆性のあるデータが必要な研究
- 薬剤疫学研究

血圧分類と共に心不全発症リスクは上昇する



Circulation. 2021 Jun 8;143(23):2244-2253.

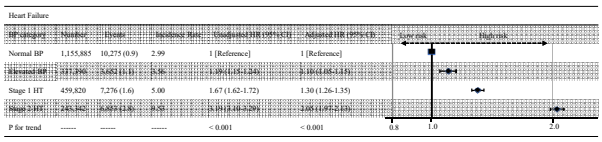
JMDCデータを用いた研究紹介（1）

背景：
2017年に米国の血圧ガイドラインが改訂され、高血圧の定義を従来の血圧140/90 mmHg以上から、130/80 mmHgに引き下げて、血圧130-139/80-89 mmHgをステージ1高血圧、血圧140/90 mmHg以上をステージ2高血圧と定義した。しかし、ステージ1高血圧が心不全や心房細動などの心血管イベントのリスクと関連するかは明らかでなかった。

Point:
ステージ1高血圧における心血管イベント発症リスクは、ステージ2高血圧と比較して低いことが予想される。したがって、ステージ1高血圧における心血管イベント発症リスクを評価するには、豊富なサンプルサイズが必要となり、JMDCデータの強みが活かせる。

Circulation. 2021 Jun 8;143(23):2244-2253.

Elevated BP/Stage 1 高血圧から心不全リスクは上昇

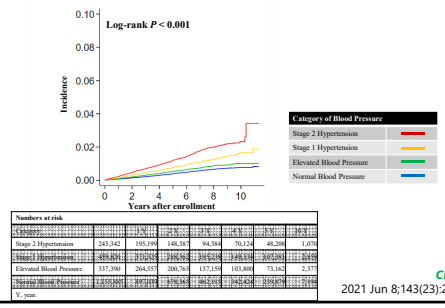


Circulation. 2021 Jun 8;143(23):2244-2253.

P	・JMDCデータベースに登録されている降圧薬を内服していない症例 ・心血管イベントの既往の無い症例 （n=2,196,437）
E	ステージ1高血圧 (SBP 130-139 mmHg or DBP 80-89 mmHg)
C	正常血圧 (SBP< 120 mmHg & DBP< 80 mmHg)
O	・心不全 ・心房細動

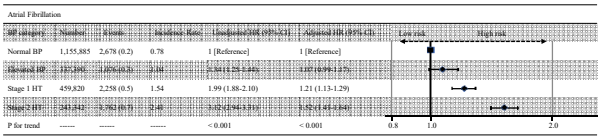
Circulation. 2021 Jun 8;143(23):2244-2253.

血圧分類と共に心房細動発症リスクは上昇する



Circulation. 2021 Jun 8;143(23):2244-2253.

Stage 1 高血圧から心房細動リスクは上昇



Circulation. 2021 Jun 8;143(23):2244-2253.

P	JMDCデータベースに登録されている20-49歳の症例で心血管イベントの既往の無い症例（n=913,235）
E/C	Life's Simple 7 Cardiovascular Health Metrics ※体重・喫煙・運動・食事・血圧・脂質・血糖で定義したModifiable Riskの指標
O	複合エンドポイント（心筋梗塞/脳卒中）

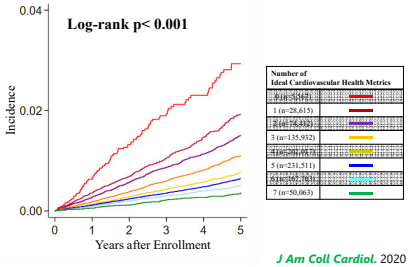
J Am Coll Cardiol. 2020 Nov 17;76(20):2414-2416.

JMDCデータを用いた研究紹介（1）

結論：
本邦の一般人口において、正常血圧と比較してステージ１高血圧が将来の心不全や心房細動発症リスクと上昇することが示された。JMDCデータベースが有する豊富なサンプルサイズを活かして、堅牢な臨床エビデンスを創出できた一例。

Circulation. 2021 Jun 8;143(23):2244-2253.

Life's Simple 7によって Young Adultの循環器疾患発症リスクが層別化される



J Am Coll Cardiol. 2020 Nov 17;76(20):2414-2416.

JMDCデータを用いた研究紹介（2）

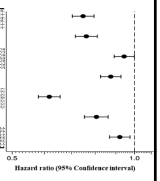
背景：
Young Adult世代の循環器疾患増加は先進国共通の疫学的課題であり、不健康な生活習慣などの"Modifiable Risk"が、Young Adult世代における生活習慣病や循環器疾患の増加に寄与すると考えられている。一方で、Young Adult世代を対象とした疫学研究は十分でなく、Modifiable Riskの指標である、Life's Simple 7 Cardiovascular MetricsによってYoung Adult世代の循環器疾患リスクの層別化が可能かは明らかでない。

Point:
Young Adult世代の循環器疾患発症リスクは、中高年世代と比較して低いため、Young Adult世代の循環器疫学研究には、医療ビッグデータ研究が有する大規模なサンプルサイズが適している。

J Am Coll Cardiol. 2020 Nov 17;76(20):2414-2416.

Life's Simple 7 Cardiovascular Health Metricsの各要素は Young Adultの循環器疾患発症リスクと独立して関連する

Ideal CHV metrics	Present	Adjusted Hazard ratio	95% Confidence Interval	P value
Ideal smoking Status	665,025	0.747	0.703-0.795	< 0.001
Ideal Body Mass Index	696958	0.761	0.715-0.810	< 0.001
Ideal Physical Activity	380,741	0.943	0.891-0.999	0.044
Ideal Diet Habits	406,379	0.874	0.825-0.926	< 0.001
Ideal Blood Pressure	523,495	0.617	0.580-0.656	< 0.001
Ideal Fasting Plasma Glucose	757,115	0.806	0.753-0.861	< 0.001
Ideal Total Cholesterol	479,164	0.920	0.868-0.975	0.005



Each Cardiovascular Health Metrics Component and Cardiovascular Event

J Am Coll Cardiol. 2020 Nov 17;76(20):2414-2416.

JMDCデータを用いた研究紹介（2）

結論：
本邦のYoung Adult世代においても、Life's Simple 7 Cardiovascular Health Metricsを用いて循環器疾患（心筋梗塞/脳卒中）発症リスクが層別化されることが示された。Young Adult世代での循環器疾患の予防でModifiable Riskの重症性を示唆する研究である。Working Age Populationを豊富に有するJMDCの強みが活かされた研究の一例。

J Am Coll Cardiol. 2020 Nov 17;76(20):2414-2416.

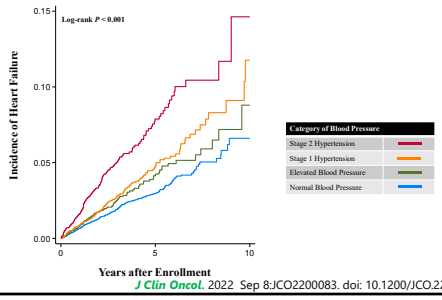
P	・JMDCデータベースに登録されている乳がん/大腸直腸がん/胃がんの既往のある症例 ・降圧薬を内服していない症例 ・心血管イベントの既往の無い症例 （n=33,991）
E	ステージ 1 高血圧 (SBP 130-139 mmHg or DBP 80-89 mmHg) ステージ 2 高血圧 (SBP ≥ 140 mmHg or DBP ≥ 90 mmHg)
C	正常血圧 (SBP< 120 mmHg & DBP< 80 mmHg)
O	・心不全

J Clin Oncol. 2022 Sep 8:JCO2200083. doi: 10.1200/JCO.22.00083.

JMDCはどんな研究に向いているか？

- 豊富なサンプル数・イベント数が必要な研究
- 悉皆性のあるデータが必要な研究
- 薬剤疫学研究

がんサバイバーにおいても血圧上昇と共に心不全リスクが上昇する



J Clin Oncol. 2022 Sep 8:JCO2200083. doi: 10.1200/JCO.22.00083.

JMDCデータを用いた研究紹介（3）

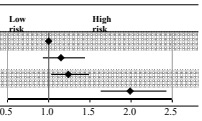
背景：
がん治療の進歩などに伴ってがん患者の長期生存が可能となり、がんサバイバーにおける循環器疾患（とりわけ心不全）が臨床的課題となっている。高血圧は循環器疾患の最大の危険因子であるが、がんサバイバーにおける高血圧と循環器疾患発症リスクの関連についてはデータが乏しかった。

Point:
“腫瘍循環器学”と呼ばれる当該分野の疫学研究には、がんと循環器疾患と言う領域をまたいだデータベースが必要であり、JMDCデータが有する悉皆性のあるデータが応用できる。

J Clin Oncol. 2022 Sep 8:JCO2200083. doi: 10.1200/JCO.22.00083.

がんサバイバーにおいても血圧上昇と共に心不全リスクが上昇する
-ステージ 1 高血圧の段階から有意に心不全リスクが上昇-

	No. of Patients	No. of Events	Model 1	Model 2	Model 3
Normal BP	17,444	301 (1.7)	1 (Reference)	1 (Reference)	1 (Reference)
Elevated BP	4,733	111 (2.3)	1.31 (1.05-1.62)	1.19 (0.95-1.48)	1.15 (0.93-1.44)
Stage 1 Hypertension	7,502	198 (2.6)	1.48 (1.24-1.77)	1.27 (1.06-1.53)	1.24 (1.03-1.49)
Stage 2 Hypertension	4,312	176 (4.1)	2.56 (2.13-3.09)	2.08 (1.72-2.52)	1.99 (1.63-2.43)



J Clin Oncol. 2022 Sep 8:JCO2200083. doi: 10.1200/JCO.22.00083.

JMDCデータを用いた研究紹介（３）

結論：
がんサバイバーを対象とした解析でも、ステージ１高血圧の段階から心不全リスクは有意に上昇し、ステージ２高血圧ではさらにそのリスクが上昇した。がんサバイバーにおいても血圧管理の重要性を示唆する研究である。本研究においては、JMDCデータが有する悉皆性のある健診データ・レセプトデータが大変有用であった。

J Clin Oncol. 2022 Sep 8:JCO2200083. doi: 10.1200/JCO.22.00083.

JMDCデータを用いた研究紹介（４）

背景：
SGLT2阻害薬は多くのRCTによって心血管イベント・腎イベント発症予防効果が報告されている。本邦では、6種類のSGLT2阻害薬が保険収載されており、我々は以前にこの6種類のSGLT2阻害薬間で心血管イベント発症リスクに差異が無い事を報告しているが（*Cardiovasc Diabetol*. 2022 May 18;21(1):67.）、腎イベントの発症リスクに差異があるかは明らかでなかった。

Point:
薬剤ブランド間でのイベントリスク比較はRCTの実装が困難である。このような研究においては、レセプトデータによる検討が有用である。

Kidney Int. 2022 Nov;102(5):1147-1153.

JMDCデータを用いた研究紹介（３）

結論：
本邦のYoung Adult世代においても、Life's Simple 7 Cardiovascular Health Metricsを用いて循環器疾患（心筋梗塞/脳卒中）発症リスクが層別化されることが示された。Young Adult世代での循環器疾患の予防でModifiable Riskの重症性を示唆する研究である。Working Age Populationを豊富に有するJMDCの強みが活かされた研究の一例。

J Am Coll Cardiiol. 2020 Nov 17;76(20):2414-2416.

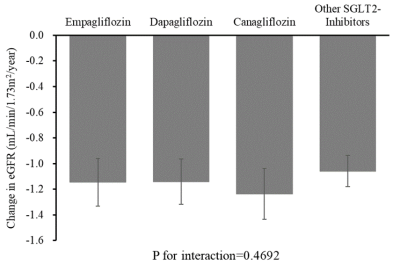
P	・JMDCデータベースに登録された糖尿病症例でSGLT2阻害薬が新たに処方された症例（n=12,100）
E	Empagliflozin
C	Dapagliflozin、Canagliflozin、その他のSGLT2阻害薬
O	年率eGFRの低下

Kidney Int. 2022 Nov;102(5):1147-1153.

JMDCはどんな研究に向いているか？

- 豊富なサンプル数・イベント数が必要な研究
- 悉皆性のあるデータが必要な研究
- 薬剤疫学研究

年率のeGFR低下速度はSGLT2阻害薬間で同等



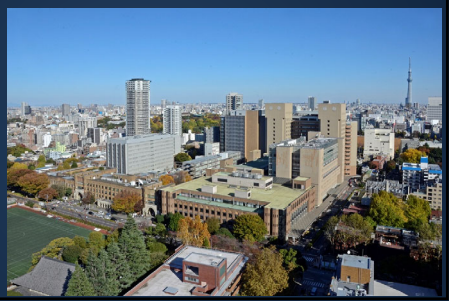
Kidney Int. 2022 Nov;102(5):1147-1153.

JMDCデータを用いた研究紹介（4）

結論：
本邦で保険収載されているSGLT2阻害薬のブラン間でeGFRの低下速度が同等であることが示された。SGLT2阻害薬の腎機能への影響がクラスエフェクトであることを示唆する結果である。レセプトデータを用いた薬剤疫学研究は、RCTが困難なClinical Questionに対しても応用可能である。

Kidney Int. 2022 Nov;102(5):1147-1153.

ご清聴ありがとうございました。



Agenda

1. JMDCデータの概要
2. 解析で注意すべき点
3. 研究紹介

総括

- JMDCデータは健診・レセプトデータを含む、本邦最大規模のリアルワールドデータベースであり、使用実績も豊富である。
- 若年成人や小児を対象とする研究、悉皆性のあるデータが必要な研究には強みがある一方で、高齢者を対象とする研究や、疾病の重症度、詳細な検査値が重要な研究には不向きである。
- 登録型コホートやRCTでは解決困難なClinical Questionの検証に有用となる可能性がある。
- データベースの利点・欠点を理解したうえでの研究立案、データハンドリング、統計解析、結果の解釈が重要である。

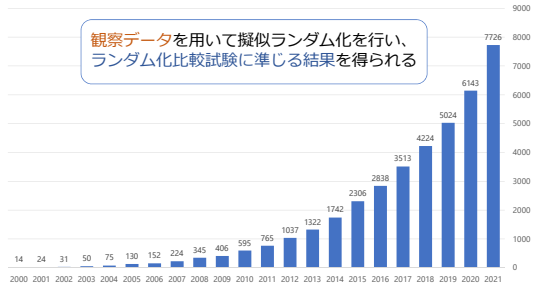
NDB・DPCデータベース研究人材育成Webinar

傾向スコア

東京大学大学院医学系研究科臨床疫学・経済学
東京大学大学院医学系研究科乳腺・内分泌外科学
小西孝明

目次

Pubmedで検索した"Propensity Score"関連論文数



目次

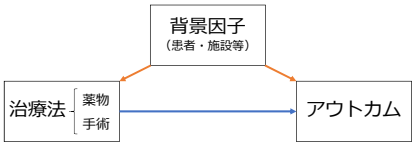
本講義の目標

➢本講義の目標

- ✓傾向スコア分析の基礎を理解する
- ✓傾向スコアマッチングについて学ぶ
- ✓傾向スコアを用いた逆確率重み付け、オーバーラップ重み付けについて学ぶ
- ✓傾向スコア分析を用いた先行研究について学ぶ
- ✓傾向スコア分析の注意点について学ぶ

概念＜観察研究の交絡＞

治療法の選択とアウトカムの両方に影響を及ぼす因子は交絡となる



軽症例に薬物療法、重症例に手術治療
→見かけ上、薬物療法の成績が良くなる

概念＜観察研究の交絡＞

治療群間で背景因子は異なるのが自然

	治療A	治療B
人数	700	700
男性	55%	60%
高血圧	70%	40%
糖尿病	60%	30%
死亡	10%	7%

7

概念＜マッチング＞

治療A			
ID	性別	高血圧	糖尿病
1	男	1	0
2	女	1	0
3	男	1	1
4	男	0	1
5	女	0	0
6	男	1	1
7	女	1	1

治療B			
ID	性別	高血圧	糖尿病
8	男	1	1
9	女	0	0
10	男	0	0
11	男	0	0
12	女	0	0
13	男	1	0
14	女	1	1

8

概念＜マッチング＞

治療A			
ID	性別	高血圧	糖尿病
1	男	1	0
3	男	1	1
5	女	0	0
7	女	1	1

治療B			
ID	性別	高血圧	糖尿病
13	男	1	0
8	男	1	1
12	女	0	0
14	女	1	1

似た背景因子の患者を選択　→　ランダム化比較試験に類似できる

変数が多い場合には完全一致は難しいので、ほぼ同じを目指すのが傾向スコア

9

概念＜傾向スコア＞

➢傾向スコアとは

- 観察された背景因子のもとで、ある治療を受ける確率
- Propensity score (PS)

➢傾向スコアの特徴

- 傾向スコアが近い人は、ほぼ同じ確率でその治療を受ける
- 傾向スコアが近い人は、ほぼ同じ特性を持っている

→傾向スコアを介して複数の交絡因子に対処し、群間のバランスをとれる

10

傾向スコアの推定＜理論＞

ある治療を受ける確率pを、背景因子X_nを独立変数とするロジスティック回帰で求める（β_0は切片、β_nは係数）

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

$$\therefore p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

この確率pが傾向スコアである

傾向スコアpは複数の背景因子が治療の割り当てに与える影響の度合いを1つの値にまとめたもので、これがほぼ等しければその背景因子はほどほどに近いといえる

11

傾向スコアの推定＜例＞

- 従属変数：治療A
- 独立変数：年齢、性別、高血圧、糖尿病、…

ID	治療	年齢 (β=.23)	性別 (β=.10)	高血圧 (β=.37)	糖尿病 (β=.36)	…	予測確率 p	
1	A	75	女	0	0	…	0.26	≒
2	B	74	女	0	0	…	0.22	
3	A	76	男	1	1	…	0.73	≒
4	B	81	男	1	0	…	0.70	

12

傾向スコアの推定 <実際>

➢統計ソフトで計算

ロジスティック回帰を行い、予測確率pを計算して保存

従属変数：治療割り当て変数（治療Aを受けたかどうか）

独立変数：背景因子　　（年齢、性別、高血圧、糖尿病、…）

通常のロジスティック回帰と異なり、独立変数に投入する背景因子に関して

- 投入しすぎを気にしなくてよい（傾向スコア分析の利点）
  - ✓過剰適合や多重共線性は考慮しなくて良い
- 治療選択より後に起きるものは投入しない（cf. 注意点）
  - ✓例、在院日数、合併症

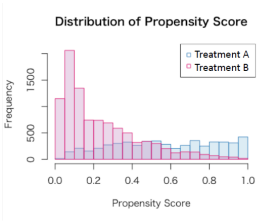
13

目次

16

傾向スコアの確認 <ヒストグラム>

- 治療Aと治療Bの傾向スコアの分布をヒストグラムで確認する
  - ✓傾向スコアが1に近いほど治療Aが選択
  - ✓傾向スコアが0に近いほど治療Bが選択
- ほどよい重なりが望ましい
  - ✓その理由は次ページに…



14

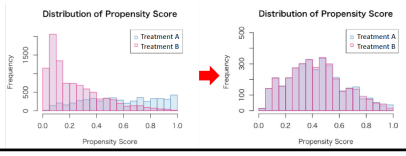
マッチングの手順

① 傾向スコアの推定

② 同じくらいの傾向スコアを持つペアを選択する

③ 二群のバランスの確認（後述）

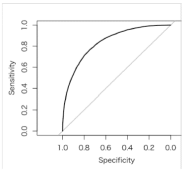
④ 二群でアウトカムを比較する（t検定、カイ二乗検定など）



17

傾向スコアの確認 <c統計量>

- 傾向スコアが治療Aの割り当てをどの程度予測するか？
  - ✓ROC（receiver operating characteristic）曲線を描く
  - ✓その曲線下の面積（AUC, area under curve）を算出する
  - ✓その値をc統計量と呼ぶ
    - 1.0　　：完全識別
    - 0.9–1.0：治療A群と治療B群の重なりが少ない（治療適応が異なるので効果比較の意味がない）
    - 0.6–0.9：適切と判断されることが多い
    - 0.5–0.6：治療A群と治療B群の重なりが多い（通常の多変量解析でも結果が変わらない）
    - 0.5　　：識別力無し



15

ペア選択のルール設定

A) マッチング方法
- ✓最近傍、最適、など
- ✓1対1、1対2、 1対3、、、

B) キャリパー（閾値）
- ✓最近傍マッチングの場合、ペアの傾向スコアの差の絶対値が閾値に収まるようにする
- ✓傾向スコア（のlogit）の標準偏差を基準
- ✓標準偏差×0.2に設定されることが多い

C) 抽出方法
- ✓復元、非復元

18

二群のバランスの確認

➢ 二群間のStandardized difference (SD)を各変数ごとに計算する
- ✓ マッチング前と後にそれぞれ算出する
- ✓ マッチング後にSDが0に近づいていれば二群の差異が減少している
- ✓ SDの絶対値が0.1未満でバランスが良いとみなすことが多い
- ✓ 0.1以上の場合、傾向スコア算出のモデルを修正する（変数を増やすなど）

p値の使用は推奨されない（nの影響を受けるため）

カテゴリ変数：

$$SD = \frac{|p_A - p_B|}{\sqrt{\frac{p_A(1-p_A) + p_B(1-p_B)}{2}}}$$

各群における割合p_A, p_Bを用いる

連続変数：

$$SD = \frac{|\bar{x}_A - \bar{x}_B|}{\sqrt{\frac{{s_A}^2 + {s_B}^2}{2}}}$$

各群における平均$\bar{x}_A$, $\bar{x}_B$と分散s_A, s_Bを用いる

19

目次

1. 傾向スコアの基礎
   - ① 概念
   - ② 傾向スコアの推定
   - ③ 傾向スコアの確認
2. 傾向スコアマッチング
3. 逆確率重み付け、オーバーラップ重み付け
4. 先行研究
5. 傾向スコア分析の注意点

22

論文中の記載例

- Method
  - ✓治療を予測する変数としてX_1, X_2,...を用いたロジスティック回帰を行った
  - ✓推定された傾向スコアを用い、非復元抽出による1対1の最近傍マッチングを行った
  - ✓キャリパーは傾向スコアの標準偏差の0.2倍に設定した
- Result
  - ✓傾向スコアマッチングでN人を抽出した
  - ✓そのc統計量は0.78であった（←c統計量は必ずしも記載しなくて良いとされる場合もある）
  - ✓Table1は、マッチング前後の背景因子である

20

傾向スコアを用いた重み付け

- 傾向スコアマッチングでは、マッチした患者のみを抽出して解析した
  - ✓ 傾向スコアの近い患者をマッチングさせて抽出している
  - ✓ そのためマッチする相手がいなかった患者は除外している
- 傾向スコアを用いた「重み」を掛けて、全ての患者を解析に利用する
  - ✓ 逆確率による重みづけ（Inverse probability of weighting, IPW）
  - ✓ オーバーラップ重み付け（Overlap weighting）

23

Table1の記載例

	マッチング前					マッチング後				
	治療A		治療B			治療A		治療B		
	n = 32,630		n = 26,764		SD	n = 23,055		n = 23,055		SD
男性	6,355	(20)	7,668	(29)	**0.213**	5,529	(24)	5,452	(24)	**−0.008**
年齢										
<55	12,427	(38)	7,528	(28)	**−0.213**	7,479	(32)	7,395	(32)	**−0.008**
55–65	10,969	(34)	8,568	(32)	**−0.034**	7,708	(33)	8,122	(35)	**0.038**
≥65	9,234	(28)	10,668	(40)	**0.246**	7,868	(34)	7,538	(33)	**−0.003**
高血圧	3,838	(12)	3,228	(12)	**0.009**	2,735	(12)	2,928	(13)	**0.003**
糖尿病	3,130	(9.6)	2,575	(9.6)	**0.001**	2,249	(9.8)	2,233	(9.7)	**−0.002**

※SDを絶対値(ASD, Absolute SD)や%を用いて表記することもある

SD, Standardized difference

21

観察している効果の違い

- 傾向スコアマッチングでは、ATTを観察している
  - ✓ ATT：治療群における平均処置効果（average treatment effect on the treated）
  - ✓ 治療A群の患者が仮に治療Bを受けた時のアウトカムの差
- 傾向スコアを用いた重み付けでは、ATEを観察できる
  - ✓ ATE：平均処置効果（average treatment effect）
  - ✓ 全患者が治療Aと治療Bを受けた時のアウトカムの差

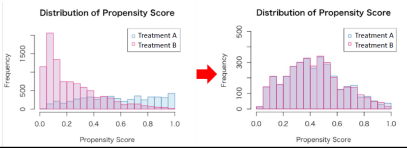
	治療A	治療B
人数	700	700
男性	55%	60%
高血圧	70%	40%
糖尿病	60%	30%
死亡	10%	7%

24

重み付けの手順

マッチングと同様の流れ

① 傾向スコアの推定
② 患者をそれぞれの傾向スコアで重み付けする
③ 二群のバランスの確認
④ 二群でアウトカムを比較する（t検定、カイ二乗検定など）



25

目次

28

重み付けの計算式

PSは治療Aを受ける確率
1-PSは治療Bを受ける確率

- 患者各々の傾向スコア(PS)を用いて重み付けをする

	治療A群	治療B群
逆確率による重み付け	1/PS	1/(1-PS)
オーバーラップ重み付け	(1-PS)	PS

- 逆確率による重み付けでの例
  - ✓ A群に傾向スコアが0.1の人が10人いたら、10/0.1＝100人いるものと見なす
  - ✓ B群に傾向スコアが0.6の人が20人いたら、20/(1-0.6)＝50人いるものと見なす
  - →重み付け後は各群の人数が増え、多くの場合非整数になる

26

提示する先行研究

- Mortality and morbidity after Hartmann's procedure versus primary anastomosis without a diverting stoma for colorectal perforation: a nationwide observational study
- A. Tsuchiya, H. Yasunaga, Y. Tsutsumi, H. Matsui, K. Fushimi
- *World Journal of Surgery* 2018;42:866–875

『DPCを用いた研究2』では、傾向スコアを用いた別の研究を例示しています

29

重み付けの利用の実際

- 主解析として利用されることはまだ少ない
  - ✓ マッチングより直感的に分かりにくい
  - ✓ 優れた解析手法ではあるが、読者に理解されにくい可能性がある
- 感度分析として利用されることが多い
  - ✓ 主解析でマッチングを行い、感度分析でいずれかの重み付けを行う
  - ✓ マッチングと重み付けを両方行うことで、ATTとATEの両方を観察できる

27

PECOと解析手法

P: 大腸穿孔に対して緊急開腹手術を受けた患者
E: 一時的にストマを作る手術（Hartman's procedure）
C: ストマを作らずに吻合手術
O: 30日死亡、合併症、術後集中治療

➢ 傾向スコアマッチング、逆確率による重み付け、操作変数法

30

Method＜傾向スコアマッチング＞

- 段落の冒頭で主解析の方法を明示
  "One-to-one propensity score matching was performed…"
- 傾向スコアの作成方法
  "To estimate the propensity score, a logistic regression model was used with the baseline independent variables (details are described in Supplemental Text)."
- マッチングのルール設定（マッチング方法、キャリパー、抽出方法）
  "Using a nearest-neighbor matching method, each patient in the PA group was matched with one patient in the HP group without replacement, with the closest estimated propensity within a caliper (0.2 standard deviations of the propensity score)."
- バランスの確認方法
  "The balance in the baseline variables between the propensity-matched HP and PA groups was examined using standardized differences, where >10% was regarded as imbalanced."

31

Method＜感度分析＞

- 感度分析として逆確率による重み付けを実施
  "We also used a propensity score method for inverse probability of treatment weighting (IPTW) using the same population as that in the propensity score matching analysis."
  "Each patient was weighted by the inverse probability of being in the observed group."
- 本研究では操作変数法も実施
  ✓操作変数法については別講義を参照ください

32

Result＜マッチングの結果＞

- 最初の段落（抽出された人数、c統計量）
  "A total of 8500 eligible patients with colorectal perforation were treated during the study period."
  "There were 5455 HP patients and 3045 PA without diverting stoma patients, from which 2800 propensity score-matched pairs were generated."
  "The C-statistic was 0.62 in the model for calculating propensity scores."
- 次の段落でTable1の説明（バランスの確認）
  "Table 1 shows the baseline characteristics of the unmatched and propensity score-matched groups…"
  "The distributions of the variables in the propensity score-matched groups were well balanced."

33

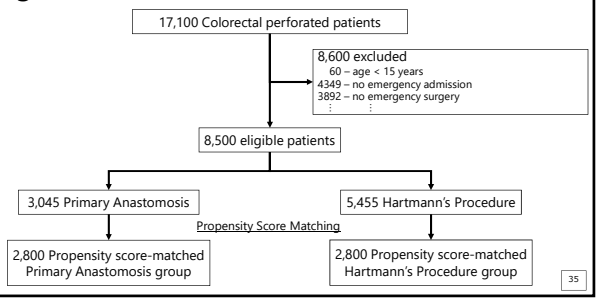
Table 1＜バランスの確認＞

	Unmatched groups					Matched groups				
	Primary anastomosis		Hartmann's procedure			Primary anastomosis		Hartmann's procedure		
	n = 3,045		n = 5,455		SD(%)	n = 2,800		n = 2,800		SD(%)
Age, years										
15–59	745	(24.5)	779	(14.3)	26.0	503	(18.0)	567	(20.3)	−5.8
60–69	542	(17.8)	1,130	(20.7)	−7.4	542	(19.4)	519	(18.5)	2.3
70–79	836	(27.5)	1,646	(30.2)	−6.0	835	(29.8)	804	(28.7)	2.4
≥80	922	(30.3)	1,900	(34.8)	−9.6	920	(32.9)	910	(32.5)	0.9
Sex male	1,680	(55.2)	2,538	(46.5)	17.5	1,460	(52.1)	2,928	(52.3)	−0.4
⋮	⋮		⋮		⋮	⋮		⋮		⋮

Data are presented as n and (%). SD, Standardized difference

34

Figure 1＜患者抽出の流れ＞



35

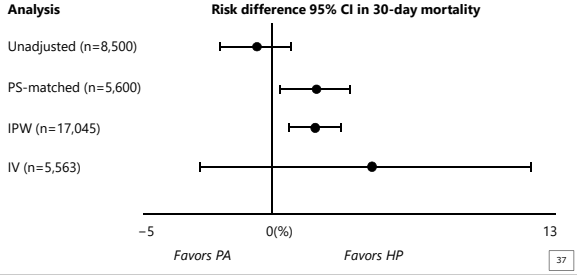
Table 2＜マッチング後のアウトカム＞

	Matched groups					
	Primary anastomosis		Hartmann's procedure		RD	95% CI
	n = 2,800		n = 2,800		(%)	
Overall postoperative complications	699	(25.0)	641	(22.9)	2.1	−0.2 to 4.3
Surgical complications						
Overall surgical interventions under GA	235	(8.4)	129	(4.6)	5.8	2.5 to 5.1
⋮	⋮		⋮		⋮	⋮

Data are presented as n and (%). RD, Risk difference

36

Figure 2＜各解析の結果＞



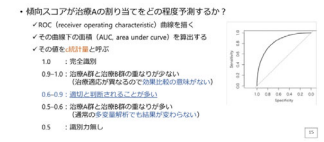
37

適度な群間差

➢治療適応の適度な一致が必要

- c統計量が大きい(0.9-1.0)
  - ✓2群で背景が大きく異なる
  - ✓一般化可能性が低い
  - ✓比較することにあまり意味がない
- c統計量が小さい(0.5-0.6)
  - ✓2群で背景が類似している
  - ✓一般的な回帰分析で十分比較可能
  - ✓傾向スコアを用いることを否定はされない

傾向スコアの確認＜c統計量＞



40

目次

1. 傾向スコアの基礎
   - ① 概念
   - ② 傾向スコアの推定
   - ③ 傾向スコアの確認
2. 傾向スコアマッチング
3. 逆確率重み付け、オーバーラップ重み付け
4. 先行研究
5. 傾向スコア分析の注意点

38

二群のバランス

➢傾向スコアによる調整後にバランス不良であった際の対処

- SD>0.1の際にはバランスがとれていない
- 傾向スコアによる調整後に以下の方法を行うとバイアスを生み出す
  - ✓SD>0.1となった（バランス不良の）変数でさらに調整する
  - ✓傾向スコアの算出に用いなかった別の交絡因子でさらに調整する

➢傾向スコア作成の際に投入する変数を工夫して対処する

- カテゴリー変数の区分を変更する
- 二乗項・交互作用項（例、年齢×BMI）を投入する

41

未測定の交絡因子

➢未測定の交絡因子には対処できない

- あくまで測定された交絡を制御しているに過ぎない
- 未測定の交絡因子によるバイアスにはさられている
  - ✓糖尿病足壊疽の研究で、HbA1cが入手できていない場合、など
- 未測定の交絡因子によって結果が反転する場合もある

操作変数法や高次元傾向スコアは別講義を参照ください

➢事後対処法はない

- 事前の研究計画・情報収集が肝要
- 感度分析（操作変数法など）やE-valueによって頑健性（robustness）を示す
- 高次元傾向スコアが有用かもしれない

39

傾向スコア解析の対象

➢治療介入（や短期の曝露）は評価可能

- 治療介入の評価のために傾向スコアは開発された
- 「治療介入より前に測定された因子」を明確に定義可能
  - ✓それらの因子によって治療介入を受ける確率を算出できる
  - ✓術式が原因で（遡って）高血圧になることはない

➢患者の状態（や長期の曝露）は評価不能？

- 「患者の状態より前に測定された因子」を明確に定義不能
  - ✓患者の状態より後に交絡因子？が発生している可能性
  - ✓慢性腎不全によって高血圧になる場合がある
- 通常の多変量解析やMatched-pair cohort解析などを検討する

42

おわりに

- 傾向スコアを用いた解析では、複数の交絡因子に対処し、群間のバランスをとることができる
- 傾向スコアマッチングでは、c統計量などを確認して記載する
- 傾向スコアを用いた重み付けでは、全患者におけるATEを見られる
- 未測定の交絡因子・適度な群間差・バランスに注意して使用する

43

NDB･DPCデータベース研究人材育成Webinar

高次元傾向スコア

high-dimensional propensity score, hd-PS

石丸美穂
東京医科歯科大学医歯学総合研究科健康推進歯学分野
プロジェクト助教

目標

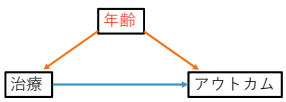
- 傾向スコアを大規模データの性質に合わせて拡張した高次元傾向スコアの基礎を理解する。
- 高次元傾向スコアの利点を理解する。
- 高次元傾向スコア分析を用いた先行研究について学ぶ。

目次

高次元傾向スコアとは？

交絡

治療法の選択とアウトカムの両方に影響を与える要因



例　・年齢若いほど治療A，高齢なほど治療Bを選択
　　・高齢なほどアウトカムが悪い
　　→ 治療Aの方が成績がよく見える

観察研究の限界

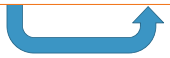
- レセプトデータ研究などの後ろ向き観察研究の場合、交絡因子が常に測定されて手に入るとは限らない
- 「未測定の交絡因子 (unmeasured confounder)」の存在は後ろ向き観察研究の最大の課題

例　疾患の重症度、検査値データ、全身状態など

未測定交絡の調整

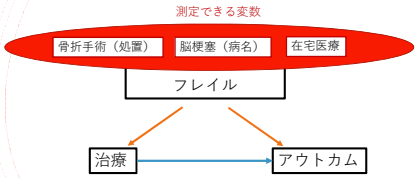
いままでは測定できる因子で未測定交絡因子の代理変数(proxy)とみなしてきた

未測定交絡因子	測定できる因子
全身状態が悪い	酸素吸入
ヘルスリテラシーの度合い	定期的な健診受診
ショック状態	カテコラミンの使用



高次元傾向スコアのコンセプト

レセプトデータなどの大規模データベースを利用した研究で
たくさんある患者情報から未測定交絡を間接的に調整する



フレイルそのものはデータベースで測定できないが、フレイルに関連がある変数は測定可能

傾向スコアからの拡張部分

- 傾向スコア推定モデルに投入する従属変数の選択方法
  → 今までは先行研究から重要だとされる変数を研究者が決定
- 高次元傾向スコアでは**一定のアルゴリズムに従って変数選択が自動**で行われる
- 多数の変数（50~1000変数）を複数の次元から選択し、傾向スコア推定モデルに投入する

次元 ＝ データの種類 （同じコーディングがされる同じ性質のもの）
病名、医薬品、診療行為、ラボデータなど

高次元傾向スコアのまとめ

- レセプトデータや電子カルテ等のリアルワールドデータを用いる研究で有用な方法
- 傾向スコアを拡張し、推定モデルに投入する変数を自動アルゴリズムで選択
- 膨大なデータを用いることで、未測定交絡を間接的に調整できる

▼ 高次元傾向スコアの活用

高次元傾向スコアの作成

(1) 次元の決定
(2) 変数コードの抽出
(3) 変数コードの出現回数の評価
(4) バイアス評価
(5) 変数の決定
(6) 傾向スコアの推定

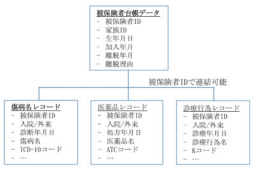
- Schneeweiss S, et al. "High-Dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data." 2009. *Epidemiology* 20 (4): 512–22.
- Schneeweiss, S. "Automated Data-Adaptive Analytics for Electronic Healthcare Data to Study Causal Treatment Effects." 2018. *Clinical Epidemiology* 10: 771–88.

高次元傾向スコアの作成(1)

次元 (dimension)の決定

- 次元＝データの種類
  （同じ構造、同じコーディング、同じ性質のも

傷病名、医薬品、診療行為
入院、外来
検査値データ　NDB等レセプトデータ



- 次元の決め方に制限はない（遺伝子データ、バイオマーカー、フリーテキスト、なんでもよい）

6 次元

高次元傾向スコアの作成(2)

変数コードの抽出

- 各次元ごとに変数コードの桁数を指定
  傷病名　ICD-10コード 4桁　E11.92
  薬剤　　ATC分類　　5桁　C08CA01
  *桁数が大きいと出現割合が小さくなりすぎる
- 出現割合の確認
  基準期間内に対象患者の何％で出現したか
  基準期間→治療A開始前の12カ月間（6カ月間）
- 各次元の上位n個の変数を候補として選択 n= 100~200

薬剤レセプトデータの例

被保険者ID	入院／外来	処方年月日	医薬品名	ATCコード
A	外来	20150124	アムロジピン錠5mg	C08CA01
A	外来	20150124	オルメテックOD錠5mg	C09CA08
B	入院	20150130	ティーエスワン配合顆粒	L01BC53
…	…	…	…	…

6 次元 × 200変数
＝1200変数

高次元傾向スコアの作成(3)

変数コードの出現回数の評価

- 患者ごとに各変数コードが何回出現したか数える
  → 3つの二値変数を作成
  ①1回以上出現
  ②散発的出現　（例：中央値以上）
  ③頻回出現　（例：75%タイル値以上）

1200変数 × 3変数
＝3600変数

ICD10コード I10(高血圧)の発生数の例

被保険者ID	I10の発生数	①	②	③
A	5	1	1	1
B	3	1	1	0
C	1	1	0	0
D	1	1	0	0
E	0	0	0	0
F	0	0	0	0
G	0	0	0	0
H	0	0	0	0
I	0	0	0	0
J	0	0	0	0

*発生している人の中で
②中央値は2、③75%タイル値は3.5

高次元傾向スコアの作成(4)

共変量のバイアス評価

各変数と治療及びアウトカムとの関連を用いて、バイアスの程度を評価

$$Bias_M = \frac{P_{c1}(RR_{cd}-1)+1}{P_{c0}(RR_{cd}-1)+1}$$

P_{c1},P_{c0}: 変数の治療群と対照群における割合
RR_{cd}：変数とアウトカムについての未調整リスク比

高次元傾向スコアの作成(4)

$$Bias_M = \frac{P_{c1}(RR_{cd}-1)+1}{P_{c0}(RR_{cd}-1)+1}$$

P_{c1},P_{c0}: 変数の治療群と対照群における割合
RR_{cd}：変数とアウトカムについての未調整リスク比

被保険者ID	I10_1	I10_2	I10_3	治療	死亡
A	1	1	1	1	1
B	1	1	0	0	0
C	1	0	0	1	1
D	1	0	0	1	0
E	0	0	0	1	1
F	0	0	0	1	0
G	0	0	0	0	1
H	0	0	0	0	0
I	0	0	0	0	0
J	0	0	0	0	0

変数名	Pc_1	Pc_0	RR_{cd}	$Bias_M$	$\lvert\log(Bias_M)\rvert$
I10_1	0.6	0.2	1.50	1.181	0.167
I10_2	0.2	0.2	1.33	1	0
I10_3	0.2	0	3.00	1.4	0.336

治療しているA,C,D,E,Fさん（5人）中、I10_1が1の人はA,C,Dさんの3人なので、**3/5=0.6**と計算

I10_1が1の群(A,B,C,Dさん）におけるアウトカム発生（A,Cさん)のリスクは2/4, I10_1が0群のリスク＝2/6, 未調整リスク比= 2/4÷2/6=1.5

I10_1の$Bias_M$=0.6(1.5-1)+1 / 0.2(1.5-1)+1 =1.3/1.1= 1.181

この3変数でバイアスが高い順は　**I10_3→ I10_1 →I10_2**

高次元傾向スコアの作成(5)

変数の決定

| log(BiasM)| を高い順から　k 個選択
→ k= 50 ~1000　今回は 500個選択 ⟸ 最終的に傾向スコア推定モデルに投入する変数

- 既知の交絡因子を傾向スコア推定モデルに追加
  → **年齢、性別、人種、医療サービスの利用**（外来受診回数、入院回数など）は投入することを強く推奨
  →　その他に先行研究から交絡因子だと言われているものはモデルに投入する

3600変数
↓
500 ＋ α

高次元傾向スコアの作成(6)

傾向スコアの推定

治療選択を従属変数としたロジスティック回帰により傾向スコアを推定

→通常の傾向スコア分析と同様に傾向スコアの分布確認

従来の傾向スコアを用いたアウトカム比較と同様に

マッチング、逆確率重み付け分析等を行う

500+α変数
↓
1

19

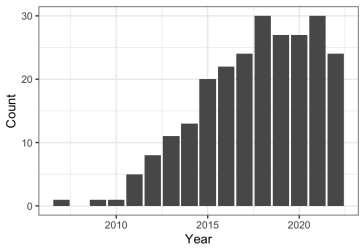
▼高次元傾向スコアを用いた先行研究

20

先行研究のレビュー

Pubmedで("high-dimensional propensity score*" OR "high dimensional propensity score*")で検索

検索結果：209編
全体では2018年ごろから横ばい



21

先行研究で使われているデータソース例

Country	US	UK	Canada	Denmark	France	Japan
データのタイプ	診療報酬データ	電子カルテデータ	診療報酬データ	診療報酬とヘルスレジストリ	ヘルスケアデータベース	診療報酬データ
データベース例	MarketScan, Optum, Medicare	CPRD, THIN	RAMQ database, MED-ECHO database	Danish National patient register	French National Healthcare System	JMDC
設定した次元	入院病名（ICD）	入院病名(THIN)	退院データから入院病名	入院病名（ICD）	入院病名（ICD）	入院病名（ICD）
	外来病名(ICD)	外来病名(THIN)	医師の診療報酬の病名コード	外来病名(ICD)	慢性疾患レジストリデータ(ICD)	外来病名(ICD)
	入院処置(ICD, CPTコード)	専門医への照会(READ codes) 入院処置(THIN)	退院データから処置コード 医師の診療報酬の処置コード		退院データから処置コード	入院処置(日本のオリジナルコード)
	外来処置(ICD, CPTコード)	外来処置(THIN)	医師の診療報酬の処置コード 診療を行う医師の専門性		医師の診療報酬の処置コード 診療を行う医師の専門性	外来処置(日本のオリジナルコード)
	外来の薬剤処方(NDC)	処方（BNFコード) 外来の薬剤使用(THIN)	外来での薬剤処方	薬剤 (ATCコード)	外来の処方(ATCコード)/入院中の高価薬剤(out of DRG cost coding system)	入院・外来での処方(日本のオリジナルコード, ATCコード)

(Rassen JA, et al. 2022.Pharmacoepidemiol Drug Saf.)

先行研究①：薬剤疫学

2型糖尿病患者にSU剤をセカンドラインで投与した場合の心臓血管疾患と低血糖のリスク

Douros A, et al. Sulfonylureas as second line drugs in type 2 diabetes and the risk of cardiovascular and hypoglycaemic events: population based cohort study. *BMJ* 2018 ;362:k2693.

Patient ：	メトホルミンで治療開始した2型糖尿病患者
Exposure:	SU剤を追加・SU剤に変更
Control：	メトホルミン単剤継続
Outcome：	心筋梗塞・全死亡・重症低血糖

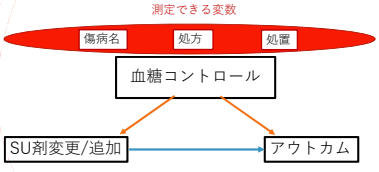
Data source: UK Clinical Practice Research Datalink (CPRD) + Hospital Episode Statistics (HES)

23

先行研究①：薬剤疫学

2型糖尿病患者にSU剤をセカンドラインで投与した場合の心臓血管疾患と低血糖のリスク

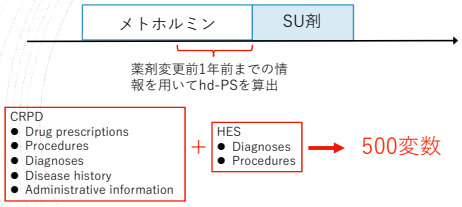
予想される未測定交絡：血糖コントロール



24

先行研究①：薬剤疫学

２型糖尿病患者にSU剤をセカンドラインで投与した場合の心臓血管疾患と低血糖のリスク



CRPD
- Drug prescriptions
- Procedures
- Diagnoses
- Disease history
- Administrative information

＋

HES
- Diagnoses
- Procedures

→ 500変数

7次元

25

先行研究②：診療行為による臨床疫学

周術期口腔機能管理とがん患者の術後合併症との関連

Ishimaru M, et al. Association between perioperative oral care and postoperative pneumonia after cancer resection. *Clin Oral Investig* 2019 ; 23:3581–3588

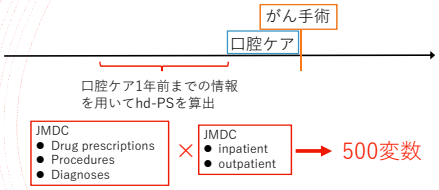
Patient：	主要ながん切除術施行患者
Exposure:	周術期口腔機能管理
Control：	周術期口腔機能管理なし
Outcome：	術後肺炎

Data source: JMDC database

26

先行研究②：診療行為による臨床疫学

周術期口腔機能管理とがん患者の術後合併症との関連



JMDC
- Drug prescriptions
- Procedures
- Diagnoses

×

JMDC
- inpatient
- outpatient

→ 500変数

6次元

27

先行研究②：診療行為による臨床疫学

周術期口腔機能管理とがん患者の術後合併症との関連

	口腔ケアあり（%）	口腔ケアなし(%)	ASD
BMI			
<18.5	7.2	6.7	2.0
18.5-24.9	54.4	44.5	19.9
>=25.0	14.5	12.2	6.7
missing	23.1	36.5	27.9
喫煙	21.3	14.7	17.1
飲酒	41.1	32.1	18.6

hd-PS マッチング

	口腔ケアあり（%）	口腔ケアなし(%)	ASD
BMI			
<18.5	7.2	5.0	9.4
18.5-24.9	54.6	52.7	0.3
>=25.0	13.9	13.8	0.5
missing	24.3	26.5	5.1
喫煙	21.0	19.8	3.0
飲酒	40.4	40.3	0.4

*特定検診からのデータは欠測が多いためPS推定に用いなかった

推定に用いていない変数についてもバランスが取れていた

28

先行研究③：ヘルスサービスリサーチ研究

予定PCI手術のケアの質に関する研究

Enders D, et al. The Potential of High-Dimensional Propensity Scores in Health Services Research: An Exemplary Study on the Quality of Care for Elective Percutaneous Coronary Interventions. *Health Serv Res*. 2018

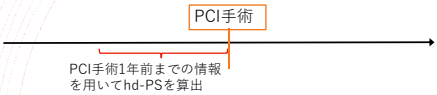
Patient：	待機的PCI(経皮的冠動脈インターベンション）手術施行患者
Exposure:	外来で施行
Control：	入院下で施行
Outcome：	死亡

Data source:
German Pharmacoepidemiological Research Database

29

先行研究③：ヘルスサービスリサーチ研究

予定PCI手術のケアの質に関する研究



- inpatient diagnoses
- ambulatory diagnoses
- inpatient procedures
- outpatient procedures
- ambulatory treatments
- ambulatory prescribed medications

→ 500変数

6次元

30

先行研究③：ヘルスサービスリサーチ研究

予定PCI手術のケアの質に関する研究

	HR
Adjusted analysis	0.58
PS matched	0.58
PS stabilized IPTW	0.45
hd-PS matched	1.18
hd-PS stabilized IPTW	1.20

先行研究やPSでは外来手術の方がHRが低くなる → 外来手術の方が健康な人が受けているという選択バイアスを調整しきれていなかった

31

▼高次元傾向スコアの利点と限界

32

高次元傾向スコアの利点

- ◆未測定交絡による残差交絡を減少させる可能性がある
- ◆レセプト・カルテデータなどに膨大にあるデータを無駄にしないで有効活用できる
- ◆研究者の変数選択時の恣意性が減少する
- ◆直感的には関係がない変数が重要な交絡因子である場合に調整できる

33

高次元傾向スコアの利点

シミュレーション研究では傾向スコアと比較して

① **推定モデルに入っていない患者背景因子もバランスが取れていた** (Guertin JR, et al. *BMC Med Res Methodol* 2016)

② **効果量がRCTのものに近似していた** (Garbe E, et al. *Eur J Clin Pharmacol* 2013)

34

高次元傾向スコアの限界

- 独立な未測定交絡因子については調整できない（医療診療情報データベースには得られるデータの系統が偏っている．SES(社会経済学的要因）が医療情報からproxyが得られるかは不明）
- 選択された変数が臨床的に解釈しづらいことがある
- 本来調整しない方が良い変数(操作変数、中間因子、M-bias、Z-biasなど）を調整している可能性がある

（over-adjustment 過剰調整の可能性が示唆されるが、もしそれらをhdPSで調整してしまっても、影響は小さいだろうとは報告されている Schneeweiss, S. 2018. *Clinical Epidemiology* :771–88.）

- さまざまな高次元データを用いた傾向スコア推定法（機械学習, Lasso回帰等)が提唱されており、今回の手法も含め、どの方法がより適しているのかについてはまだ明らかではない

図: 中間因子

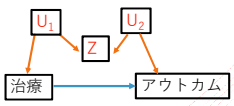


図: Mバイアス

35

▼高次元傾向スコア研究のチェックリスト

36

高次元傾向スコア研究のチェックリスト

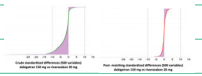
重要な考慮事項を記載したチェックリストを提供

高次元傾向スコアに馴染みのない研究者や読者が高次元傾向スコアを用いた研究を理解し、解釈するための支援ツールとして利用できる

▼

Rassen, JA, et al. 2022. "High-Dimensional Propensity Scores for Empirical Covariate Selection in Secondary Database Studies: Planning, Implementation, and Reporting." *Pharmacoepidemiology and Drug Safety*, November.

高次元傾向スコア研究のチェックリスト（発表者訳）

Guidance	Key concepts & considerations	Check
研究デザイン、プロトコル、統計解析プラン、hdPSの実行		
統計解析計画で共変量を特定・選択するためのパラメータを指定する。 ●次元の設定 ●変数と代理変数の同定 ●変数の優先順位付	●分析の前に、共変量をどのように同定し、ランク付けし、選択するかを事前に特定・決定する ●データソースごとの特徴を考慮する。	
研究者が事前に決めた変数の同定	研究プロトコルには事前に記載する。また、age, sex,race, health service utilization の変数は含めることを推奨	
IV(操作変数）やコライダー（合流点）となる変数は除外する		
PSの計算をするソフトウェア環境と、PSをどの手法で治療効果を推定するのかを記載		
計画された診断名ともしも異常が感知された時の手段を記載	●ベースラインの患者特性や、絶対標準化平均差などのサマリー指標に許容できる閾値を「表1」として作成し、選択した変数の検査を行う ●PS分布図や変数追加によるプロットなどを出力する	

Rassen, JA, et al. 2022. *Pharmacoepidemiology and Drug Safety.*

高次元傾向スコア研究のチェックリスト（発表者訳）

Guidance	Key concepts & considerations	Check
報告および透明性		
交絡調整の成功とhdPSの性能を示す診断表とグラフを提示する	●PSの分布や標準化差をプロットしたり、優先される共変量が順次追加されていく様子を示すことは、hdPSの性能を示すのに有効な視覚化手法である ●2つの治療群間の患者のベースライン特性を示す「表1」を作成する。	
STaRT-RWE*フレームワークの付録表3Fを完成させ、再現性と透明性を確保するために、共変量を同定・選択するために使用するパラメータを指定することを検討する。 *STaRT-RWE:Structured template and reporting tool for real world evidence(再現性のあるRWE研究をデザインし実施するためのガイドツール） Wang SV, et al. BMJ. 2021	●STaRT-RWE の構造化テンプレートは、研究手法の全体的な計画や報告を支援する。 ●STaRT-RWEの補足表3Fは、共変量定義のアルゴリズム、共変量評価期間、コードタイプ、診断位置を含む主要なパラメータを指定することを推奨する。 ●可能であれば、hdPS 法の透明性を高めるために、変数の詳細なリストと解釈可能な説明を補足付録の表で提供すること。	

Rassen, Jeremy A, et al. 2022. *Pharmacoepidemiology and Drug Safety.*

まとめ

- 高次元傾向スコアは大規模データの性質に合わせて傾向スコア推定の変数選択を自動化した方法である。
- 高次元傾向スコアは未測定交絡を間接的に調整できる可能性がある。
- 高次元傾向スコアは主に薬剤の効果比較研究に用いられており、様々な国レセプトデータベースで活用されている

NDB・DPCデータベース研究人材育成Webinar

操作変数法

東京大学大学院医学系研究科イートロス医学

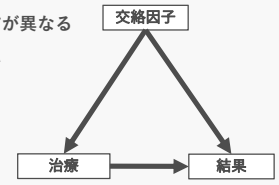
大野幸子

本日の内容

- 交絡
- 傾向スコア
- ランダム化比較試験
- 操作変数法
- 操作変数の例
- 操作変数法の仮定
- 単調性
- 操作変数法の推定の実際
- 操作変数法の実例
- 操作変数法checklist
- 操作変数法の限界

治療と結果の両方に影響

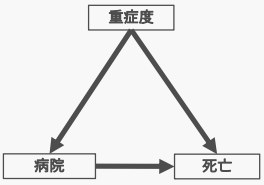
- 結果に影響を与える
- 曝露/治療の有無によって分布が異なる
- 曝露/治療から影響を受けない



専門病院Aを受診すると死にやすい？

- 専門病院A　20%が死亡
- 一般病院B　10%が死亡

交絡
専門病院Aには一般病院で治療できない**重症**な患者が集まる



未測定の交絡因子が存在しないとき

- 層別解析
- 回帰モデル
- **傾向スコア**
- etc.

未測定の交絡因子が存在しないという仮定で、

傾向スコア計算

治療を受ける確率を背景因子から予測

分析

- 層別化
- 回帰
- マッチング
- 重み付け

結果

ランダム化比較試験と同様の結果

ただし、
未測定の交絡がなければ

因果推論いろいろ 7

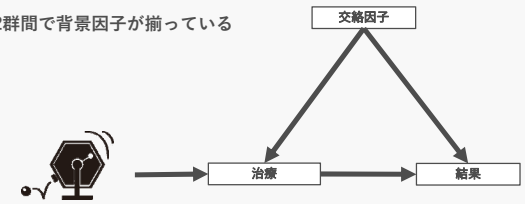
未測定の交絡が存在するとき

- **ランダム化比較試験(RCT)**
- 操作変数
- 差の差の分析
- 回帰不連続
- etc.

ランダム化比較試験 8

未測定の交絡因子に影響を受けない

- くじ引きで治療を割り当てる
- 2群間で背景因子が揃っている



ランダム化比較試験 （割付を完全に遵守した場合） 9

アウトカムの差は治療の効果による

割付	治療群	対照群
治療を受けた割合	100%	0%
年齢	73	73
女性	21%	20%
重症度スコア	9	9
未測定の交絡	?	?
アウトカムの割合	20%	50%

両群で揃う

治療の効果

ランダム化比較試験 10

いつでもRCTができるわけではない

- 費用（高）
- 人数（たくさん集めるのは大変）
- 倫理（ランダムに割り付けられない治療も存在）

未測定の交絡があるとき 11

RCTができないときの選択肢

- ~~ランダム化比較試験(RCT)~~
- **操作変数法**
- 差の差の分析
- 回帰不連続
- etc.

操作変数法 12

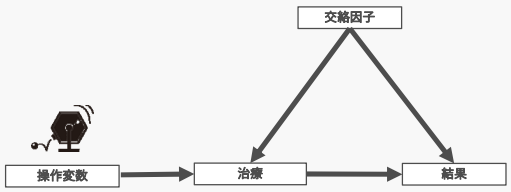
くじ引きの働きをする変数で未測定の交絡に対処

- 操作変数がRCTのくじ引きの役割を果たす
- 操作変数は治療を確率的に割り付ける
- 未測定の交絡因子も調整できる

操作変数法 13

くじ引きとなるような変数を使う

- RCTと同じイメージ図



操作変数 14

アウトカムの差は治療の割合の差に起因

操作変数の割付	治療されやすい群	治療されにくい群
治療を受けた割合	75%	25%
年齢	73	73
女性	21%	20%
重症度スコア	9	9
未測定の交絡	?	?
アウトカムの割合	20%	30%

両群で揃う

治療の効果

操作変数法 15

治療の受けやすさに影響するくじ引き変数の例

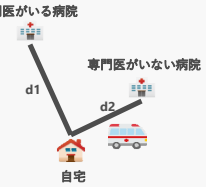
- ある治療を積極的に行う病院と消極的に行う病院までの距離の差(Differential distance)
- 医師/病院の治療の好み
- 地域の治療実施割合
- 曜日、日付
- 遺伝的変異(Mendelian randomization)

操作変数法 16

Differential distance
病院が備える特性を見るための操作変数

- (自宅から専門病院までの距離(d1))-(自宅から一般病院までの距離(d2))

d1-d2 **小** → 専門治療を受けやすい
d1-d2 **大** → 専門治療を受けにくい



操作変数法 17

曜日(Friday admission)
早期介入の効果を見るための操作変数

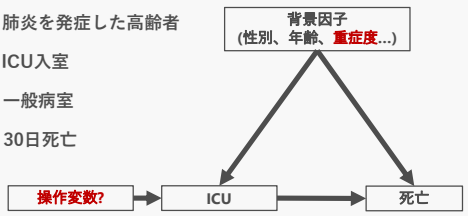
- 金曜：土日を挟むため、3日目以降に手術/検査が遅れやすい
- 翌日か翌々日が平日なので2日以内に手術/検査を受けやすい

月 → 火 → 水 → 木 → 金 → 土 → 日 → 月

操作変数の例 JAMA 2015;314:1272–9 18

高齢者の肺炎はICUで管理するべきか？

- P：肺炎を発症した高齢者
- E：ICU入室
- C：一般病室
- O：30日死亡



操作変数の例　JAMA 2015;314:1272–9　19

高齢者の肺炎はICUで管理するべきか？

- P：肺炎を発症した高齢者
- E：ICU入室
- C：一般病室
- O：30日死亡
- **操作変数：differential distance**

(自宅からICUに入室させる病院とそうでない病院までの距離の差)

操作変数の仮定　20

基本の仮定は3つ

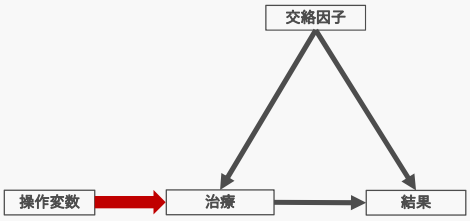
仮定①
治療と強く関連している

仮定②
結果と直接関連しない

仮定③
結果との共通原因を持たない

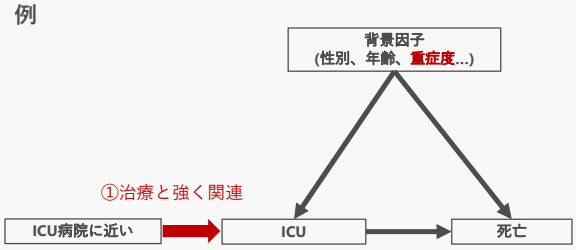
操作変数法の仮定①　21

①治療と強く関連している



操作変数法の仮定　22

例



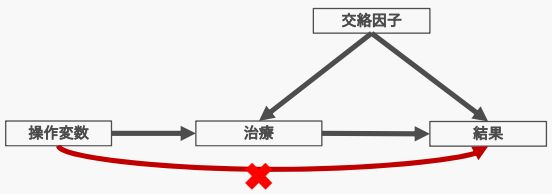
操作変数法の仮定　23

①治療と強く関連している

- くじ引きの役割「操作変数が1なら治療を受ける」(100%ではない)
- 治療を操作変数で回帰した際のF値で評価 (F > 10が目安)
- F値 ≤10: 弱い操作変数(weak instrument)で誤った結果を導く

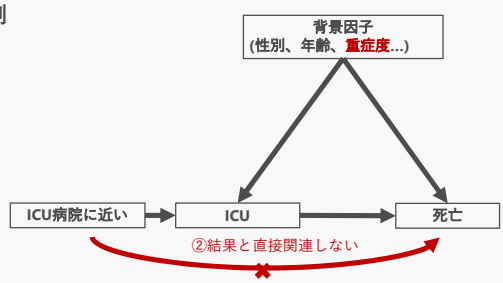
操作変数法の仮定②　24

②結果と直接関連しない(治療を介してのみ関連あり)



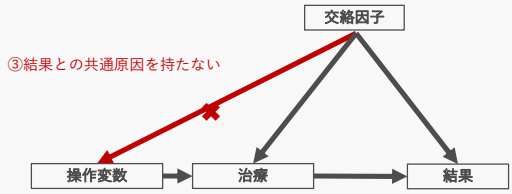
操作変数法の仮定② 25

例



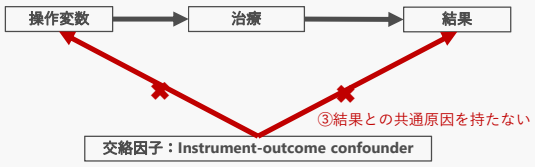
操作変数法の仮定③ 26

③結果との共通原因を持たない



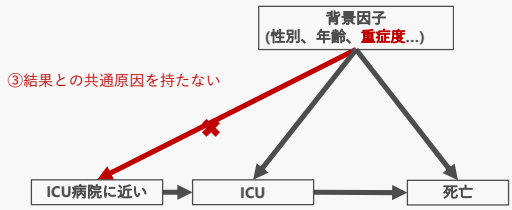
操作変数法の仮定③ Ann Intern Med. 2014;161:131-138 27

③結果との共通原因を持たない



操作変数法の仮定③ 28

③結果との共通原因を持たない



操作変数法の仮定②③ 29

②結果と直接関連しない
③結果との共通原因を持たない

- ②③は証明できない
- 臨床的にメカニズムを判断
- ③はIVのカテゴリごとに測定されている背景因子を比較し傍証とすることもある

操作変数の仮定 30

仮定を検証できるのは①のみ

仮定①
治療と強く関連している

仮定②
結果と直接関連しない

仮定③
結果との共通原因を持たない

操作変数の仮定 31

仮定を満たさない効果変数は推定にバイアス

効果推定量

操作変数の②③の仮定を満たさないとバイアス

$$\frac{\text{（操作変数1の時のアウトカム）} - \text{（操作変数0の時のアウトカム）}}{\text{（操作変数1の時の治療）} - \text{（操作変数0の時の治療）}}$$

仮定①を満たさず、弱い操作変数でゼロに近づく

操作変数の仮定 32

追加の仮定は1つ

仮定①	仮定②
治療と強く関連している	結果と直接関連しない
仮定③	**追加の仮定④**
結果との共通原因を持たない	単調性

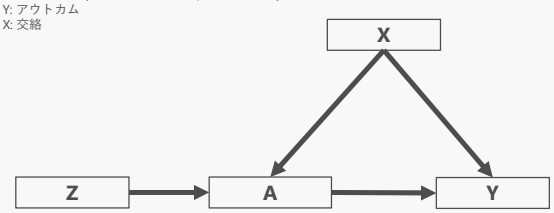
操作変数法の仮定④ 33

④単調性

ICU病院近い	ICU病院遠い	タイプ	
一般病室	一般病室	Never-taker	
ICU	一般病室	Complier	
一般病室	ICU	Defier ✗	割り当てと逆の治療を受ける人は存在しないと仮定
ICU	ICU	Always-taker	

単調性が必要な理由 34

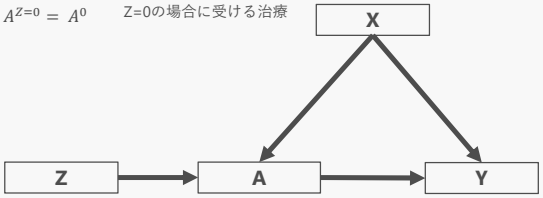
Z: 操作変数 (治療群に割付けられる場合を1, それ以外を 0)
A: 受ける治療 (治療を受る場合は1,それ以外を 0)
Y: アウトカム
X: 交絡



単調性が必要な理由 35

$A^{Z=1} = A^1$ Z=1の場合に受ける治療

$A^{Z=0} = A^0$ Z=0の場合に受ける治療



単調性が必要な理由 36

4つのタイプ

A^0	A^1	タイプ
0	0	Never-taker
0	1	Complier
1	0	Defier
1	1	Always-taker

Target Inference
Complier average causal effect（CACE）

Complier average treatment effect (CATE)
Local average treatment effect (LATE)

$$E(Y^{Z=1}|A^0=0,A^1=1)-E(Y^{Z=0}|A^0=0,A^1=1)$$
$$=\mathrm{E}(Y^{Z=1}-Y^{Z=0}|\text{compliers})$$
$$=\mathrm{E}(Y^{a=1}-Y^{a=0}|\text{compliers})$$

操作変数法で得られる結果はcomplierの中での効果
操作変数が治療の割り当てに影響しないalways-takerとnever-takerでの効果は不明
defierがいてはなぜいけないのか？

実際に観察されるデータからはタイプを決定できない

Z	A	A^0	A^1	タイプ
0	0	0	?	Complier Never-taker
0	1	1	?	Always-taker Defier
1	0	?	0	Never-taker Defier
1	1	?	1	Always-taker Complier

割り当てと逆の治療を受ける人は存在しないと仮定

A^0	A^1	タイプ
0	0	Never-taker
0	1	Complier
1	0	Defier ✗
1	1	Always-taker

Defierがいないと仮定すると、、

Z	A	A^0	A^1	タイプ
0	0	0	?	Complier Never-taker
0	1	1	1	Always-taker ~~Defier~~
1	0	0	0	Never-taker ~~Defier~~
1	1	?	1	Always-taker Complier

Target Inference
Complier average causal effect（CACE）

$$\mathrm{E}(Y^{a=1}-Y^{a=0}|\text{compliers})$$

まずは観察データで計算できるものから、

$$\mathrm{E}(Y^{Z=1}-Y^{Z=0})=\mathrm{E}(Y|Z=1)-\mathrm{E}(Y|Z=0)$$

単調性を仮定した場合

$$\mathrm{E}(Y^{Z=1}-Y^{Z=0})=\mathrm{E}(Y|Z=1)-\mathrm{E}(Y|Z=0)$$

ここで

$$\mathrm{E}(Y|Z=1)=\mathrm{E}(Y|Z=1,\text{always-taker})\mathrm{P}(\text{always-taker})+\mathrm{E}(Y|Z=1,\text{never-taker})\mathrm{P}(\text{never-taker})+\mathrm{E}(Y|Z=1,\text{complier})\mathrm{P}(\text{complier})$$

$$\mathrm{E}(Y|Z=0)=\mathrm{E}(Y|Z=0,\text{always-taker})\mathrm{P}(\text{always-taker})+\mathrm{E}(Y|Z=0,\text{never-taker})\mathrm{P}(\text{never-taker})+\mathrm{E}(Y|Z=0,\text{complier})\mathrm{P}(\text{complier})$$

never-takerとalways-takerはZに影響を受けない

$$E(Y^{Z=1} - Y^{Z=0}) = E(Y|Z=1) - E(Y|Z=0)$$

ここで

$$E(Y|Z=1) = E(Y|\cancel{Z=1}, \text{always-taker})P(\text{always-taker}) + E(Y|\cancel{Z=1}, \text{never-taker})P(\text{never-taker}) + E(Y|Z=1, \text{complier})P(\text{complier})$$

$$E(Y|Z=0) = E(Y|\cancel{Z=0}, \text{always-taker})P(\text{always-taker}) + E(Y|\cancel{Z=0}, \text{never-taker})P(\text{never-taker}) + E(Y|Z=0, \text{complier})P(\text{complier})$$

同じ

$$E(Y|Z=1) - E(Y|Z=0) = E(Y|Z=1, \text{complier})P(\text{complier}) - E(Y|Z=0, \text{complier})P(\text{complier})$$

両辺をP(complier)で割って

$$\frac{E(Y|Z=1) - E(Y|Z=0)}{P(\text{complier})}$$

$$= E(Y|Z=1, \text{complier}) - E(Y|Z=0, \text{complier})$$
$$= E(Y^{a=1}|\text{complier}) - E(Y^{a=0}|\text{complier})$$ ← CACE

Target Inference
Complier average causal effect（CACE）

$$\text{CACE} = \frac{E(Y|Z=1) - E(Y|Z=0)}{P(\text{complier})}$$

$$P(\text{complier}) = E(A|Z=1) - E(A|Z=0)$$

always-takerまたはcomplierの割合

always-takerの割合

単調性の仮定により観察データからCACEを得ることができる

$$\text{CACE} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(A|Z=1) - E(A|Z=0)}$$

仮定を検証できるのは①のみ

仮定① 治療と強く関連している	仮定② 結果と直接関連しない
仮定③ 結果との共通原因を持たない	追加の仮定④ 単調性

操作変数法ののいろいろな推定方法

- **Wald推定量**
- **2-stage least squares（TSLS）**
- **2-stage residual inclusion（TSRI）**
- 2-stage prediction substitution（TSPS）
- etc.

推定の実際 49

Wald推定量

$$\frac{E(Y|Z=1)-E(Y|Z=0)}{E(A|Z=1)-E(A|Z=0)}$$

(操作変数1の時のアウトカム) − (操作変数0の時のアウトカム)

(操作変数1の時の治療) − (操作変数0の時の治療)

推定の実際(共変量があるとき) 50

2SLS

- 線形モデルを仮定
- アウトカムが連続変数の時に適用されやすい
- パッケージが存在

推定の実際(共変量があるとき) 51

2SLS

① 治療の変数を操作変数と背景因子で線形回帰する (F値確認)

$$A = \alpha_0 + \alpha_1 Z + \alpha_2 C + \varepsilon_1$$

② ①で得られた係数をもとに治療の予測値を求める

$$\hat{A} = \widehat{\alpha_0} + \widehat{\alpha_1} Z + \widehat{\alpha_2} C$$

③ アウトカムを②で求めた予測値と背景因子で線形回帰する

$$Y = \beta_0 + \beta_1 \hat{A} + \beta_2 C + \varepsilon_2$$

分散は修正が必要(Palmer et al., AJE 2017)

推定の実際(共変量があるとき) 52

2SRI

- 非線形モデルを仮定
- アウトカムが2値変数に適用されることが多い
- 2回一般化線形モデルを使う

推定の実際(共変量があるとき) 53

2SRI

① 治療の変数を操作変数と背景因子で回帰する

$$A = \alpha_0 + \alpha_1 Z + \alpha_2 C + \varepsilon_1$$

② 実際の治療の値と①から計算できる予測値の差(残差) を計算

$$\widehat{\varepsilon_1} = A - \hat{A}$$

③ アウトカムを②で求めた残差と治療と背景因子で回帰する

$$Y = \beta_0 + \beta_1 A + \beta_2 C + \widehat{\varepsilon_1}$$

分散は修正が必要(Palmer et al., AJE 2017)

過剰識別制約 54

複数の操作変数を投入する場合

- 1つの治療効果に対して2つ以上操作変数を投入できる
- F値が高くなる可能性があり、弱い操作変数でなくなる
- 過剰識別制約検定が必要
  「全ての操作変数が内生変数ではない」ことを検定
  (1つでも操作変数が内生変数でなければよい)
- 内生変数：説明変数だが、別な説明変数で説明されてしまう変数
- 操作変数が内生変数だと条件を満たさないため不適当

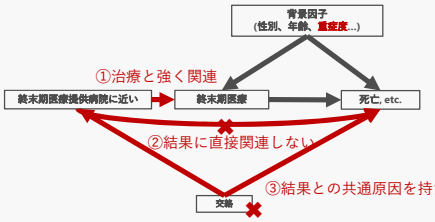
操作変数法の実例① *Crit Care Med.* 2019;47(11):1591-1598. 55

終末期医療提供はICU患者のアウトカムに影響するか？

- P：ICU患者
- E：終末期医療提供病院
- C：通常医療提供病院
- O：**院内死亡/ホスピス退院/** 増加すると予測
  在院日数/治療強度 減少すると予測
- 操作変数：Differential distance

操作変数法の実例① *Crit Care Med.* 2019;47(11):1591-1598. 56

終末期医療提供はICU患者のアウトカムに影響するか？



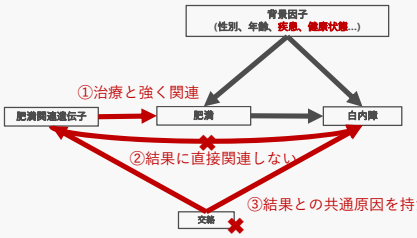
操作変数法の実例② *Am J Clin Nutr.* 2019;110(4):969-976. 57

肥満は白内障と関連するか？

- P：49歳以上の成人コホート
- E：肥満
- C：肥満でない
- O：白内障
- 操作変数：肥満関連遺伝子（Mendelian randomization）

操作変数法の実例② *Am J Clin Nutr.* 2019;110(4):969-976. 58

肥満は白内障と関連するか？



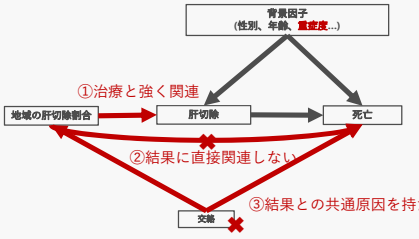
操作変数法の実例③ *Ann Surg.* 2019;270(4):692-700. 59

大腸がん肝転移患者の肝切除は生命予後を改善するか?

- P：大腸がん肝転移患者
- E：肝切除
- C：非切除
- O：1, 2, 3, 4, 5年生存
- 操作変数：地域の肝切除施術割合

操作変数法の実例③ *Ann Surg.* 2019;270(4):692-700. 60

大腸がん肝転移患者の肝切除は生命予後を改善するか?



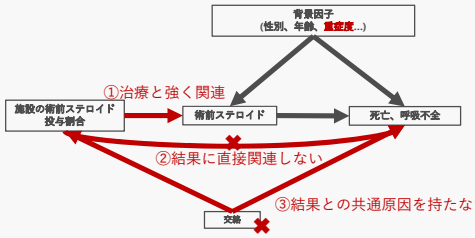
操作変数法の実例④ Ann Surg. 2022. in press 61

術前ステロイド投与は食道切除術の予後を改善するか?

- P：食道がん患者
- E：術前ステロイド投与
- C：術前ステロイド投与なし
- O：入院死亡、呼吸不全
- 操作変数：施設の術前ステロイド投与割合

操作変数法の実例④ Ann Surg. 2022. in press 62

術前ステロイド投与は食道切除術の予後を改善するか?



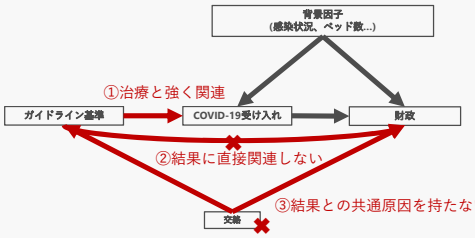
操作変数法の実例⑤ J Jpn Int Econ. 2022;65: 101218. 63

COVID-19患者の受け入れは病院の財政に影響するか?

- P：東京のCOVID-19患者受け入れ可能な病院
- E： COVID-19患者受け入れ
- C： COVID-19患者受け入れなし
- O：1ベッドあたりの利益/月
- 操作変数：COVID－19受け入れに関するガイドライン（病床あたりの呼吸器専門医数、病床あたり個室数）

操作変数法の実例⑤ J Jpn Int Econ. 2022;65: 101218. 64

COVID-19患者の受け入れは病院の財政に影響するか?



操作変数checklist Epidemiology 2013;24: 363–369 65

論文を読むとき、書くときの注意点

- 推定する対象集団への効果であること（CACE）や仮定を明記
- 操作変数と治療の関連をF値を使って明記
- 測定可能な交絡因子と操作変数や治療の関係を記載
- 2値のアウトカム、治療、操作変数の組合せを明記
- 分散は修正が必要

操作変数の限界 66

実は限界が多い

- 適切な操作変数を見つけることが難しい
- ①操作変数と治療に強い関連がないと、誤った結果を導く
- ②操作変数と結果に関連がないことを確実に証明できない
- ③操作変数と結果が共通の原因を持たないことを確実に証明できない
- 仮定が多い（操作変数の条件①－③、④単調性）

バイアス増幅の危険性

CACE

$$\frac{E(Y|Z=1)-E(Y|Z=0)}{E(A|Z=1)-E(A|Z=0)}$$

操作変数の仮定②③を満たさないとバイアス

仮定①弱い操作変数でゼロに近づく

実は限界が多い

- 重要な交絡が欠落しているときに操作変数法単独使用を検討
- 操作変数法単独の研究は多くはない（4割程度）
- 傾向スコア分析などの方法と併用されることが多い
- 傾向スコア分析と操作変数法で、結果が異なるということも
- 交絡因子が全て測定されている仮定 VS. 操作変数の仮定

本日の内容

- 交絡
- 傾向スコア
- ランダム化比較試験
- 操作変数法
- 操作変数の例
- 操作変数法の仮定
- 単調性
- 操作変数法の推定の実際
- 操作変数法の実例
- 操作変数法checklist
- 操作変数法の限界

- 操作変数法は未測定の交絡因子を調整できる
- 適切な操作変数を見つけることは難しい
- 仮定が厳しい
- 仮定を満たさない、弱い操作変数でバイアスを増幅
- 単独で用いることは少ない

NDB・DPCデータベース研究人材育成Webinar

不連続回帰デザイン・差の差分析

自治医科大学データサイエンスセンター
山名隼人

Agenda

テーマ
- 不連続回帰デザイン（regression discontinuity）
- 差の差分析　（difference in differences）

内容
- 研究デザインの背景
- 研究デザインの基礎・注意点
- どのような臨床疑問の検証に適しているか
- 活用例

Agenda

テーマ
- 不連続回帰デザイン（regression discontinuity）
- 差の差分析　（difference in differences）

内容
- 研究デザインの背景
- 研究デザインの基礎・注意点
- どのような臨床疑問の検証に適しているか
- 活用例

臨床研究の種類

- 介入研究
  - 実験的な環境
  - ２群間の比較を行う場合、背景を揃えることが可能
  - 例：ランダム化比較試験
- 観察研究
  - 実際に行われた診療を“見るだけ”
  - ２群間の比較を行う場合、“結果的に”治療を受けた２群を比較する
  - 例：保健医療データベースを用いた研究

臨床研究の種類

例：「降圧薬の使用有無による脳卒中予防効果」
- ランダム化比較試験の場合
  - 基礎疾患、血圧などが揃った２群で比較が可能
- 観察研究の場合
  - 患者背景は２群で異なるのが当然
  - ベースラインの血圧は降圧薬あり群で高い

連続値と閾値

- 臨床的判断は不連続
  特定の閾値を境に治療方針が変わることがある
  例：収縮期血圧≥140mmHgの場合に降圧薬が開始
- 血圧は連続値
- 血圧が1mmHg上がっても脳卒中のリスクはほぼ不変
- 測定時の1mmHgの誤差は偶然による

Regression discontinuity

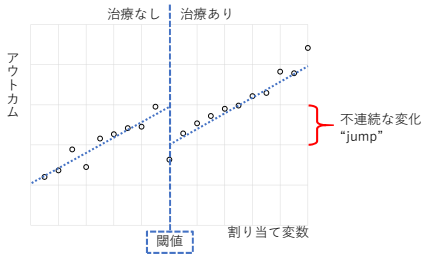
- 治療の判断に使われる**割り当て変数**（連続）に**閾値**がある場合に適用
- 閾値の近傍で、閾値以上 vs 閾値未満でアウトカムを比較する
- ランダム割り付けに近い状況を作り出す
- 例

  収縮期血圧140mmHgで降圧薬が開始された群

  v s

  収縮期血圧139mmHgで降圧薬が開始されなかった群

Regression discontinuity



Regression discontinuity

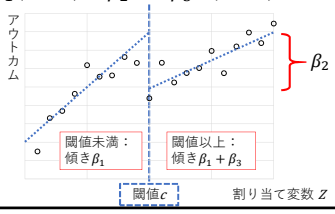
回帰式を用いる場合

$$Outcome = \beta_0 + \beta_1(Z - c) + \beta_2 D + \beta_3 D(Z - c)$$

Z: 割り当て変数

c: 閾値

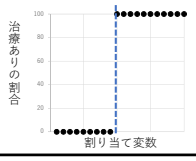
D: $Z \geq c$なら1, $Z < c$なら0



Sharp/fuzzy regression discontinuity

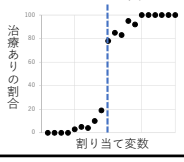
- Sharp RD

  閾値により割り当てが「全か無か」で決まる場合



- Fuzzy RD

  閾値前後で割り当てが不連続に変わるが、完全に２分されるわけではない場合



Fuzzy RDの場合

- 閾値を超えたかどうかは、
  - 治療と強く関連
  - 結果と直接は関連しない
  - 結果との交絡が存在しない

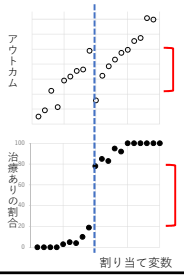
  操作変数と同じ
- 対象者は下記に分類される：
  - 閾値を超えたかに関わらず、治療を受けたであろう人 (always-taker)
  - 閾値を超えたかに関わらず、治療を受けなかったであろう人(never-taker)
  - 閾値をまたいだことで、治療を受けるようになった人 (complier)
  - ~~閾値をまたいだことで、治療を受けなくなった人~~ (defier)

Fuzzy RDの場合

操作変数として考える

- アウトカムのjumpはintention to treat (ITT)に相当
- 単調性の仮定のもとで、Wald推定量を用いてcomplier average causal effect (CACE)を計算できる

$$\text{CACE} = \frac{\text{閾値前後のアウトカムのjump}}{\text{閾値前後の治療割合のjump}}$$



不連続回帰デザインの活用法（1）

検査値等と治療

- 閾値前後で治療（介入）が変化する場合に、介入の効果を検証する

例

- 介入：　降圧薬　輸血　手術/保存的
- 割り当て変数：　血圧　Hb　年齢

不連続回帰デザインの活用法（2）

時期による治療方針の変化

- 制度変更やガイドラインの更新などにより治療方針（介入）が急に変化した場合

例

- 介入：　治療方針・ワクチン接種
- 割り当て変数：　時期・年度

不連続回帰デザインの活用法（3）

医療政策の評価

- 政策の対象者が限定され、一定の基準がある場合に、政策の効果を検証する

例

- 介入：　保険・助成
- 割り当て変数：　収入・年齢・地域

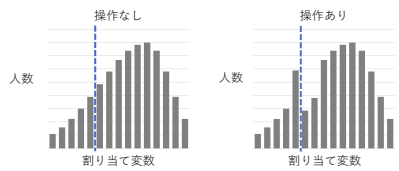
不連続回帰デザインの前提（1）

1．割り当てのルールと閾値が明確である
- 検査値等が一定の値を超える（or下回る）と介入
- 個人によって閾値が異なる場合は×

2．介入以外の要因は閾値付近で連続である
- 他の要因も同時にjumpをしてはいけない
- 血圧の治療と同時に脂質・血糖値の治療も開始される場合は×

不連続回帰デザインの前提（2）

3．対象者が介入前に割り当て変数を操作できない（manipulation）
- メタボ健診の腹囲測定は×



不連続回帰デザインの利点と限界

利点

- ランダム化比較試験に準じて因果関係が推定できる
- リアルワールドにおける効果を検証できる

限界

- 閾値付近の結果を閾値から遠い対象者にあてはめることには限界がある（外的妥当性）
- 閾値付近の人数は少ないため、比較的大きなN数が必要

不連続回帰デザインの例 (1)

HIV患者に対する抗レトロウイルス療法開始

- 対象者　南アフリカ　HIV陽性患者
- 介入　抗レトロウイルス療法開始
- 割り当て変数　初診時CD4 count
- 閾値　200 cells/μL
- アウトカム　死亡

Bor et al. Epidemiology 2014;25:729-37.

不連続回帰デザインの例 (2)

HPVワクチン接種と性行動

- 対象者　カナダ　オンタリオ州の女性
- 介入　HPVワクチン接種の政策
- 割り当て変数　中学２年生になった年
- 閾値　2008年
- アウトカム　妊娠、STD

Smith et al. CMAJ 2015;187:E74-81.

不連続回帰デザインの例 (3)

生活保護費の支給額と医療費

- 対象者　日本　生活保護世帯
- 介入　生活保護費支給額
- 割り当て変数　第１子の年齢
- 閾値　3歳（児童養育加算が減額）
- アウトカム　世帯の医療費

Nishioka et al. J Epidemiol Community Health 2022;76:505-11.

不連続回帰デザイン：まとめ

- 割り当てに関わる連続変数に閾値が存在する場合に、閾値近傍でのアウトカムの不連続な変化を比較する手法
- 検査値により治療が変わる場合や、時期により治療方針が変化した場合などの検証に適している
- 前提条件
  - 割り当てのルールと閾値が明確である
  - 介入以外の要因は閾値付近で連続である
  - 対象者が介入前に割り当て変数を操作できない

Agenda

テーマ

- 不連続回帰デザイン（regression discontinuity）
- 差の差分析　(difference in differences)

内容

- 研究デザインの背景
- 研究デザインの基礎・注意点
- どのような臨床疑問の検証に適しているか
- 活用例

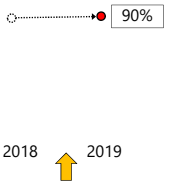
差の差分析：背景

- 経時変化が関係するclinical questionの例
  - ガイドライン改訂により患者のアウトカムは改善したか？
  - 医療の質を改善するプログラムを導入した病院で、臨床評価指標（quality indicator）は改善したか？
  - 新しい保健医療政策の導入により格差は是正されたか？
- 想定されるデザイン：前後比較
- 課題
  - 自然な経時変化の効果を見ているだけではないのか
  - 対照群がない

経時変化を評価する

例：医療の質を改善する“介入”の効果を検証する

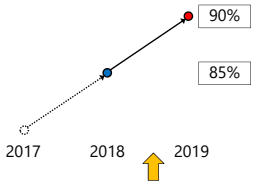
- 方法１：
  導入後の指標を見る
  ⇒ 介入前より改善?悪化?



経時変化を評価する

例：医療の質を改善する“介入”の効果を検証する

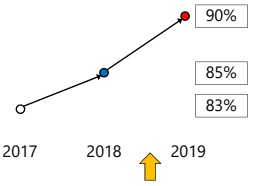
- 方法２：
  導入前後の指標を比較
  ⇒ 介入がなくても自然と改善しているのでは？



経時変化を評価する

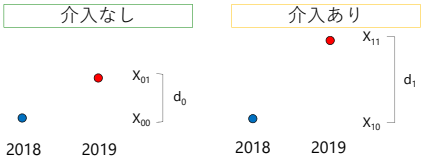
例：医療の質を改善する“介入”の効果を検証する

- 方法３：
  導入前のトレンドを考慮
  ⇒ 介入がなかった対照群ではどうだったか？



差の差（difference in differences）

介入の影響を受けない対照群を見つけ、経時変化を比較



$D = d_1 - d_0 = (X_{11} - X_{10}) - (X_{01} - X_{00})$

差の差分析の実際

交互作用を導入した回帰モデルを利用

アウトカム
$= \alpha + \beta_1 \times (\text{介入}) + \beta_2 \times (\text{時点}) + \beta_3 \times (\text{介入}) \times (\text{時点})$

介入：介入群=1、対照群=0
時点：介入後=1、介入前=0
介入×時点：交互作用項

差の差分析の実際

交互作用を導入した回帰モデルを利用

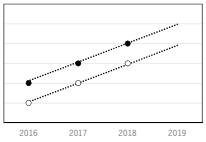
アウトカム
$= \alpha + \beta_1 \times (\text{介入}) + \beta_2 \times (\text{時点}) + \beta_3 \times (\text{介入}) \times (\text{時点})$

	介入前(時点=0)	介入後(時点=1)	後-前	
介入群 (介入=1)	$\alpha + \beta_1$	$\alpha + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$	差の差 β_3
対照群 (介入=0)	α	$\alpha + \beta_2$	β_2	

差の差分析の仮定 (1)

仮定１．平行トレンド (parallel trends)

- 介入群において介入がなかったと仮定した場合、アウトカムの経時変化は対照群と同じ
- 検証方法
  - 介入前の両群のトレンドを図示
  - 介入前で、時期×群の交互作用がないことを確認



差の差分析の仮定 (2)

仮定２．共通ショック (common shocks)

- 介入と無関係にアウトカムに影響する外的要因(shock)が発生した場合、影響の程度が介入群と対照群で同じ

  例：
  - 効果を検証したい政策と同時に施行された、全く別の制度変更
  - 天災、新興感染症

- 検証方法：なし

差の差分析の活用法

研究テーマ

- 医療制度改革の効果
- 院内での取り組みの効果　（前後比較と同様）
- ガイドライン改訂の影響

注意点

- 適切な対照群を選択する　（２つの仮定を満たすか）
- 介入前から、複数時点のデータが必要
- 各群・時点で同等のデータ　（データベース研究に適している）

差の差分析の例 (1)

医師の働き方改革と患者アウトカム

- データ：　米国Medicareデータ
- 介入群：　Teaching hospitals　（レジデントが多い病院）
- 対照群：　Non-teaching hospitals　（レジデントが少ない病院）
- 時期：　2011年のACGME duty hour reforms前後
- アウトカム：　入院患者の30日以内死亡率・再入院率

ACGME = Accreditation Council for Graduate Medical Education

Patel et al. JAMA 2014;312:2364-73.

差の差分析の例 (2)

東日本大震災が「避けられる入院」に与えた影響

- データ：　厚生労働科学研究DPC研究班データベース
- 介入群：　岩手県・宮城県・福島県の病院
- 対照群：　それ以外の病院
- 時期：　2010/7~2011/2 vs 2012/7~2013/2
- アウトカム：　各種ambulatory care sensitive conditionによる入院

Sasabuchi et al. J Epidemiol Community Health 2017;71:248-52.

差の差分析の例 (3)

子ども医療費助成と歯科受診・口腔健康状態

- データ：　熊本県　国民健康保険
- 介入群：　自治体の子ども医療費助成が終了
- 対照群：　〃　継続
- 時期：　2015年3月の前後
- アウトカム：　歯科受診（初回・全受診）、重症歯科疾患の発生

Ono et al. Community Dent Oral Epidemiol 2022 (epub).

差の差分析：まとめ

- 介入群における経時変化と対照群における経時変化の差をとることで、介入の影響を検証する手法
- 分析には、介入×時点の交互作用を加えた回帰モデルを用いる
- 適切な対照群を選択する必要がある
  - 平行トレンドの仮定：　介入群において介入がなかった場合アウトカムの経時変化は対照群と同じ
  - 共通ショックの仮定：　外的要因が発生した場合、影響の程度は介入群と対照群で同じ

NDB・DPCデータベース研究人材育成Webinar

時間依存性交絡と周辺構造モデル

東京大学大学院　医学系研究科　臨床疫学・経済学
大邉寛幸

目次

1. 時間依存性交絡について
2. 時間依存性交絡の対処法と周辺構造モデル
3. 周辺構造モデルの逆確率重み付け法に必要な仮定
4. 周辺構造モデルの逆確率重み付け法の実際の手順

目次

1. 時間依存性交絡について
2. 時間依存性交絡の対処法と周辺構造モデル
3. 周辺構造モデルの逆確率重み付け法に必要な仮定
4. 周辺構造モデルの逆確率重み付け法の実際の手順

周辺構造モデルの逆確率重み付け法を用いた最初の論文

HIV陽性の男性に対するZidovudineは死亡を減らすか？

Design	観察研究
Patient	HIV陽性男性で、AIDSを発症しておらず、治療薬も投与されていない
Exposure	Zidovudineの投与
Control	Zidovudineの非投与
Outcome	死亡

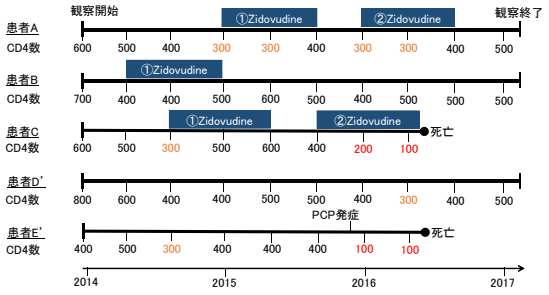
Hernán MA, Brumback B, Robins JM. Epidemiology. 2000;11(5):561-70.

HIV陽性の男性に対するZidovudineは死亡を減らすか？

1. AIDS発症していない場合
   (1)CD4陽性Tリンパ球数が350/uLより多い
   →治療を行わず経過観察する
   (2)CD4陽性Tリンパ球数が350/uL以下
   →治療を開始する
2. AIDS発症している場合
   →治療を開始するが、カリニ肺炎など重篤な場合はその治療を優先する場合がある

抗HIV治療ガイドラインより

HIV陽性の男性に対するZidovudineは死亡を減らすか？



粗解析

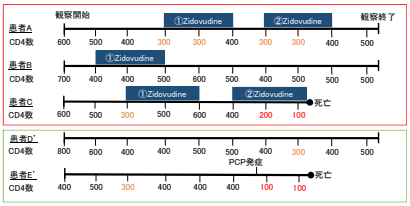
Crude

アウトカム：死亡

治療群：患者A,B,C

コントロール群：患者D’,E’

共変量：なし



Zidovudineは死亡率をハザード比3.6倍上昇させる

Hernán MA, Brumback B, Robins JM. Epidemiology. 2000;11(5):561–70.

標準的な統計解析方法

Cox比例ハザードモデル

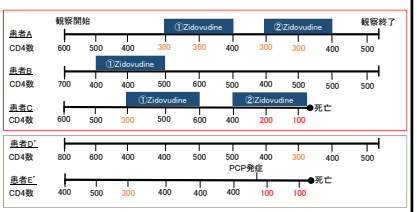
アウトカム：死亡

治療群：患者A,B,C

コントロール群：患者D’,E’

共変量：

観察開始時点での年齢,性別,既往症,CD4数など



Zidovudineは死亡率をハザード比2.3倍上昇させる

Hernán MA, Brumback B, Robins JM. Epidemiology. 2000;11(5):561–70.

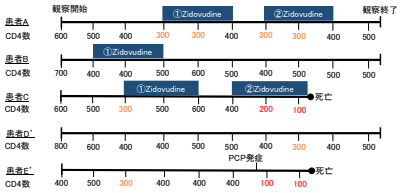
標準的な統計解析ではバイアスを生じる

時間依存性交絡(=CD4数)が存在するため

Hernán MA, Brumback B, Robins JM. Epidemiology. 2000;11(5):561–70.

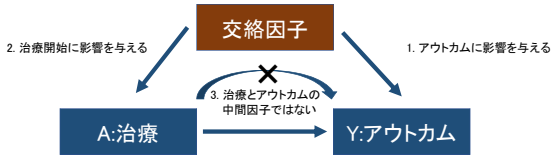
時間依存性変数(Time varying variables)

観察期間中に何度も測定し、時間に依存して変化していく変数



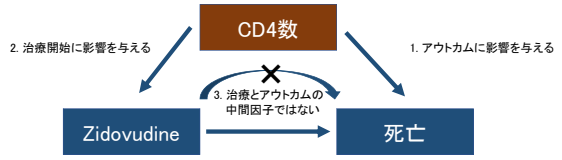
交絡因子(Confounder)

興味あるアウトカム（例えば死亡）のリスク因子（予測因子）かつ、治療開始に影響する（予測する）変数で中間因子でない



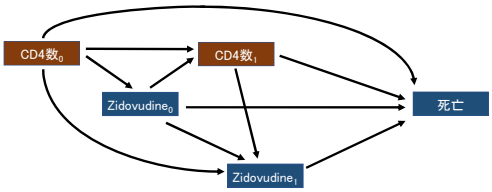
交絡因子(Confounder)

CD4数はZidovudineの開始を予測し、
かつ死亡のリスク因子であるので交絡因子である



時間依存性交絡(Time dependent confounder)

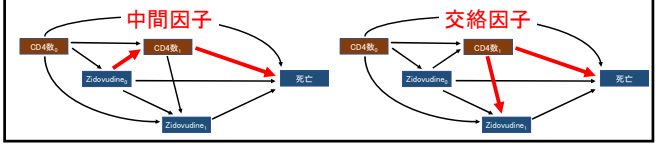
時間依存性交絡＝時間依存性変数＋交絡因子



時間依存性交絡の問題①過調整バイアス

治療によって変化する中間結果（CD4数$_1$）は

- 中間因子なので
  →調整すると治療効果の推定にバイアス（過調整バイアス）
- 交絡因子なので
  →調整しないと治療効果の推定にバイアス

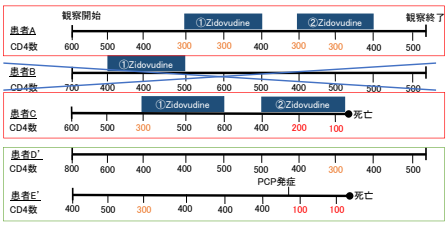


時間依存性交絡の問題②選択バイアス

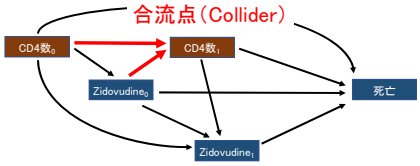
CD4数が300以下の患者のみ含める

↓

患者Bのように初回Zidovudine効果を無視



時間依存性交絡の問題②選択バイアス



*合流点層別バイアス（Collider stratification bias）とも呼ぶ

時間依存性交絡の問題まとめ

①過調整バイアスや②選択バイアスのため
通常の層別解析・回帰モデル・傾向スコア法では
バイアス補正ができない

BMJ. 2017 Oct 16;359:j4587.
Epidemiology. 2009;20(4):488–95.
Epidemiology. 2003;14(3):300–6.

時間依存性交絡の例1

例1: 高齢腎不全患者に対する慢性維持透析と死亡との関連

高齢の腎不全患者に対して慢性維持透析を行う場合と行わない場合のどちらが、生命予後が良いか？

慢性維持透析はeGFRの値を基準に導入を決定する。

eGFRは死亡と関連し、透析という治療によりその値が変化。

つまり、eGFRは時間依存性交絡因子、維持透析の開始は時間依存性治療、となる。

Kidney Int. 2018;94(3):582–8.

時間依存性交絡の例2

例2:炎症性腸疾患に対する免疫抑制剤と重症感染症との関連

炎症性腸疾患に対する免疫抑制剤（thiopurine単独, anti-TNF単独, 併用）の種類により、重症感染症の発症頻度は変わるか？

免疫抑制剤の実施は、炎症性腸疾患の活動性により決定される。

炎症性腸疾患の活動性は重症感染症と関連し、免疫抑制剤という治療により変化する。

つまり、炎症性腸疾患の活動性は時間依存性交絡因子、免疫抑制剤治療は時間依存性治療、となる。

Gastroenterology. 2018;155(2):337-346.e10.

目次

1. 時間依存性交絡について
2. 時間依存性交絡の対処法と周辺構造モデル
3. 周辺構造モデルの逆確率重み付け法に必要な仮定
4. 周辺構造モデルの逆確率重み付け法の実際の手順

時間依存性交絡の対処法

1. 周辺構造モデルの逆確率重み付け法（MSM-IPW）
2. パラメトリックG-formula
3. 構造ネストモデルのG推定法

Int J Epidemiol. 2017 01;46(2):756-62.

時間依存性交絡の対処法の比較

	利点	欠点
MSM-IPW	標準的な統計手法と似ていて理解しやすい ほとんどの統計ソフトで利用可能 暴露割付けの理由がわかっている場合に有用	極端な重みの存在下では不安定 曝露と時間変化する交絡因子の間の相互作用を研究するためにはあまり有用ではない
G-formula	複数のリスク因子に対する介入（共同介入）および動的介入を検討する研究に最適 リスク比やリスク差など、関心のある因果関係の測定値を計算する	計算量が多く、余分なプログラミングを必要とし、過適合問題を引き起こす可能性がある 結果と同様に交絡因子のモデルが必要 “G null paradox”
G推定法	曝露と時間的に変化する交絡因子の間の相互作用を研究するのに有用	計算量が多く、余分なプログラミングを必要とし、適合性の問題を引き起こす可能性がある

時間依存性交絡の対処法

1. 周辺構造モデルの逆確率重み付け法（MSM-IPW）
2. パラメトリックG-formula
3. 構造ネストモデルのG推定法

時間依存性交絡の対処としてMSM-IPWが最も広く使用され臨床医にも理解しやすい

重要用語の説明

- 潜在アウトカム
- 周辺構造モデル
- 因果推論
- 疑似集団

潜在アウトカム

	実際に治療を受けたかどうか	治療を受けた場合のアウトカム	治療を受けなかった場合のアウトカム
Aさん	YES	1	?
Bさん	NO	?	1
Cさん	NO	?	0

もし"?"を知りたければタイムマシンを使って治療受けるかどうか以前に戻らないといけないが、実世界では観測できない（神のみぞ知る）。
それぞれの治療パターンにおけるアウトカムを潜在アウトカムと呼ぶ。

潜在アウトカム

	実際に治療を受けたかどうか	治療を受けた場合のアウトカム	治療を受けなかった場合のアウトカム
Aさん	YES	1	?
Bさん	NO	?	1
Cさん	NO	?	0

"?"は実世界では観測できない（神のみぞ知る）ため個人に対して、直接因果を比較することはできない。
個人は調べられないから集団を対象として、疑似集団を作成して因果推論を行う。

周辺構造モデル(Marginal Structural Model)

- 因果推論のモデルの1つ
- 潜在アウトカムモデル

→時間依存性交絡の調整に応用

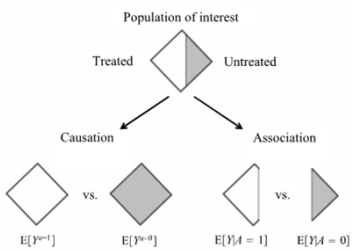
Epidemiology 2000;11:550–60.

周辺構造モデル(Marginal Structural Model)

- 周辺（マージナル）=『潜在アウトカムの周辺分布モデル』
- 構造（ストラクチャー）=『潜在アウトカムに対する回帰モデル』

潜在アウトカムという概念を使用した
疑似集団全体での回帰モデル

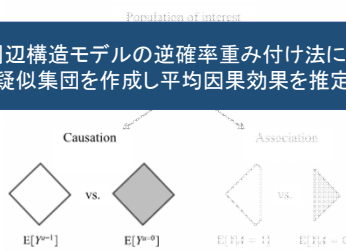
因果推論



Hernán MA, Robins JM (2020). Causal Inference.

因果推論

周辺構造モデルの逆確率重み付け法によって
疑似集団を作成し平均因果効果を推定する



Hernán MA, Robins JM (2020). Causal Inference.

目次

1. 時間依存性交絡について
2. 時間依存性交絡の対処法と周辺構造モデル
3. 周辺構造モデルの逆確率重み付け法に必要な仮定
4. 周辺構造モデルの逆確率重み付け法の実際の手順

周辺構造モデルの逆確率重み付け法に必要な仮定

1. Consistency –治療方法は単一であり、かつその効果は単一である
2. Exchangeability –未測定の交絡因子が無い
3. Positivity –全ての患者層で、治療を受けた人と受けなかった人が存在する必要がある
4. No misspecification of the model –特に重み作成のモデルが正しい

Pharmacoepidemiol Drug Saf. 2014 June ; 23(6): 560–571.
Am J Epidemiol. 2008;168(6):656–664

周辺構造モデルの逆確率重み付け法に必要な仮定

1. Consistency –治療方法は単一であり、かつその効果は単一である
2. Exchangeability –未測定の交絡因子が無い
3. Positivity –全ての患者層で、治療を受けた人と受けなかった人が存在する必要がある
4. No misspecification of the model –特に重み作成のモデルが正しい

Pharmacoepidemiol Drug Saf. 2014 June ; 23(6): 560–571.
Am J Epidemiol. 2008;168(6):656–664

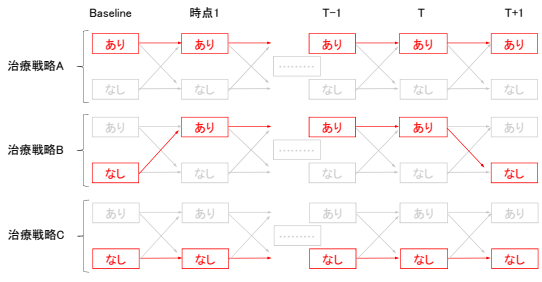
時間依存性治療(Time varying treatment)

- 対義語：Time fixed treatment　0 時点で治療状況を決定する
- 2つの時間依存性治療戦略とRandom treatment strategy
  1. Dynamic treatment strategy（動的な治療戦略）
     時間依存性共変量に依存して治療が決定される（**Treatment-by-indication**）
     “CD4数が300以下になった時に治療を開始する”など
  2. Static treatment strategy（静的な治療戦略）
     時間依存性共変量に依存しない
     “1ヶ月置きに治療を行う”、”最初の月以外は治療を行う”など

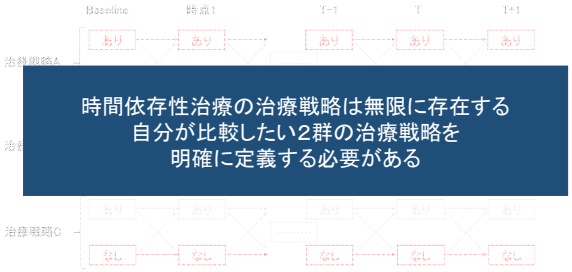
時間依存性治療(Time varying treatment)

- Random treatment strategy
  - より一般化して、特定の値で治療に割りつけるのではなく、治療確率を考慮するストラテジー（静的にも動的にもなり得る）
    - 例: CD4が300なので、確率0.7で治療を行い0.3で治療を行わない

時間依存性治療(Time varying treatment)



時間依存性治療(Time varying treatment)



周辺構造モデルの逆確率重み付け法に必要な仮定

1. Consistency –治療方法は単一であり、かつその効果は単一である
2. Exchangeability –未測定の交絡因子が無い
3. Positivity –全ての患者層で、治療を受けた人と受けなかった人が存在する必要がある
4. No misspecification of the model –特に重み作成のモデルが正しい

Pharmacoepidemiol Drug Saf. 2014 June ; 23(6): 560–571.
Am J Epidemiol. 2008;168(6):656–664

Exchangeability

- Exchangeability = No unmeasured confounding
- この仮定は観測されたデータからは検証できない
- Unmeasured confoundingに対する感度解析が推奨されるが臨床上重要な交絡をきちんと測定し、解析の中に組み込むことが大事

Stat Med. 2004 Mar 15;23(5):749–67

周辺構造モデルの逆確率重み付け法に必要な仮定

1. Consistency –治療方法は単一であり、かつその効果は単一である
2. Exchangeability –未測定の交絡因子が無い
3. Positivity –全ての患者層で、治療を受けた人と受けなかった人が存在する必要がある
4. No misspecification of the model –特に重み作成のモデルが正しい

Pharmacoepidemiol Drug Saf. 2014 June ; 23(6): 560–571.
Am J Epidemiol. 2008;168(6):656–664

Positivity

- 全ての交絡因子の患者層において、治療を受けた人と受けなかった人が存在している必要がある
- Positivityの存在下では重みが極端な値をとってバイアスを生じる
- 多くの交絡因子をいれることでPositivityの可能性が上昇
  →Bias–variance tradeoff
- Positivityのチェックを行う
- 重みをTruncationした感度解析（99%位、95%位など）を行う

Am J Epidemiol. 2008;168(6):656–664

周辺構造モデルの逆確率重み付け法に必要な仮定

1. Consistency –治療方法は単一であり、かつその効果は単一である
2. Exchangeability –未測定の交絡因子が無い
3. Positivity –全ての患者層で、治療を受けた人と受けなかった人が存在する必要がある
4. No misspecification of the model –特に重み作成のモデルが正しい

Pharmacoepidemiol Drug Saf. 2014 June ; 23(6): 560–571.
Am J Epidemiol. 2008;168(6):656–664

No misspecification of the model

- 重みにより、測定された交絡(measured confounders)と独立なpseudo-population(擬似集団)を作り出す
- 重みの分布を描く事がとても大事
- MSM-IPWの信頼性は、重みを正確にモデルできているかに依存(治療の重み、打ち切りの重み)

Am J Epidemiol. 2008;168(6):656-664

目次

1. 時間依存性交絡について
2. 時間依存性交絡の対処法と周辺構造モデル
3. 周辺構造モデルの逆確率重み付け法に必要な仮定
4. 周辺構造モデルの逆確率重み付け法の実際の手順

MSM-IPWの実際の手順

1. 治療に対する逆確率の重み(IPTW)を作成する
   - 1_1. 安定化された重み(Stabilized treatment weight)の作成
   - 1_2. 打ち切りの重み(censoring weight)の作成
   - 1_3. 最終的な重み(overall weight)の作成
2. IPTWで重み付けしたアウトカム-治療回帰モデルを適用する

The Stata Journal 2004;4:402-420.

MSM-IPWを用いた最初の論文(再掲)

HIV陽性の男性に対するZidovudineは死亡を減らすか？

Design	観察研究
Patient	HIV陽性男性で、AIDSを発症しておらず、治療薬も投与されていない
Exposure	Zidovudineの投与
Control	Zidovudineの非投与
Outcome	死亡

Hernán MA, Brumback B, Robins JM. Epidemiology. 2000;11(5):561-70.

1. 治療に対する逆確率の重み(IPTW)を作成する

1_1. 治療の重み(treatment weight)の作成

時間非依存性変数

年齢
西暦
観察開始時の検査値
(CD4数,CD8数,WBC,RBC,Plt)
観察開始時の症状の有無
(発熱,口腔内カンジダ,下痢,
体重減少,口唇ヘルペス)

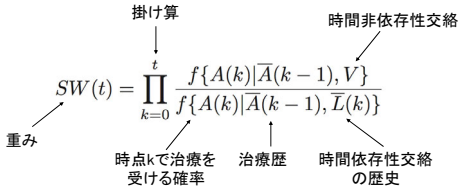
時間依存性変数

検査値
(CD4数,CD8数,WBC,RBC,Plt)
症状の有無
(発熱,口腔内カンジダ,下痢,
体重減少,口唇ヘルペス)
AIDS疾患の有無

1. 治療に対する逆確率の重み(IPTW)を作成する

1_1. 治療の重み(treatment weight)の作成

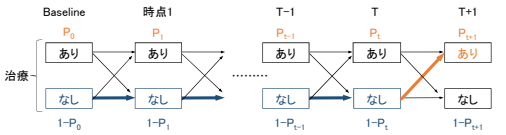
治療を受ける時点ごとに傾向スコアを算出し、最終的に掛け算



1. 治療に対する逆確率の重み(IPTW)を作成する

1_1. 治療の重み（treatment weight）の作成

治療を受ける時点ごとに傾向スコアを算出し、最終的に掛け算



時点T+1で治療を受ける確率＝$(1-P_0) \times (1-P_1) \times \cdots\cdots \times (1-P_{t-1}) \times (1-P_t) \times P_{t+1}$

$$『SW(t)』= \frac{P_0 \times (t+1)}{(1-P_0) \times (1-P_1) \times \cdots\cdots \times (1-P_{t-1}) \times (1-P_t) \times P_{t+1}}$$

1. 治療に対する逆確率の重み(IPTW)を作成する

1_2.打ち切りの重み（censoring weight）の作成

Failure timeより時点kは小さいという条件

時間非依存性交絡

$$SW^{\dagger}(t) = \prod_{k=0}^{t} \frac{\Pr\{C(k)=0 | \overline{C}(k-1)=0, \overline{A}(k-1), V, T>k\}}{\Pr\{C(k)=0 | \overline{C}(k-1)=0, \overline{A}(k-1), \overline{L}(k-1), T>k\}}$$

重み

時点kでCensoringしない確率

時点k−1でCensoringしていない歴史

治療歴

時間依存性交絡の歴史

1. 治療に対する逆確率の重み(IPTW)を作成する

1_3. 最終的な重み（overall weight）の作成

$$SW(t) \times SW^{\dagger}(t)$$

治療の重み（treatment weight）　打ち切りの重み（censoring weight）

1. 治療に対する逆確率の重み(IPTW)を作成する

1_3. 最終的な重み（overall weight）の作成

重みの分布の確認とTruncationを考慮

治療の重み（treatment weight）　打ち切りの重み（censoring weight）

2. IPTWで重み付けしたアウトカム−治療回帰モデルを適用する

- 『最終的な重み(obverall weight)』により、擬似集団を作り出す
- その疑似集団にどのような周辺構造モデルを当てはめるかで、求まる治療効果は異なる
- 周辺構造ロジスティックモデル
  →Marginal Structural Logistic Model
- 周辺構造Cox比例ハザードモデル
  →Marginal Structural Cox proportional hazards Model

2. IPTWで重み付けしたアウトカム−治療回帰モデルを適用する

- 同じ人が何回も出てくるので、クラスターを考慮
- 治療の重みを作成する際に分子に投入した変数は、全てアウトカムモデルに投入

Stat Med. 1990 Dec;9(12):1501−15.

HIV陽性の男性に対するZidovudineは死亡を減らすか

	ハザード比（95%信頼区間）
Crudeハザード比	3.55 (2.95–4.27)
Cox比例ハザードモデル	2.32 (1.92–2.81)
周辺構造Cox比例ハザードモデル	0.74 (0.57–0.96)

Hernán MA, Brumback B, Robins JM. Epidemiology. 2000;11(5):561–70.

まとめ

- 疫学的な曝露（治療）の多くは時間依存性
- 時間依存性交絡の存在下では、通常の解析方法では結果にバイアス
- 時間依存性交絡・時間依存性治療の対処法は周辺構造モデルの逆確率重み付け法（MSM–IPW）、パラメトリックG–formula、構造ネストモデルのG推定法
- 周辺構造化モデルとは、潜在アウトカムという概念を使用した疑似集団全体での回帰モデルのこと

まとめ

- 周辺構造モデルの逆確率重み付け法は”Consistency”, ”Exchangeability”, “Positivity”, ”No misspecification of the model”を仮定している
- 疑似集団の作成には逆確率重み付け法を用いる
- 重みの作成の後には、重み付き回帰モデルを適用
- 重みの分布の確認を

NDB・DPCデータベース研究人材育成Webinar

生存時間分析における競合リスクモデル

東京大学大学院医学系研究科ヘルスサービスリサーチ講座
道端伸明

本日の内容

- 生存時間分析とは
- 従来の生存時間分析
  - カプランマイヤー法・ログランク検定
  - Cox回帰分析
  - 比例ハザード性の確認
- 競合リスクを考慮した生存時間分析
  - 原因別ハザード比の推定
  - 部分分布ハザード比の推定（Gray検定、Fine and Grayモデル）

なぜ生存時間分析？

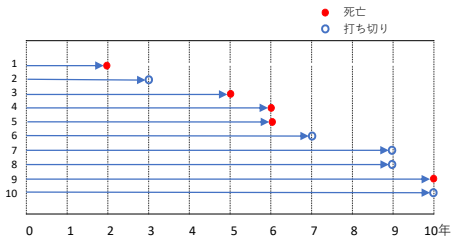
- アウトカムを在院死亡としたロジスティック回帰分析
  入院3日で死亡、入院3か月で死亡。同じアウトカム!?

→ 生存時間解析　アウトカムが発生するまでの時間を考慮

“生存時間”分析　→　生存時間 (survival time) 生死だけでない
例：がんの再発、術後退院までの時間など

用語

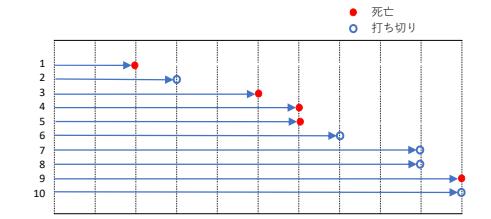
- 生存時間 survival time
  観察開始からアウトカム（イベント・エンドポイント）が発生するまでの時間
- ハザード hazard
  観察期間のある点までアウトカムを発生していなかった人々が次の瞬間アウトカムを発生する確率
- 打ち切り censoring
  研究終了期間前にアウトカムが発生せずに脱落（転居など）したり、研究終了時までアウトカムが発生しない場合を打ち切りとする



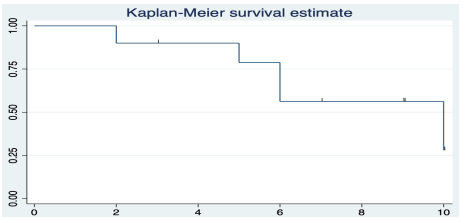
死亡の発生割合　5人/10人 = 50%
死亡の発生率　5人/67人年

カプランマイヤー(Kaplan-Meier)法

- ある時点までに生存している者の割合（生存割合）を表す生存関数をグラフ化する方法。生存曲線。
- アウトカムが発生する度に、その時点での生存割合を算出し、前時点までの生存割合に、時点での生存割合を掛けた累積生存割合を算出する。
- ノンパラメトリック



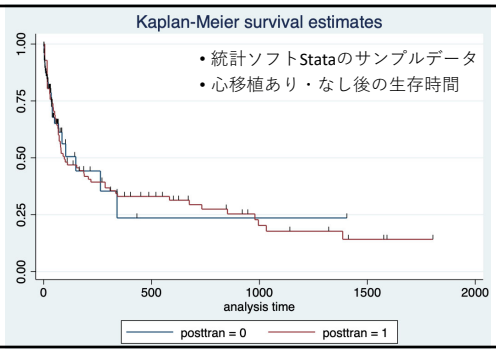
人数	10	10	10	9	8	8	7	5	4	4	2
生存	10	10	9	9	8	7	5	5	4	4	1
時点生存割合	1	1	0.9	1	1	0.875	0.714	1	1	1	0.5
累積生存割合	1	1	0.9	0.9	0.9	0.788	0.563	0.563	0.563	0.563	0.281



人数	10	10	10	9	8	8	7	5	4	4	2
生存	10	10	9	9	8	7	5	5	4	4	1
時点生存割合	1	1	0.9	1	1	0.875	0.714	1	1	1	0.5
累積生存割合	1	1	0.9	0.9	0.9	0.788	0.563	0.563	0.563	0.563	0.281

カプランマイヤー(Kaplan-Meier)法

- 2群間の比較



ログランク(log-rank)検定

```
. sts test posttran

        failure _d:  died
  analysis time _t:  t1
            id:  id

Log-rank test for equality of survivor functions

           |  Events       Events
posttran   |observed     expected
-----------+------------------------
0          |      30        31.20
1          |      45        43.80
-----------+------------------------
Total      |      75        75.00

                chi2(1) =     0.13
                Pr>chi2 =   0.7225
```

- 2群間に有意差は無い。
- 心移植の有無と生存時間に有意な関連があるとは言えない

Cox 回帰分析

- 生存時間分析における多変量回帰モデル
- 比例ハザード性を仮定している
  - −ハザードは時間と共に変化して良い
  - −比較する2つのハザードの比が一定
- セミパラメトリック

Cox 回帰分析

```
Cox regression -- Breslow method for ties

No. of subjects =          103                 Number of obs    =       172
No. of failures =           75
Time at risk    =      31938.1
                                               LR chi2(3)       =     14.38
Log likelihood  =   -291.12672                 Prob > chi2      =    0.0024

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    posttran |   1.001496   .3058315     0.00   0.996     .5504453    1.822152
         age |   1.033265   .0144652     2.34   0.019     1.005299    1.062009
     surgery |   .3305094   .1420734    -2.58   0.010     .1423262    .7675082
------------------------------------------------------------------------------
```

比例ハザード性の確認

- シェーンフィールド残渣
- 二重対数プロット

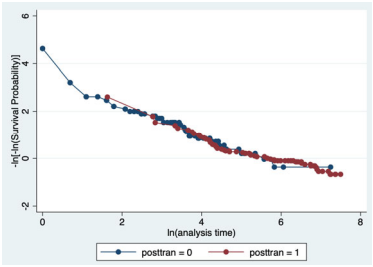
シェーンフィールド残渣 (Schoenfeld residual)

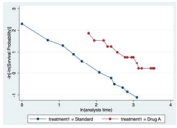
```
Test of proportional-hazards assumption

Time:  Time
                |       rho       chi2       df       Prob>chi2
----------------+-------------------------------------------------
posttran        |   -0.00521       0.00        1          0.9638
age             |    0.12258       1.43        1          0.2314
surgery         |    0.30630       6.18        1          0.0129
----------------+-------------------------------------------------
global test     |                  7.78        3          0.0508
```

※帰無仮説：比例ハザード性が一定 → 棄却してはいけない

二重対数プロット (log-log plot)





白血病寛解研究（Stataサンプルデータ）

比較する2群が視覚的に並行であることを確認する

競合リスクを考慮した生存時間分析

競合リスク(competing risk)とは

- 関心のあるアウトカム（イベント・エンドポイント）以外のアウトカムが生じることにより関心のあるアウトカムが観察できなくなる場合、**競合リスク**があるという。

例１：がんによる死亡がアウトカム
　　心血管障害で死亡→　競合リスク
例２：維持透析患者の死亡
　　腎移植により維持透析から離脱→　競合リスク

心血管障害による死亡　→→→　心血管障害で死亡しなければ、がんで死亡していたはず

従来の生存時間解析との違い

- カプランマイヤー法、Cox回帰はランダムな打ち切りを仮定
- 競合リスクが有る場合はランダムな打ち切りと仮定できない

- 心血管障害で死亡した患者と、観察を継続できた患者とは
  関心のあるアウトカムの発生割合が異なる可能性がある

競合リスクが有る場合のバイアス

- 競合リスクによる打ち切りを受けた参加者が
  （打ち切りを受けずに観察を継続できたとした場合）
  関心のあるアウトカムの発生割合が
  その後、他の観察を継続できている参加者の発生割合と

1. 同じ場合
   → バイアスはなし
2. 比較して高い場合
   → 生存割合を過大評価してしまう
3. 比較して低い場合
   → 生存割合を過小評価してしまう

競合リスクの対処法

- 複合エンドポイント (composite endpoint)
  → 心血管イベントまたはがんによる死亡

- 統計解析による対処法
  –原因別ハザード比の推定
  –部分分布ハザード比の推定

生存時間分析における統計手法

解析目的	従来の生存時間解析	競合リスクを考慮した生存時間解析	
		原因別ハザード比の推定	部分分布ハザード比の推定
リスクの推定（グラフの描出）	カプランマイヤー法による生存割合、累積発生割合	累積発生割合	累積発生関数（cumulative incidence function, CIF）
2群（以上）の単純比較	ログランク検定	ログランク検定	Gray検定
多変量回帰モデル	Cox回帰モデル	原因別Cox回帰モデル	Fine and Gray モデル

康永秀生、森田光治良ら. 超絶解説医学論文の難解な統計手法が手に取るようにわかる本. 東京: 金原出版; 2019. より改変

原因別ハザード比の推定

- 競合リスク(アウトカム・イベント)の発生をランダムな
  打ち切りとして扱う方法
- 次のことを仮定
  – 関心があるアウトカム以外は発生しない
  – 競合リスクは互いに独立（関連が無い）
- 実際独立でなく、互いに正の相関があるときには、
  関心があるアウトカムの発生割合を過大評価する
- 原因別ハザードは、ある時点まで、どのアウトカム
  （関心あるアウトカム・競合リスクを含む）も発生していない時、
  次の瞬間に関心のあるアウトカムが発生する確率

- 予後比較よりも、アウトカムの**リスク要因**に適した解析手法

部分分布ハザード比の推定

- 競合リスク(アウトカム・イベント)が発生した場合、
  打ち切りとせず解析上は観察集団に残り続けるよう扱う方法
- 2群（以上）の単純（単変量）比較には、各群の累積発生関数
  (cumulative incidence function, CIF)を推定しGray検定を行う
- 多変量解析（交絡因子を調整した回帰分析）には
  部分分布ハザード比を推定するFine and Grayモデルを用いる
- 部分分布ハザードは、ある時点まで、関心あるアウトカムが
  発生していない時、次の瞬間に関心のあるアウトカムが発生する確率

- アウトカム・イベント毎の**発生確率の推定・予測モデル**作成
- 2群間の**予後を比較**したい場合に適した解析手法

統計ソフトの対応について

- 原因別ハザード比の推定
  - → Stata, R, SASで通常の生存時間分析のコマンドで可能
- 部分分布ハザード比の推定
  - → R, SASは対応している
  - → StataはGray検定以外には対応

原因別ハザード比と部分分布ハザード比の比較

ヨーロッパ腎臓透析移植学会のデータベースに登録された患者（N=73382）を5年間追跡したデータ

関心のあるアウトカム：維持透析患者の死亡

競合リスク：腎移植

- Noordzij M, et al. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant. 2013;28:2670–7.

原因別ハザード比と部分分布ハザード比の比較

5年間時点の累積発生割合	原因別ハザード比	部分分布ハザード比
死亡	60%	51%
腎移植	33%	24%
その他	25%	25%
合計	118%	100%

Noordzij M, et al. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant. 2013;28:2670–7.

原因別ハザード比と部分分布ハザード比の比較

維持透析患者死亡をアウトカムにした多変量解析	原因別ハザード比	部分分布ハザード比
男性	1.04 (1.02-1.07)	1.03 (0.87-1.23)
高齢者	2.57 (2.52-2.63)	3.47 (3.39-3.55)

性別は競合リスクである腎移植と関連が低いためどちらの方法でもほぼ同じ結果。
一方、高齢者は腎移植を受けにくくなるため、原因別ハザード比では過小評価される。

Noordzij M, et al. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant. 2013;28:2670–7.

論文紹介

P: 介護老人保健施設入所者（N=342758）

O: 在宅復帰

競合リスク：死亡、病院へ入院、他介護施設への移動

在宅復帰に関連するリスク要因の探索が目的、原因別ハザード比の推定を行った。

結果：追跡期間中央値は137日で、19%の入所者が在宅復帰した。

原因別Cox回帰分析の結果、高齢、高い要介護度、複数の合併症、介護施設の運営形態が私立、施設規模が大きいなどが有意に在宅復帰を妨げる要因として関連していた。

- Morita K, et al. Factors Affecting Discharge to Home of Geriatric Intermediate Care Facility Residents in Japan. J Am Geriatr Soc. 2018;66:728–34.

論文紹介

P: 初回冠動脈バイパス術/弁置換・形成術を受けた65歳以上心房細動患者（N=10524）

E: 左心耳閉鎖術併用あり（n=3892）

C: 左心耳閉鎖術併用なし（n=6632）

O: 血栓塞栓症の発症

競合リスク：死亡

結果：追跡期間平均2.6年、全体の5.4%に血栓塞栓症発症し、全体の21.5%が死亡。

Fine and Grayモデルを用いた調整済み部分分布ハザード比は、0.67 (95%信頼区間 0.56-0.81)と、

左心耳閉鎖術併用は、血栓塞栓症の発症を減少させる可能性を示唆した。

- Friedman DJ, et al. Association Between Left Atrial Appendage Occlusion and Readmission for Thromboembolism Among Patients With Atrial Fibrillation Undergoing Concomitant Cardiac Surgery. JAMA. 2018;319:365–74.

参考文献

1. 康永秀生、森田光治良ら. 超絶解説医学論文の難解な統計手法が手に取るようにわかる本. 東京: 金原出版; 2019.
2. Noordzij M, et al. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant. 2013;28:2670–7.
3. Morita K, et al. Factors Affecting Discharge to Home of Geriatric Intermediate Care Facility Residents in Japan. J Am Geriatr Soc. 2018;66:728–34.
4. Friedman DJ, et al. Association Between Left Atrial Appendage Occlusion and Readmission for Thromboembolism Among Patients With Atrial Fibrillation Undergoing Concomitant Cardiac Surgery. JAMA. 2018;319:365–74.

NDB・DPCデータベース研究人材育成Webinar

欠損値の取扱と多重代入法

東京大学大学院医学系研究科　生物統計情報学講座

麻生将太郎

Contents

欠損値

例えばカルテからデータ収集する…
既往歴から高血圧の記載は、
「高血圧あり」、「高血圧なし」、「」（聞いていない）のいずれか

入院前のADLの記載は、
「自立」、「一部介助」、「全介助」、「」（聞いていない）のいずれか

受診時の呼吸数は、
「1～∞の数値」、「」(測定していない)

ID	高血圧	ADL	呼吸数
1	「あり」	「自立」	「15」
2	「なし」	「」	「」
3	「」	「一部介助」	「」
4	「あり」	「」	「18」
5	「あり」	「全介助」	「」
…	…	…	…

「」が欠損値

欠測が生じるには理由がある

理由で欠損値の種類が決まる

欠損値のメカニズムによる分類

- MCAR
- MAR
- MNAR

MCAR

- Missing completely at random (完全にランダムな欠測)
- データの欠測が完全にランダムに発生している状況

MCARの例

- 体重計が壊れた
  - → 体重が測れない
  - → 体重の値が欠測
- 体重の値、他の変数(年齢や性別)で欠測したわけではない（偶然）

ID	体重	年齢	性別
1	40	30	男
2	(35)	60	女
3	60	65	女
4	(65)	35	男

MAR

- Missing at random (ランダムな欠測)
- データの欠測が完全にランダムではない
- 欠測した変数自体で欠測した理由は説明できない
- しかし、他の変数で欠測を説明できる

MARの例

- 若年者で体重の欠測が多く発生
  - → 若年者は体重測定しない
- 体重の値自体で欠測したわけではない
- しかし、年齢によって欠測が発生

ID	体重	年齢	性別
1	(35)	30	女
2	40	60	男
3	60	65	女
4	(65)	35	男

MNAR

- Missing not at random (ランダムではない欠測)
- データの欠測する理由が、欠測した変数自体で説明できる

MNARの例

- 肥満の人の体重の欠測が多く発生
  - → 肥満の人は測定しない傾向
- 体重の値自体で欠測が生じる

ID	体重	年齢	性別
1	40	30	男
2	(80)	60	女
3	45	65	女
4	(90)	35	男

欠損値の種類を見分ける

- MCAR、MAR、MNARを客観的に判断するのは難しい
- 欠損値を予測する変数を増やすことでMARの仮定を満たす確率上昇
- Little's test → MCARの判断 → 帰無仮説がMCAR

  → P≧0.05でMCAR（P<0.05でMAR or MNAR）

対処法

- 完全ケース分析（complete case analysis）
- 欠損値を表す指標用いる方法
- 単一代入法
- 多重代入法 (multiple imputation)

完全ケース分析（complete case analysis）

- 欠測のない対象者のみで解析
- MCARでは結果は変わらない

欠点

- 母集団を代表していない
- MNARの欠損値を除外するとバイアス
- 症例数が少 → 検出力が低下
  → 信頼区間が広くなる

ID	体重	年齢	性別
1	40	30	男
2	(80)	60	女
3	45	65	女
4	(90)	35	男

欠損値を表す指標を用いる方法

- 欠損値に適当な値を代入してカテゴリーとするような方法

ID	高血圧	ADL	呼吸数
1	1	1	15
2	0		
3		2	
4	1		18
5	1	3	
…	…	…	…

→

ID	高血圧	ADL	呼吸数
1	1	1	15
2	0	9	99
3	9	2	99
4	1	9	18
5	1	3	99
…	…	…	…

変数	オッズ比	標準誤差	P値
高血圧			
なし	Reference		
あり	2.42	3.07	0.488

→

変数	オッズ比	標準誤差	P値
高血圧			
なし	Reference		
あり	2.42	3.07	0.488
欠損	9.15	6.32	0.001

- 高血圧あり、なしの2群 → あり、なし、欠損の3群にする
- 欠損群は高血圧なし群に比べて、オッズ比9倍？
- 解釈困難な結果

推奨しない

代入法

欠測値を何らかの値に置き換えてデータを疑似完全状態にする

- 単一代入法
- 多重代入法

単一代入法

- 平均値代入法
- 比率代入法
- Hot-deck法
- Cold-deck法
- 回帰代入法
- LOCF法

平均値代入法

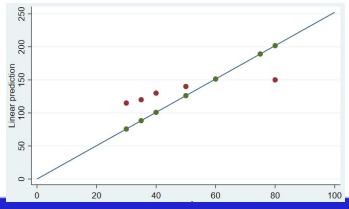
- データのある集団から平均値を求め、欠損値を平均値に置き換える
- 百害あって一利なし!!!

ID	高血圧	ADL	呼吸数
1	1	1	15
2	0	2	16.5
3	0.75?	2	16.5
4	1	2	18
5	1	3	16.5

比率代入法

- 欠損値を変数と関連する変数を切片なしで単回帰して予測した値を代入
- 1つの変数との関連のみ

ID	年齢	血圧
1	50	140
2	35	120
3	60	
4	80	150
5	75	
6	40	130
7	30	115



ID	年齢	血圧
1	50	140
2	35	120
3	60	151
4	80	150
5	75	189
6	40	130
7	30	115

Hot-deck法

- 欠損値のある属性に近いデータを同じデータセットから見つけ、その値を代入
- 例えば、血圧が欠損の65歳女性は、同じデータセットの中から

  65歳女性を探し出して、その血圧の値を欠損値を代入
- 同じ属性が複数の場合は、その中から無作為に1件を選択

ID	性別	年齢	血圧
1	男	65	160
2	女	30	90
3	女	65	130
4	男	40	120
5	女	65	

➡

ID	性別	年齢	血圧
1	男	65	160
2	女	30	90
3	女	65	130
4	男	40	120
5	女	65	

➡

ID	性別	年齢	血圧
1	男	65	160
2	女	30	90
3	女	65	130
4	男	40	120
5	女	65	130

Cold-deck法

- 欠損値のある属性に近いデータを異なるデータセットから見つけ、その値を代入
- 例えば、血圧が欠損の65歳女性は、異なるデータセットの中から

  65歳女性を探し出して、その血圧の値を欠損値を代入
- 同じ属性が複数の場合は、その中から無作為に1件を選択

データセットB

ID	性別	年齢	血圧	体重
1	男	20	120	60
2	女	40	100	70
3	男	65	150	80
4	女	65	140	50
5	男	80	160	60

データセットA

ID	性別	年齢	血圧	脈拍
1	男	65	160	80
2	女	30	90	70
3	女	65	130	100
4	男	40	120	60
5	女	65		80

データセットA

ID	性別	年齢	血圧	脈拍
1	男	65	160	80
2	女	30	90	70
3	女	65	130	100
4	男	40	120	60
5	女	65	140	80

回帰代入法

- 欠損値のある変数を他の変数で回帰分析してモデルを作成
- モデルから推定値を出して欠損値に代入する
- モデルに誤差をランダムに加える場合もある（確率的代入法）

呼吸数＝a×性別+b×年齢+c×血圧+d×脈拍+e×体温+f

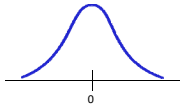
ID	性別	年齢	血圧	脈拍	呼吸数	体温
1	男	30	120	60	15	36.5
2	女	45	120	80	16.9	37.0
3	男	50	140	70	10	38.0
4	女	40	110	65	20	37.5
5	男	60	150	90	7.5	36.0

呼吸数＝a×0+b×45+c×120+d×80+e×37.0+f

呼吸数＝a×1+b×60+c×150+d×90+e×36.0+f

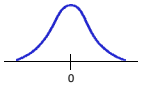
確率的回帰代入法

- 回帰代入法で得られた予測式に誤差項を加えた値を代入
- 誤差項は平均値0、分散が回帰モデルの残差分散の正規分布から無作為に抽出



※残差分散：実測値回帰の予測値の差の平方和を(データ数−変数の数−1)で割った値

呼吸数＝a×性別+b×年齢+c×血圧+d×脈拍+e×体温+f+誤差項



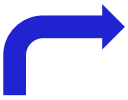
ID	性別	年齢	血圧	脈拍	呼吸数	体温
1	男	30	120	60	15	36.5
2	女	45	120	80	14.9	37.0
3	男	50	140	70	10	38.0
4	女	40	110	65	20	37.5
5	男	60	150	90	9	36.0

呼吸数＝a×0+b×45+c×120+d×80+e×37.0+f-2

呼吸数＝
a×1+b×60+c×150+d×90+e×36.0+f+1.5

LOCF法

- Last observation carried forward法
- 経時データで欠測が起きた場合、最後に観察された値で欠損値を埋める
- 「脱落後の変数の推移は最後に観察された値のまま変化しない」
  強い仮定が必要
- 推定精度を過小評価
- 最悪値で補完する方法もあるが、保守的で強い仮定が必要



ID	時点1	時点2	時点3	時点4	時点5
1	200	1000	5000	3500	2000
2	5000	8000	10000	15000	15000
3	1000	3000	3000	4000	1000
4	10000	8000	8000	8000	6000
5	100	100	100	100	100

ID	時点1	時点2	時点3	時点4	時点5
1	200	1000	5000	3500	2000
2	5000	8000	10000	15000	
3	1000	3000		4000	1000
4	10000	8000			6000
5	100				

単一代入法

- 個々の欠測データを1つの値で代入する方法
- 欠測値5%未満だと良好な結果を得ることもある
- 1つのモデルで推定した推定値を代入するため、不確実性が高い
  (代入値が測定された値に近いか不明)
- 推定値の分散を過小評価（信頼区間狭くなる）→ 誤った結果

推奨しない

多重代入法

なぜ多重代入法か

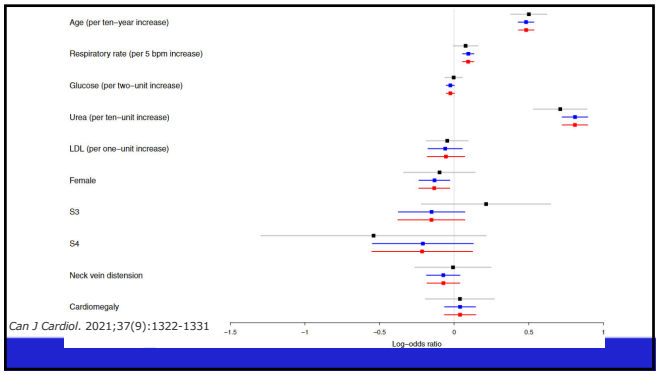
- 欠損値によるバイアスは代入法で是正可能
- 代入法では完全に欠損値を復元できるわけではない(単一でも多重でも)
- 代入法の目的はバイアスを減少し、真の値を推定
- 代入法の誤差を適切に評価する必要

なぜ多重代入法か

- 単一代入法だと代入モデルによって代入した値は変わる
- 単一代入法で得られた代入値の背後には様々な可能性が隠されている
- 一つの値を代入して解析することは不確実性が高い
- 複数の値を代入して不確実性を減少することが必要

多重代入法のメリット

- 不確実性下げる
- バイアスを減らし精度を高める
- Complete case analysisより検出率向上、信頼区間も広くならない



Can J Cardiol. 2021;37(9):1322-1331

多重代入法の手順

① 代入

② 解析

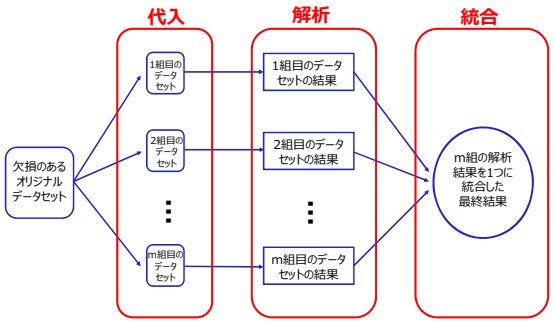
③ 統合

多重代入法の手順

① 代入

② 解析

③ 統合



①代入

- オリジナルデータから欠損を埋めた完全データをm個作成
- 欠損値はMARを仮定

アルゴリズム

- DA (Data Argumentation)法
- EMB (Expectation-Maximization with Bootstrapping)法
- FCS (Fully Conditional Specification)法

アルゴリズム

- ベイズ統計学の枠組みを使用
- どの方法もMARを仮定
- DA法とEMB法は正規分布を仮定
- DA法とFCS法はマルコフ連鎖モンテカルロ法に基く
- 少しだけFCS法の精度が良い

FCS法

- MICE（multivariate imputation by chained equations）
- 柔軟な方法で、多変量の正規分布の仮定が不要（条件付き分布）
- 変数ごとに回帰モデルを作成
- 2値変数、順序変数をアウトカムとしたlogistic回帰などのモデルも使用可

FCS法

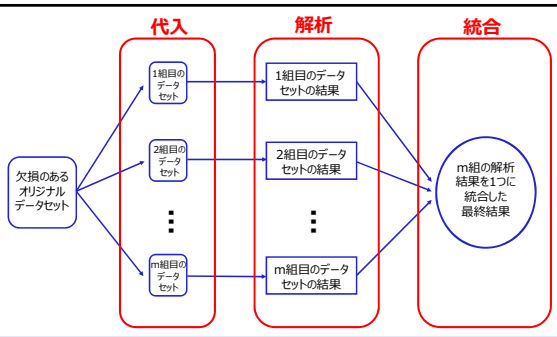
- 「各変数の欠測に対する補完モデルとして、欠測を補完したい変数以外の他のすべての変数によってお互いに説明可能」という強い仮定
- この仮定を満たせない場合、推定値の分散が小さい（信頼区間が狭くなる）

多重代入法の手順

① 代入

② 解析

③ 統合



②解析

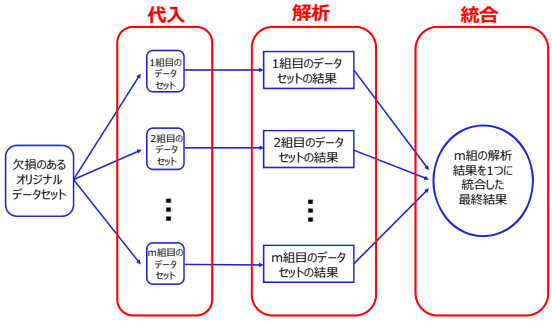
- 作成したm個のデータセットを各々別々に解析
- m個のデータセットをまとめて1つのデータセットとして解析してはいけない

多重代入法の手順

① 代入

② 解析

③ 統合



③統合

- 各データセットで得られた結果を統合する
- 統合の方法はRubin's ruleに従う

Rubin's rule

推定値はm個のデータセットの推定値の平均　$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$

分散はm個のデータセットの分散の平均　$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}U_i$

データセット間の分散　$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i - \bar{Q}\right)^2$

全体の分散　$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$

95%信頼区間　$\bar{Q} \pm 1.96 * \sqrt{T}$

m	データセットの数
$\bar{Q}$	推定値の平均
$\hat{Q}_i$	各データセットの推定値
T	全体の分散
$\bar{U}$	分散の平均
U_i	各データセットの分散
B	データセット間の分散

データセット	推定値	分散
1	10	5
2	15	3
3	7	2
4	6	10
5	12	7

$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$　推定値＝(10+15+7+6+12)/5 ＝ 10

$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}U_i$　分散＝(5+3+2+10+7)/ 5＝ 5.4

$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i - \bar{Q}\right)^2$　データセット間の分散
＝((10-10)^2+(15-10)^2+(7-10)^2
+(6-10)^2+(12-10)^2)/(5-1) ＝ 13.5

$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$　全体の分散＝5.4+(1+1/5)*13.5 ＝ 21.6

Rubin's rule

全体の分散　$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$

$\bar{U} < B$　データセット数を増やすと精度が上がる

$\bar{U} > B$　データセット数を増やしても精度は上がらない

データセット	推定値	分散
1	10	5
2	15	3
3	7	2
4	6	10
5	12	7

推定値＝10

分散＝ 5.4　　データセット間の分散＝ 13.5

$\bar{U} < B$

全体の分散＝5.4+(1+1/5)*13.5 ＝ 21.6

全体の分散　$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$

データセット数を増やすと全体の分散が小さくなる

データセットの数は？

- 開発された当初は5セットで十分とされていた
- しかし、最近では50セット、100セット以上必要という主張も
- ガイドラインでは少なくとも20セット作成が推奨

代入モデルで用いる変数は？

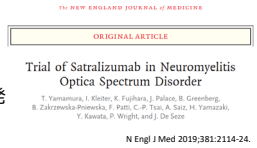
- MARで欠測が起きる状況を説明できる変数（補助変数）を使用
- できるだけ多くの補助変数と欠測を説明するのに重要な変数は必須
- 解析で使用する変数だけでなく、
  解析に使用しなくても欠損に関連する変数も投入

代入モデルで用いる変数は？

- MNARでも、補助変数を多数投入 → MARを満たす確率が上昇
- 説明変数が欠損値でも、代入モデルにアウトカム変数の投入が推奨
- 解析のステップでは代入モデルに含まれていない変数の使用×

アウトカム変数の欠損に使えるか？

- 使えます
- RCTでアウトカムの追跡不能問題から開発
- MARの仮定が重要
- 従属変数と説明変数を数学的に区別しない


THE NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Trial of Satralizumab in Neuromyelitis Optica Spectrum Disorder

T. Yamamura, I. Kleiter, K. Fujihara, J. Palace, B. Greenberg, B. Zakrzewska-Pniewska, F. Patti, C.-P. Tsai, A. Saiz, H. Yamazaki, Y. Kawata, P. Wright, and J. De Seze

N Engl J Med 2019;381:2114-24.

使用例

Antenatal blood pressure for prediction of pre-eclampsia, preterm birth, and small for gestational age babies: development and validation in two general population cohorts

BMJ　2015;101:h15948

妊婦の平均血圧が子癇前症、早産、胎内発育遅延を

予測するモデルの開発と妥当性を研究

記載例

For both the ALSPAC and SWS datasets there were missing data on maternal characteristics and also for the derived values of initial mean arterial pressure and mean arterial pressure at 20, 25, 28, 31, 34, and 36 weeks for some women.

データセットには母体の背景因子と平均血圧に欠損値が含まれる

To increase power and minimise selection bias in the second stage prediction models, we next used multivariable multiple imputation of missing data.

検出力を高め、選択バイアスを最小化するために多重代入法を用いた

For each imputation model, we included all early pregnancy characteristics and derived mean arterial pressure values as well as the outcomes.

モデルには患者背景、平均血圧、アウトカムも投入

記載例

For each imputation we generated 20 imputed datasets and combined coefficient estimates across these using Rubin's rules.

20の完全なデータセットを作成 Rubin's ruleで結果を統合

Imputation of missing data in this way assumes that data are missing at random—that is, that any reasons for missingness are explained by the observed variables included in the imputation model.

欠損値はMARを仮定して代入 MARは他の変数で欠損が説明できる

The assumption seems reasonable here as most women will not be aware of their current blood pressure (only their previous values) so this is unlikely to influence their decision to have their blood pressure measured at any given gestational age.

MARの仮定は妥当である説明

最近は、当然の如く、あっさり説明

Childhood Cardiovascular Risk Factors and Adult Cardiovascular Events

All primary analyses were performed after multiple imputation of missing values by means of chained equations with fully conditional specification (10 replications) in PC-SAS software (version 9.4, SAS Institute); data were assumed to be missing at random.[11]

多重代入法を行い解析している(10セット複製)。MARを仮定している。

N Engl J Med. 2022;386(20):1877-1888.

まとめ

- 欠測値の種類（MCAR, MAR, MNAR）
- 単一代入法
- 多重代入法

NDB・DPCデータベース研究人材育成Webinar

マルチレベル分析

東京大学大学院医学系研究科リアルワールドエビデンス講座　特任准教授
笹渕　裕介

Agenda

- **マルチレベル分析とは**
- **マルチレベル分析の統計モデル**
- **マルチレベルモデルを利用した研究紹介**

Agenda

- **マルチレベル分析とは**
- マルチレベル分析の統計モデル
- マルチレベルモデルを利用した研究紹介

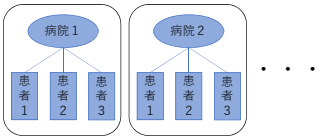
マルチレベル分析とは

階層構造をもつデータを分析する方法

階層構造

複数のレベルをもつデータ

✓病院　ー　患者
✓地域　ー　住民
✓患者　ー　時点
✓病院　ー　患者　ー　時点



・・・

なぜマルチレベル分析が必要なのか

階層構造を考慮しない回帰分析
• 各個人のデータは独立であることを仮定

階層構造をもつデータ
• 同じ集団に属する個人は独立であると言えない
• 同じ患者から得られたデータは似通っている
　　グループ内のデータには相関がある（級内相関）

なぜマルチレベル分析が必要なのか

✓病院　ー　患者

施設によって
- 専門性
- スタッフの習熟度
- 最新技術が利用可能
- 管理体制

などが異なる

⇨

背景が一致する患者を
同じ病院から選ぶと

異なる病院から選ぶ
ときと比較して

アウトカムは似通う

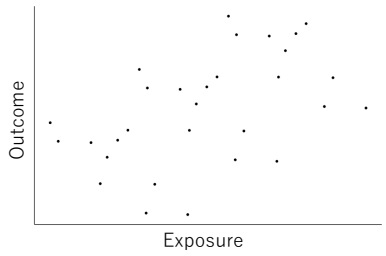
なぜマルチレベル分析が必要なのか

✓患者　ー　時点

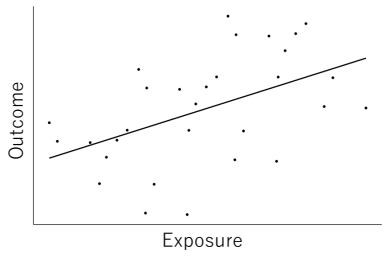
患者の重症度を複数時点にわたって測定

- 最初に測定した重症度はその後の重症度に影響がある
- どこかの時点で差があれば他の時点でも差がつきやすい

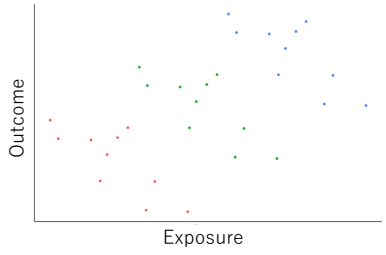
階層を考慮しない分析



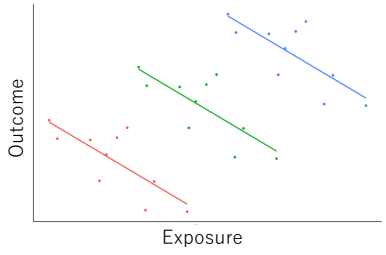
階層を考慮しない分析



階層を考慮した分析



階層を考慮した分析



階層構造を考慮しないと

- 誤った結果を導く
- 第一種の過誤の可能性が増大する

→サンプルをコピーして検出力を上げているかのよう

マルチレベル分析を行う目的

➢階層構造を考慮する
　多施設研究やアウトカムの繰り返し測定
➢グループの特性が個人に与える影響を調べる
　施設の特性が個人に与える影響を知りたい
➢グループ間のばらつきを調べる
　グループ間でのばらつき、あるいはその要因を調べる

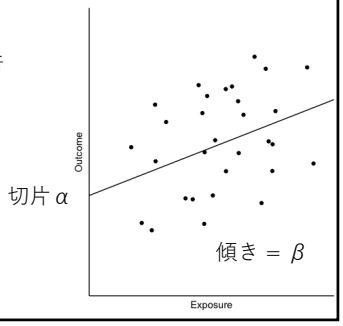
Agenda

- マルチレベル分析とは
- **マルチレベル分析の統計モデル**
- マルチレベルモデルを利用した研究紹介

統計モデル

階層構造を考慮しない回帰分析

$$Y = \alpha + \beta \times x + \varepsilon$$

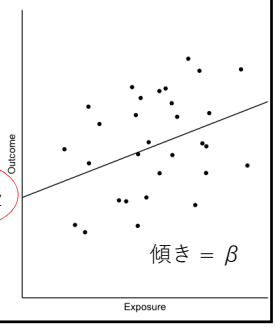


統計モデル

階層構造を考慮しない回帰分析

$$Y = \alpha + \beta \times x + \varepsilon$$

値が一つに定まる



統計モデル

マルチレベルモデル

A病院

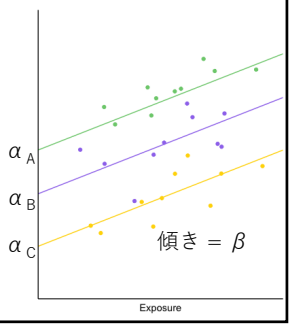
$$Y_A = \alpha_A + \beta \times x + \varepsilon$$

B病院

$$Y_B = \alpha_B + \beta \times x + \varepsilon$$

C病院

$$Y_C = \alpha_C + \beta \times x + \varepsilon$$



統計モデル

マルチレベルモデル

A病院

$Y_A = \alpha_A + \beta \times x + \varepsilon$

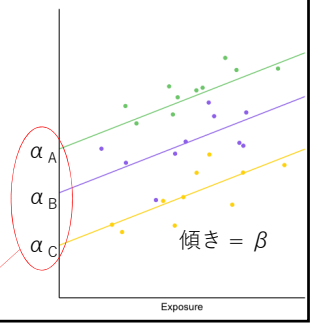
B病院

$Y_B = \alpha_B + \beta \times x + \varepsilon$

C病院

$Y_C = \alpha_C + \beta \times x + \varepsilon$

切片の全体平均＋施設ごとの誤差
全体平均を中心にランダムに分布



統計モデル

マルチレベルモデル

$Y = \alpha + \beta \times x + \varepsilon$

$\alpha = \gamma + u$

γ：切片の全体平均

u：切片のグループごとの誤差

ランダム切片モデル

結果の解釈

$Y = \alpha + \beta \times x + \varepsilon$

$\alpha = \gamma + u$

γ：切片の全体平均

u：切片のグループごとの誤差

係数β：xが1単位増加したときのアウトカムの変化
切片α：各グループのアウトカムの大きさ
誤差uの分散：切片のばらつきの大きさ
分散が0でないならばグループごとに切片が異なる

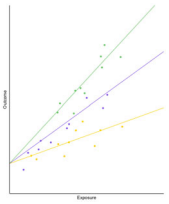
様々なマルチレベルモデル

$Y = \alpha + \beta \times x + \varepsilon$

$\beta = \gamma + u$

γ：傾きの全体平均

u：傾きのグループごとの誤差



ランダム傾きモデル

様々なマルチレベルモデル

$Y = \alpha + \beta \times x + \varepsilon$

$\alpha = \gamma_0 + u_0$

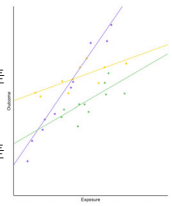
γ_0：切片の全体平均

u_0：切片のグループごとの誤差

$\beta = \gamma_1 + u_1$

γ_1：傾きの全体平均

u_1：傾きのグループごとの誤差



ランダム切片・傾きモデル

Agenda

- マルチレベル分析とは
- マルチレベル分析の統計モデル
- **マルチレベルモデルを利用した研究紹介**

マルチレベル分析の実例

1. 腰椎すべり症に対する硬膜外ステロイド投与の効果
2. 小児の下気道感染による入院後の再入院割合

Epidural Steroid Injections for Management of Degenerative Spondylolisthesis

J Bone Joint Surg Am. 2020;102(15):1297-1304.

背景

- 腰椎変性すべり症は神経性跛行や腰痛の一般的な要因
- 理学療法や硬膜外腔へのステロイド投与が一般的
- ステロイド投与
  - 椎間板ヘルニア→疼痛緩和・手術回避
  - 脊柱管狭窄症→QOL低下と関連
  - 腰椎変性すべり症→効果はわかっていない

研究目的

- 腰椎変性すべり症に対するRCTと前向きコホート研究のデータ

P　腰椎変性すべり症
E　硬膜外ステロイド投与
I　ステロイド投与なし
O　疼痛・身体機能（最大4年まで反復測定）

統計モデル

- 個人－時点
- ランダム切片モデル

  疼痛・身体機能 $= \alpha + \beta \times$ 交絡要因 $+ \varepsilon$

  $\alpha = \gamma + u$

  γ：切片の全体平均

  u：切片の個人毎の誤差

結果

- 非手術患者
  - 疼痛・身体機能とも硬膜外ステロイド投与により悪化
- 手術患者
  - 疼痛・身体機能とも硬膜外投与の有無で有意差なし

Pediatric Readmissions After Hospitalizations for Lower Respiratory Infections

Pediatrics. 2017 Aug;140(2):e20160938.

背景

- 下気道感染
  - 小児科入院の最も一般的な理由の一つ
  - 多くの再入院につながっている
- 再入院率
  - 20－30%は避けられる可能性

研究目的

- 診療報酬請求（Medicaid）データ
- 2007年12月1日～2009年11月30日

下気道感染で入院した18歳以下の患者

30日以内の予定外の入院に影響する要因
再入院が病院によって異なるかどうか

統計モデル

- 病院－患者
- ランダム切片モデル
  - 再入院 $= \alpha + \beta \times$ 患者背景 $+ \varepsilon$
  - $\alpha = \gamma + u$
    - γ：切片の全体平均
    - u：病院ごとの誤差

統計モデル

患者背景の係数
- 再入院に与える影響の大きさ

各病院の切片
- 各病院の再入院率の大きさ

切片の分散
- 再入院率のばらつき

結果

再入院のリスク因子
- 年齢・性別・慢性疾患

各病院の調整後再入院確率
- 5.2%(3.8 - 8.8%)
  - 1SD上の病院→6.5%
  - 1SD下の病院→4.2%

まとめ

- 階層構造を考慮しない分析を行うと誤った結果を導く
- マルチレベル分析は3つの目的に利用する
  - ➢階層構造を考慮する
  - ➢グループの特性が個人に与える影響を調べる
  - ➢グループ間のばらつきを調べる
- 切片・傾きがグループごとに異なるモデル

NDB・DPCデータベース研究人材育成Webinar

自己対照ケースシリーズ (Self-Controlled Case Series: SCCS)

東京大学大学院医学系研究科臨床疫学・経済学
橋本 洋平

本日の内容

1. イントロダクション
2. SCCS研究とは
3. 古典的な研究デザインとの比較
4. SCCSの必要条件
5. よく受ける質問

コホート研究/ケースコントロール研究

- ランダム化比較試験 (RCT) は倫理的・費用的制約が大きい
- そこで・コホート研究やケースコントロール研究が行われる・
- しかし・未測定交絡因子の問題が残る・
- たとえ交絡因子が全て測定できたとしても・
  因果パス (causal pathway) を完璧に描くのは不可能である・
- 結果・残差交絡が生じる

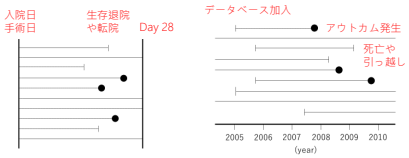
自己対照研究デザイン (SCCS: self-controlled study design)

- Event (アウトカム) が起きた人のみを対象とする・
- 同一人の過去および未来の期間と比較する・
- 同一人内の比較のため・時間で変化しない因子
  (性別・遺伝子・生活習慣・社会経済的背景) は自然に相殺される・
- 因果の追求のため (だけ) に用いる・

	Exposure-indexed	Outcome-indexed
個人内比較	SCCS	Case-crossover
個人間比較	コホート	ケースコントロール

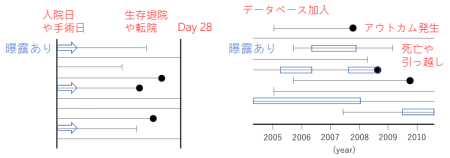
入院データ vs. 外来 (一般住民) データ

	入院データベース (例：DPC, MDV)	**外来 (一般住民) データベース (例：JMDC, NDB)**
(i) コホートの特性	Closed	Open
(ii) 曝露因子の特性	(多くが) time-invariant	(多くが) time-variant
(iii) アウトカムの特性	(多くが) 死亡	様々なアウトカム



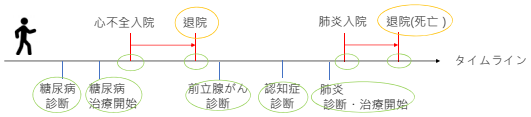
入院データ vs. 外来 (一般住民) データ

	入院データベース (例：DPC, MDV)	外来 (一般住民) データベース (例：JMDC, NDB)
(i) コホートの特性	Closed	Open
(ii) 曝露因子の特性	(多くが) time-invariant	(多くが) time-variant
(iii) アウトカムの特性	(多くが) 死亡	様々なアウトカム



入院データ vs. 外来 (一般住民) データ

	入院データベース (例：DPC, MDV)	外来 (一般住民) データベース (例：JMDC, NDB)
(i) コホートの特性	Closed	Open
(ii) 曝露因子の特性	(多くが) time-invariant	(多くが) time-variant
(iii) アウトカムの特性	(多くが) 死亡	様々なアウトカム



記述的 vs. 分析的

記述的な研究

実態を把握したい
- 疾病の負担
- 地域格差
- 時代による変化

分析的な研究

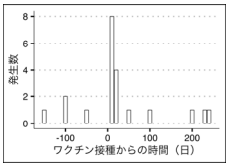
因果を追求したい

本日の内容

1. イントロダクション
2. SCCS研究とは
3. 古典的な研究デザインとの比較
4. SCCSの必要条件
5. よく受ける質問

なぜSCCSが生まれたか？

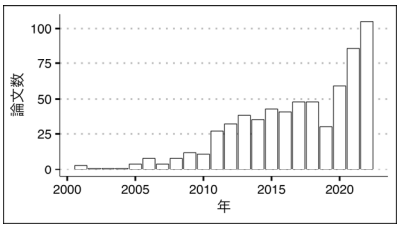
- 1988年・イギリスでは・
  浦部株を含むMMR（麻疹・おたふくかぜ・風疹）ワクチンが導入された・
- もともとこのワクチンは無菌性髄膜炎を起こす可能性が指摘されていた・
- 無菌性髄膜炎になった児を調べると・
  明らかにワクチン接種「後」に発症が集積していた・



1) Farrington P et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.

- 時間的関係から因果関係はありそう・
- コホート研究は・イベント発症頻度が低すぎるため実行困難・
- ケースコントロール研究は・コントロールを選ぶ源集団の設定が困難・
- なんとかして・ケース（無菌性髄膜炎患者）のみから・リスク比やハザード比のような統計量を出すことができないか？
- 開発された手法がSCCS・

最近の論文数の推移

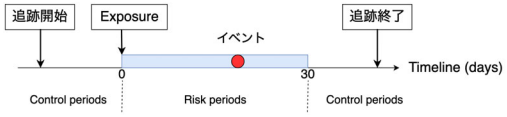


self-controlled case series [Title/Abstract]で検索（2022/11/17, PubMed）

SCCSに必要な情報

- ケースのみのデータで良い
  1. 追跡開始日
  2. 追跡終了日
  3. Exposureを受けた日
  4. Event発生日
- Timelineはageで考えるのが一般的（日単位）.
- 観察開始日の年齢が20歳なら，365.25×20=7305（日）
- イベント発生日が20歳2ヶ月なら，7305+30.5×2= 7366（日）

SCCSで求められるもの

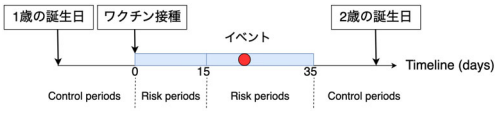


- Control periodsにおける発生率を基準とした，risk periodsにおける発生率の比が求められる（発生率比，incidence rate ratio [IRR]）．
- 条件付きポアソン回帰が用いられる．

SCCSのメリット・デメリット

- 同一人内の比較のため，時間で変化しない因子（性別，遺伝子，生活習慣，社会経済的背景）は自然に相殺される
- ケースのみを用いた手法なので，コントロールを選んでくる必要がない
  → コントロール選択に伴う選択バイアスが発生しない
- 絶対的指標（発生率）を求めることはできない．
  → コホート研究を同時に行い，発生率を提示すると説得力が増す．

MMRワクチンと熱性けいれん・無菌性髄膜炎



Control periodと比べたrisk periodの発生率比 (IRR)は，
6–11日後　3.04 (2.27-4.07)
15–35日後 1.51 (1.21-1.90)

Farrington P, et al. *Lancet*. 1995.

本日の内容

1. イントロダクション
2. SCCS研究とは
3. 古典的な研究デザインとの比較
4. SCCSの必要条件
5. よく受ける質問

SCCSとコホート研究の比較 ①

–MMRワクチンと熱性けいれん– [1)]

	サンプルサイズ	Risk period	Relative incidence
コホート研究[2)]	679942人 （487人event発生）	8-14 days	**2.83 (1.44–5.55)**
SCCS[3)]	952人 (全員event発生)	6-11 days	**3.04 (2.27–4.07)**

- 点推定値は両者でほぼ同じ．
- SCCSの方がイベント数が多いため信頼区間が狭い．
- SCCSの方が全体のサンプル数は小さく，研究遂行が容易．

1) Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.
2) Barlow WE, et al. *N Engl J Med*. 2001.
3) Farrington P, et al. *Lancet*. 1995.

SCCSとコホート研究の比較 ②

–MMRワクチンと自閉症– [1]

	サンプルサイズ	Risk period	Relative incidence
コホート[2]	537303人 (316件event発生)	Indefinite	**0.92 (0.68–1.24)**
SCCS[3]	357人 (全員event発生)	Indefinite	**0.88 (0.40–1.95)**

- 点推定値は両者でほぼ同じだが・SCCSの方が信頼区間が広い・
- Risk periodがindefiniteの時は・SCCSは検出力低くなる・
- SCCSの方が全体のサンプルサイズが小さく研究遂行が容易・

1) Farrington P et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.
2) Madsen KM, et al. *N Engl J Med*. 2002.
3) Farrington P, et al. *Journal of the Royal Statistical Society: Series C*. 2006.

SCCSとコホート研究の比較 ③

–クロピドグレル内服中患者における・プロトンポンプ阻害薬 (PPI) と心筋梗塞– [1]

	サンプルサイズ	Risk period	Relative incidence
コホート	29025人 (734件event発生)	PPI内服期間	**1.30 (1.12–1.50)**
SCCS	774件event発生	PPI内服期間	**0.75 (0.55–1.01)**

- 点推定値は両者で異なる方向・
- コホート研究では残差交絡が起きていた可能性がある・

1) Douglas IJ, et al. *BMJ*. 2012.

SCCSとコホート研究の比較 ④

–コロナワクチン接種と眼有害事象 (ぶどう膜炎・視神経炎など) – [1]

	サンプルサイズ	Risk period	Relative incidence
コホート	99718人 (107件event発生)	接種後3ヶ月	**1.8 (1.2–2.9)**
SCCS	115件event発生	接種後3ヶ月	**0.9 (0.7–1.1)**

- 複数回のexposure (ワクチン1回目と2回目) を区別した解析
- 点推定値は両者で異なる方向・
- コホート研究では残差交絡が起きていた可能性がある・
  (肥満・血圧値の情報が不明)

1) Hashimoto Y, et al. *Ophthalmology*. 2022.

本日の内容

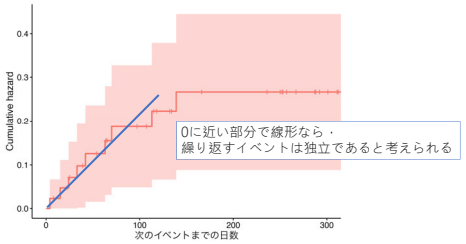
1. イントロダクション
2. SCCS研究とは
3. 古典的な研究デザインとの比較
4. SCCSの必要条件
5. よく受ける質問

条件①： 反復するeventの場合は・互いに独立している
反復しないeventの場合は・その確率がレアである

- 心筋梗塞は1回発生すると・2回目以降も起きやすいはず・
- 初回の心筋梗塞のみを解析に含めた解析を行う[1]・
- 2回目以降の心筋梗塞も含めた解析とほぼ同じ結果なら・
  robustと言える・
- 心筋梗塞の発生頻度はレアであるという仮定をおいている・

1) Minassian C, et al. Annals of Internal Medicine. 2010.

Cumulative hazardでイベントが独立であることを確認



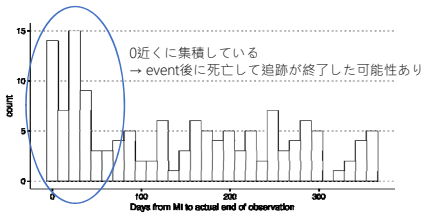
Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.

条件②：Event発生は後の追跡期間に影響を及ぼさない

- 短期死亡率が高いeventで問題となる．
- eventが心筋梗塞の場合，その後早期に死亡してしまう確率が高く，後の追跡期間に影響を及ぼす（極端な例は，event = 死亡）．

25

Eventから追跡終了までの日数のヒストグラム



Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018. 26

Terminal risk periodの導入

- Ⅰ. baseとなるSCCSモデル
- Ⅱ. 追跡終了～30日前をterminal risk periodとして入れたモデル
- モデルⅠ・Ⅱを尤度比検定し有意となれば，eventが追跡期間に影響を及ぼしている可能性が高い．
- Eventが後の追跡期間に影響を及ぼしていると考えられる際は，RのSCCS package内の eventdepenobs()関数で対応可能．

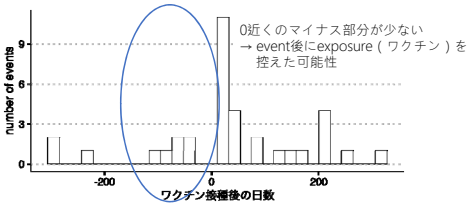
Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018. 27

条件③：Event発生は後のexposureに影響を及ぼさない

- 薬剤の副作用を検討する場合が代表的．
- ワクチンとけいれんの関連を見たい場合，けいれんを起こした患者へのワクチン投与は控えられるため，後のexposureに影響を及ぼす．

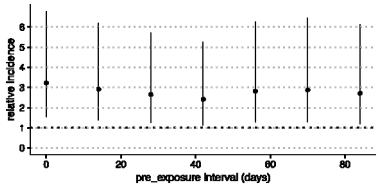
28

Exposureからeventまでの日数のヒストグラム



Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018. 29

Pre-exposure periodをモデルへ組みこむ



Pre-exposure periodを動かしても点推定値は大きく変化しない→ robust
Robustではない場合は，RのSCCS package内のeventdepenexp()関数で対応可能．

Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018. 30

条件④：Exposureは後のevent捕捉に影響を及ぼさない

- 薬剤有害事象の自己申告型調査が代表例．
- 薬剤A内服後の下痢をeventとする場合．はじめから薬剤Aと下痢の関連が疑われたケース（内服後早期の下痢）のみが含まれる．SCCSの結果・有意に関連がでるのは当然である．
- ケースコントロール研究と同様に．源集団を明確に意識する．（薬剤Aとの時間的関係によらず下痢患者を集める）

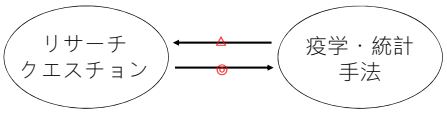
- 4つの条件全てを満たすのは困難である．
- コホート研究では未測定交絡因子がないことを前提にしているが．それも現実的には無理な話である．
- コホート研究・ケースコントロール研究といった古典的な手法に加え．SCCS結果も提示するのが好ましい．
- 両者は異なるassumptionの上に成立している
  → 結果が一致すれば．robustと言える
  → 結果が一致しなければ．より深く結果を考察することが可能となる

本日の内容

1. イントロダクション
2. SCCS研究とは
3. 古典的な研究デザインとの比較
4. SCCSの必要条件
5. よく受ける質問

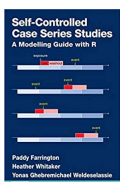
Q1: SCCSを使ってみたい．どんなリサーチクエスチョンが良いか？

- 「SCCSを使う」ことを目的にするのは．おすすめできません．
- まずはリサーチクエスチョンを設定し．曝露とアウトカムの特性がSCCSを使用するための条件を満たしている場合．是非トライしてみてください．

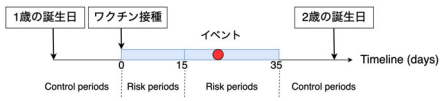


Q2: 自己対照研究デザインの統計コマンドを教えてほしい．

- 下記ウェブサイトにR・STATA・SASの練習用データセットとコマンド例（データセット作成から条件付きポアソン回帰分析まで）が載っています．
  https://sccs-studies.info/about-us.html
- Rユーザーは右の本がおすすめです．
  (Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.)



Q3: サンプルサイズはどれくらい必要か？




```
samplesize(  eexpo=2.5,    # 発生率比
             risk=21,      # Risk period (ワクチン接種後15-35日)
             astart=366,   # 観察開始日
             aend=730,     # 観察終了日
             p=1)          # 集団の中でexposureを受けた割合
→ 110例のイベント数が必要
```

Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.

Q4: 解析に用いる症例は・eventを生じ・かつexposureを受けた人の情報のみでよいか（exposureを受けていない人は除くか）？

- 時間依存性交絡因子を調整しない場合には・exposureを受けた人のみに絞っても良い・
- 時間依存性交絡因子を調整する場合は・exposureを受けた人も受けなかった人も全員解析対象とすべき（効率が上がる）[1]・
- 計画的にexposureを受けた人のみに絞った例もある[2]

1) Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.
2) Kwong JC, et al. *New England Journal of Medicine*. 2018.

37

サンプルサイズ計算（Exposureの割合を50%にしてみる）

```
samplesize(  eexpo=2.5,    # 発生率比
             risk=21,      # Risk period (ワクチン接種後15-35日)
             astart=366,   # 観察開始日
             aend=730,     # 観察終了日
             p= 0.5)       # 集団の中でexposureを受けた割合
```

→ **210例**のイベント数が必要

- p = 1.0→ 必要なサンプルサイズは110例だった・
- P = 0.5→ 必要なサンプルサイズは210例（約2倍）に増加・
- Exposureなしの人の情報量は乏しい
  （時間依存性交絡因子を考慮しない場合）

本日のまとめ

- SCCSは・eventが起きた人のみを対象に・その人の過去および未来の期間と比較したりする手法である・
- コホート研究よりも効率が良い・
- 時間非依存性交絡因子は自動的にキャンセルアウトされる・
- 発生率比を求めることで・因果推論に絶大な効果を発揮する・
- 絶対リスクを求めることはできない・
- 4つのassumptionの上に成立していることを意識する・
- コホート/ケースコントロール研究と併行して行うことで説得力が増す・

39

NDB・DPCデータベース研究人材育成Webinar

機械学習

東京大学大学院医学系研究科糖尿病・生活習慣予防講座
岡田 啓

本日の内容の対象者

- 機械学習を学びたいと思っているが
  ハードルが高く、導入の本すら読めずにいる人
- 機械学習に興味があるけれども
  どう論文作成に活かせばよいかわからない人

本日の目標

①機械学習で用いられる方法・用語を説明出来る
Lasso回帰、決定木、ランダムフォレスト、 XGBoost
用語：交差検証法、正則化、過学習

②予測と推論で、機械学習を使い分け、論文を読み書き出来るようになる

本日の内容

1. 機械学習の導入（Lasso回帰を中心に）
2. 機械学習を用いた論文の紹介と使われ方
3. ほかの機械学習（決定木 → ランダムフォレスト/XGBoost）

はじめに

基本的な用語説明

機械学習とは

- 大量のデータから法則性を見つけ出す一連の流れのこと
- 説明変数をX・応答変数をY・ランダム誤差をεとする
- Xと Yの間に何らかの関係があると仮定する

$$Y = f(X) + \varepsilon$$

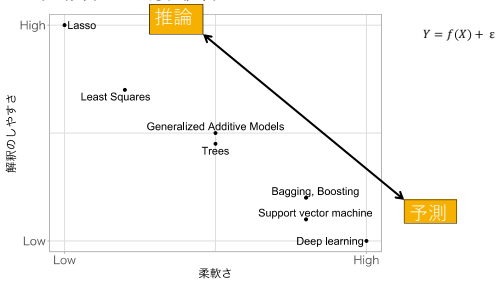
- fを推定するのが機械学習

予測 (prediction) と推論 (inference)

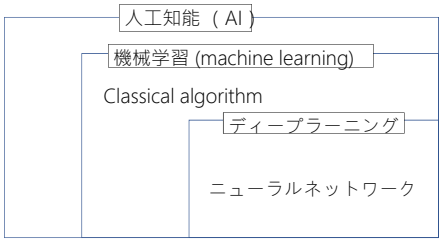
$$Y = f(X) + \varepsilon$$

- 予測
  - Xから Yを予測することが最重要・
    fの形はブラックボックスで良い・
- 推論
  - Xと Yの関係を知ることが最重要・
    X が変化すると Yはどのように変化するのか知りたい・
    fの形はブラックボックスというわけにはいかない.

柔軟性と解釈性のトレードオフ



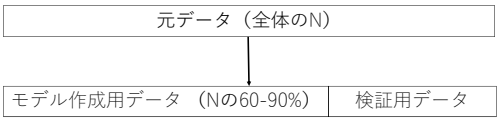
用語の説明
AI・機械学習・ディープラーニング



Introduction

多変量回帰からみた、Lasso回帰について

用語の確認



一般的に予測モデルでは、モデル作成用データと検証用データが違うデータセットの方が良いが、難しい場合、元データをモデル作成用と検証用に分けるのが一般的である。

身近な機械学習①
多変量回帰　～ 一般線形モデル ～

①最小二乗法

$\{\begin{pmatrix}x_1\\y_1\end{pmatrix}, \begin{pmatrix}x_2\\y_2\end{pmatrix}, \cdots \begin{pmatrix}x_n\\y_n\end{pmatrix}\}$というデータがあるとして

$$Predicted\ y_i = F(x) = ax_i + b$$

とすると

実測値とのずれの和

$$J = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

が最小となるaとbを求める

身近な機械学習②
多変量回帰　～ 一般化線形モデル ～

②link関数を用いたモデルの作成例：ロジスティック回帰
　リンク関数はlogitなので、

$$logit(y) = \frac{\ln(y)}{1 - \ln(y)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

実測値

$$\{\begin{pmatrix} x_{11} \\ \cdots \\ x_{1n} \end{pmatrix}, \begin{pmatrix} x_{21} \\ \cdots \\ x_{2n} \end{pmatrix} \cdots \begin{pmatrix} x_{N1} \\ \cdots \\ x_{Nn} \end{pmatrix}\}$$

①②を組み合わせて

$$J = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{n} x_{ij} \beta_j \right)^2$$

が最小となるように計算する

予測($\hat{y}$)を考える

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

これは計算出来そうだし、予測出来そう
では1000個のパラメータを使う場合は？

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{100} x_{100} + \beta_{101} x_{101} + \cdots + \beta_{1000} x_{1000}$$

つまり、yを1000個のパラメータで予測する

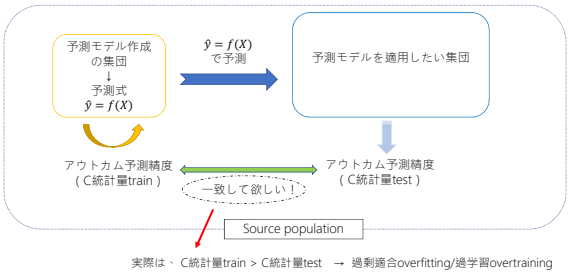
臨床疫学研究としての問題は？
①予測能は高いが、推論が難しい
②統計学的な問題も・・・

多変量回帰での主な問題

• 過剰適合overfitting/過学習overtraining/optimism
提供データ（訓練データ）に対して学習されているが
未知データ（テストデータ）に対しては適合できていない状態
　　← 過多の回帰変数、過多の自由度など

（・複数の多重共線性multiple collinearity）

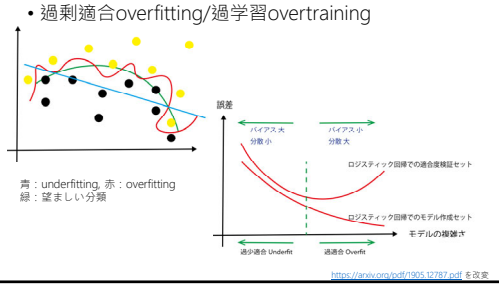
予測モデルのoptimism



実際は、C統計量train > C統計量test　→ 過剰適合overfitting/過学習overtraining

過適合の概念

• 過剰適合overfitting/過学習overtraining



https://arxiv.org/pdf/1905.12787.pdf を改変

過適合への対応

対策：
①交差検証Cross-validation
　訓練データ内でもサンプリングとテストセットを作り、それを何度も繰り返して適合度を確認する。例：k-fold法など（次スライド）

→ 外的妥当性generalizabilityを高める

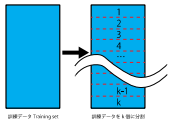
②正則化Regularization（→ 縮小推定shrinkage）
　誤差関数に正則化項を追加して、モデルの複雑度・自由度を抑制し、変数を減らす。

→ わかりやすさを重視する（より「推論」に寄せる）

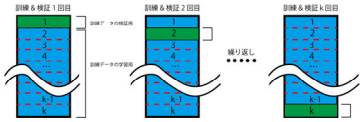
K-fold法

方法

①まず、k個に等分割する



②そのうち、1つを検証用として、残りのk-1個を訓練データとして予測式を作成する。



③k回のバリデーションの誤差を平均し、予測式の性能を評価する。

用語の確認（再掲）

元データ（全体のN）

モデル作成用データ （Nの60-90%） | 検証用データ

Trainingと validation

クロスバリデーションやk-fold法はこの部分の話

罰則項の追加による正則化

回帰で最小にしたい値

$$J = \sum_{i=1}^{N}\left(y_i - \sum^{n}_{j=1} x_{ij}\,\beta_j\right)^2$$

に罰則項

$$\lambda\sum_{j=1}^{n}|\beta_j|$$

絶対値のみ：Lasso回帰
二乗のみ：Ridge回帰
その両方：Elastic net

を加えた

$$\Sigma_{i=1}^{N}\left(y_i - \Sigma_{j=1}^{n} x_{ij}\,\beta_j\right)^2 + \lambda\Sigma_{j=1}^{n}|\beta_j|$$

を最小にするようにβを推定する。λは値を振ってみて適切な値を決める。

Lasso回帰で行われていること

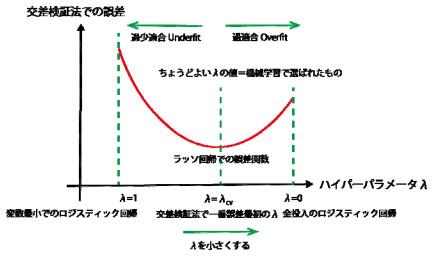
$$\Sigma_{i=1}^{N}\left(y_i - \Sigma_{j=1}^{n} x_{ij}\,\beta_j\right)^2 + \lambda\Sigma_{j=1}^{n}|\beta_j|$$

①ハイパーパラメータλの値を100個くらい振る

②k-fold法での誤差関数を値を計算

③誤差関数の最も小さいところのλを採用
（交差検証法で得られたλ）

Cf. Hyperparameterとは、自動計算で出されたparameterと違い、自分で設定するパラメータなのでhyper- + parameter



λが大きい → 罰則項の影響大 → 変数少ない
λが小さい → 罰則項の影響小 → 変数多い

小括

1. 推論/予測
2. Lasso回帰のキーワード：過学習、K-fold法、罰則項

Lasso回帰論文例：シンプルなモデル

Diabetes Care 1

A Machine Learning–Based Predictive Model to Identify Patients Who Failed to Attend a Follow-up Visit for Diabetes Care After Recommendations From a National Screening Program

Akira Okada,[1] Yohei Hashimoto,[2,3] Tadahiro Goto,[2,4] Satoko Yamaguchi,[1] Sachiko Ono,[5] Kayo Ikeda Kurakawa,[1] Masaomi Nangaku,[6] Toshimasa Yamauchi,[7] Hideo Yasunaga,[2] and Takashi Kadowaki[1,7,8]

https://doi.org/10.2337/dc21-1841

背景

①糖尿病スクリーニング後の受診率は低い

②受診中断の因子は数々の研究があるが、
受診勧奨後未受診の因子などを検討した論文は殆ど無い

既存研究では、重要と思われる13因子で回帰

→ BMI、HbA1c値、蛋白尿、降圧剤処方、脂質異常症薬処方、抗うつ剤処方などが、未受診に関連していた。

③機械学習などの新たな手法の出現の変化もあり、より「良く」「効率的に」糖尿病受診勧奨後未受診を予測すれば政策立案に役立つ可能性がある

④目的：機械学習を用いて、受診勧奨後の未受診予測モデルの構築を試みた

方法

「既存ロジスティックモデル（Diabetes Res Clin Pract 2014;105:176–184）」

VS

「既知の糖尿病関連マーカーなどを含めたものからlasso回帰にて予測因子を選択したモデル」

モデル作成・評価方法：

①trainデータ (80%)とtestデータ(20%)に分割

②交差検証法により、trainデータで予測モデル作成（lasso回帰の1SE ruleを適応）

③testデータでを用いてc統計量（ROC曲線のグラフ下面積: AUC）で予測能の評価（Delongテストで検定）を行う

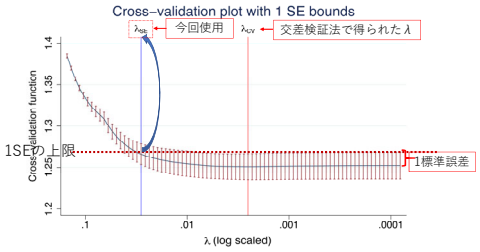
PECOによる概要

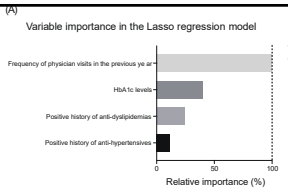
P: 過去12ヶ月に糖尿病関連レセプトがなく、健診でHbA1c ≧6.5% and 空腹時血糖値 ≧126 mg/dLを満たした患者

E/C: 受診勧奨後受診に関わりそうな変数からlasso回帰で選択された変数の曝露有無

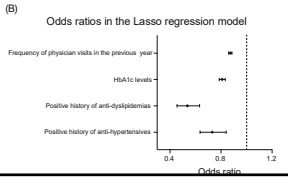
O: 健診後6ヶ月以内の糖尿病関連レセプトの発生

ハイパーパラメータλの設定



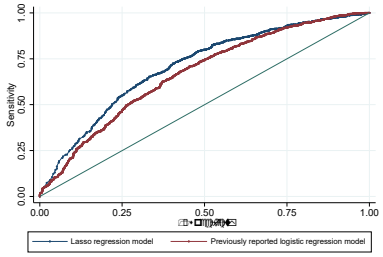


変数の重要度



係数

C統計量の比較



	criteria	C-statistics	95% CI	P value
Lasso model	10820.17	0.71	0.69 - 0.73	<0.001
Previous model	11105.58	0.67	0.65 - 0.69	

結論

既存モデル（13因子）のモデルよりも、機械学習で選ばれた4因子のモデルの方が受診勧奨後未受診をより正確に予測

機械学習を用いたメリット

よりわかりやすい（少ない変数で）モデルを作成するために、Lasso回帰を用いた。特に、変数を減らすために、1SE ruleを用いた。推論に特に有用である。

論文例②：スコア作成＋ほかの機械学習

Surgery 171 (2022) 1036–1042

Contents lists available at ScienceDirect

Surgery

journal homepage: www.elsevier.com/locate/surg

New machine learning scoring system for predicting postoperative mortality in gastroduodenal ulcer perforation: A study using a Japanese nationwide inpatient database

Takaaki Konishi, MD[a,b,*], Tadahiro Goto, MD, MPH, PhD[b,c], Michimasa Fujiogi, MD[b,d], Nobuaki Michihata, MD, MPH, PhD[e], Ryosuke Kumazawa, MPharm[b], Hiroki Matsui, MPH[b], Kiyohide Fushimi, MD, PhD[f], Masahiko Tanabe, MD, PhD[a], Yasuyuki Seto, MD, PhD[a,g], Hideo Yasunaga, MD, PhD[b]

背景と目的

- 胃十二指腸潰瘍穿孔は一般的な外科的救急疾患である
- 術後死亡予測としてASA (American Society of Anesthesiology) スコア等が知られる
- しかし、それらの予測能は不十分である

Anbalakan K et al. *Int J Surg* 2015; 14: 38–44.

本研究では、DPCデータベースを用いて下記を行うことを目的とする

① 機械学習による予測モデルの構築

② 臨床現場で使いやすい整数スコアの作成

③ ASAスコアと比較した①②の予測能の評価

34

方法＜予測モデルの構築と評価＞

① 機械学習により術後在院死亡を予測するモデルを構築

✓予測因子：入院時情報64項目

✓予測手法：XGBoost、Lasso回帰

② Lasso回帰で同定された因子を用いて整数スコアを作成

Training set (70%)

XGBoost

✓ 決定木モデルの発展形

✓「勾配ブースティング」

✓ 連続変数も投入できる精度の高い機械学習

Lasso回帰

✓ ロジスティック回帰の発展形

✓ 重要な因子のオッズ比を呈示

✓ 二値変数とカテゴリー変数を投入可能

35

方法＜予測モデルの構築と評価＞

① 機械学習により術後在院死亡を予測するモデルを構築

✓予測因子：入院時情報64項目

✓予測手法：XGBoost、Lasso回帰

② Lasso回帰で同定された因子を用いて整数スコアを作成

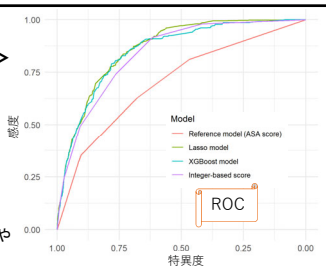
③ 機械学習モデルと整数スコアの予測能を評価

- ✓ 既存のASAスコアと比較
- ✓ Receiver Operating Curveの曲線化面積（AUC）で比較

Training set (70%) → Test set (30%)

36

結果＜予測能＞



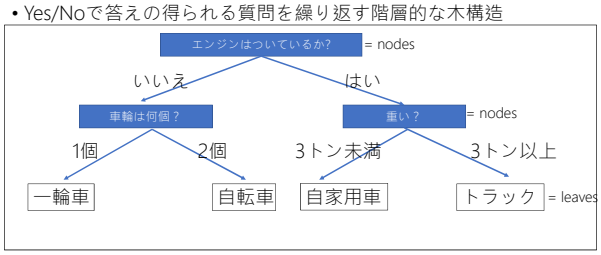
ASAスコアより機械学習モデルや整数スコアの方が予測能が高い

		C統計量 (95%信頼区間)	正確度 (95%信頼区間)	感度	特異度	陽性的中率	陰性的中率	P値
既存	ASAスコア	0.70 (0.67-0.73)	0.68 (0.67-0.69)	0.63	0.68	0.09	0.97	基準
今回作成	Lasso回帰	0.86 (0.85-0.88)	0.76 (0.75-0.77)	0.80	0.76	0.14	0.99	<0.001
	XGBoost	0.85 (0.83-0.87)	0.76 (0.75-0.76)	0.81	0.75	0.13	0.99	<0.001
	整数スコア	0.84 (0.82-0.86)	0.63 (0.62-0.64)	0.92	0.62	0.10	0.99	<0.001

決定木とその発展版
XGBoost、ランダムフォレスト

決定木 (decision tree)

- Yes/Noで答えの得られる質問を繰り返す階層的な木構造



決定木の特徴

メリット

分類・回帰とも用いることができる

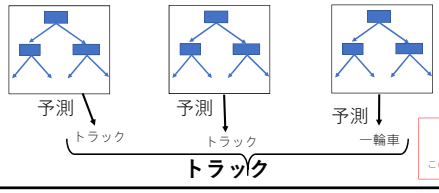
- 説明変数はカテゴリカル変数・連続変数とも対応可
- 得られた結果の意味を解釈しやすい

デメリット

外れ値やノイズに弱い (= overfittingしやすい)

ランダムフォレスト

- 決定木を複数作り・多数決を取ることでoverfittingを防ぐ
  1. サンプル数nのデータから・n個のサンプルを復元抽出する
  2. このデータセットを用いて決定木を作成する
     （いくつかの説明変数を選択する = 捨てる説明変数が発生する）
  3. k回繰り返す
  4. k個の決定木から・多数決によりクラスを予測する

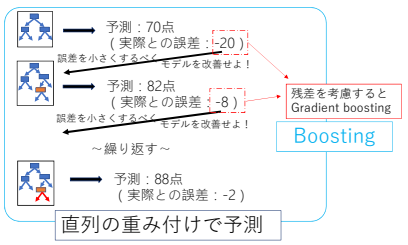


多数決により、決定木の弱点のoverfittingを防ぐこの多数決のことを「バギング」と呼ぶ

XGBoostなどのBoosting

- 最初の決定木から誤差を求め、より誤差の少ない（精度のよい）決定木に、徐々に洗練させていき、重みづけて予測する方法

例：ある生徒の英語のテスト成績
前回の点数：80点
性別：女性
数学の点数：90点
国語の点数：70点
本当の英語の点数：90点



XGBoost: eXtreme Gradient Boosting

本日のまとめ（目標の再掲）

①機械学習で用いられる方法・用語を説明出来る

Lasso回帰、決定木、ランダムフォレスト、 XGBoost

用語：交差検証法、正則化、過学習

②予測と推論で、機械学習を使い分け、論文を読み書き出来るようになる

NDB・DPCデータベース研究人材育成Webinar

データベース研究における バリデーション研究

自治医科大学データサイエンスセンター
山名隼人

Agenda

- バリデーション研究とは
  なぜバリデーション研究が必要か
- バリデーション研究の活用例
- バリデーション研究の例と実際
- 現状と課題

大規模データベース研究の特徴

保健医療介護ビッグデータ
- NDB、DPC、介護レセプトデータ
- JMDC、MDV

など

利点
- データの量、N数
- 全国規模、集団代表性
- データ取得が（比較的）容易
- 実際の臨床現場を反映（リアルワールド）

大規模データベース研究の特徴

研究の特徴
- 後ろ向き研究 ↔ 前向きにデータ取得
- 観察研究 ↔ RCTなどの介入研究

→ 研究デザイン・統計手法を駆使、
観察研究から重要なエビデンスを生み出す

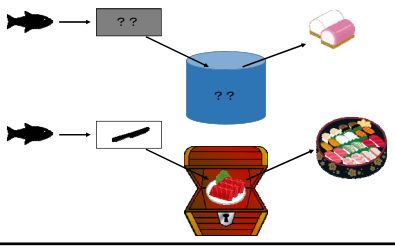
- 既存の情報を二次利用

→ 情報の「質」はどうか？

情報の質

情報の妥当性(validity)
- 査読者コメントで非常に多いもの：“Is the data validated?”
- データベースに格納されている情報と、研究者がイメージする情報の差
- 例：傷病名
  - DPCデータの主病名・併存症名、“レセプト病名”
  - 緊急入院の場合の併存症名
- 100%の正確性は現実的ではないが、定量化することでデータを活用する研究の質を向上させる

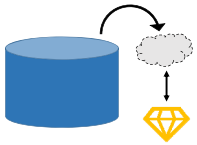
情報の質



バリデーション研究

データベースに記録された情報の妥当性を検証する

- データベースのうちの一部について、より正確な別の情報と比較する（至適基準、reference standard）
- データベース全体の妥当性を担保する



妥当性の指標

感度・特異度

至適基準（reference standard）との比較が必要

例：「脳梗塞」の病名

- 至適基準＝カルテ上の病名
- 検証したいもの＝データベース上の病名

の場合、

- **感度**： カルテ上、脳梗塞がある人のうち、データベース上の病名でも脳梗塞ありとされる人の割合
- **特異度**：カルテ上、脳梗塞がない人のうち、データベース上の病名でも脳梗塞なしとされる人の割合

妥当性の指標

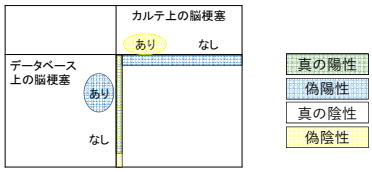
感度・特異度・陽性的中率・陰性的中率

		カルテ上の脳梗塞		
		あり	なし	
データベース上の脳梗塞	あり	真の陽性 a	偽陽性 b	陽性的中率 =a/(a+b)
	なし	偽陰性 c	真の陰性 d	陰性的中率 =d/(c+d)
		感度 =a/(a+c)	特異度 =d/(b+d)	

- 感度90%, 特異度90%, 実際の割合5%→陽性的中率は32%

妥当性の指標

感度・特異度



- 感度90%, 特異度90%, 実際の割合5%→陽性的中率は32%

バリデーション研究の必要性

感度・特異度が研究結果に与える影響

- A薬 vs B薬でアウトカムを比較したい
- 実際：

A薬 100/2000 (5%)　　B薬 200/2000 (10%)

＜データベースにおけるアウトカム特定の感度50%・特異度80%の場合＞

A薬		真のアウトカム		
		あり	なし	
観測されるアウトカム	あり	50	380	430
	なし	50	1520	1570
		100	1900	2000

B薬		真のアウトカム		
		あり	なし	
観測されるアウトカム	あり	100	360	460
	なし	100	1440	1540
		200	1800	2000

バリデーション研究の必要性

感度・特異度が研究結果に与える影響

- A薬 vs B薬でアウトカムを比較したい
- 実際：

A薬 100/2000 (5%)　　B薬 200/2000 (10%)

＜データベースにおけるアウトカム特定の感度50%・特異度80%の場合＞

- 観測：

A薬 430/2000 (21.5%)　　B薬 460/2000 (23%)

Real-World??

バリデーション研究の活用例（1）

Angiotensin converting enzyme inhibitors and risk of lung cancer.

BMJ 2018;363:k4209

- Database　UK Clinical Practice Research Datalink (CPRD)
- Patient　高血圧で治療開始
- Exposure　アンジオテンシン変換酵素阻害薬
- Control　アンジオテンシン受容体拮抗薬
- Outcome　肺癌の発症

バリデーション研究の活用例（1）

Angiotensin converting enzyme inhibitors and risk of lung cancer.

BMJ 2018;363:k4209

- Method内の記載
  “CPRDの記録は妥当であり質が高いことが示されている[引用]。さらに、CPRDにおける肺癌の診断名はUK National Cancer Data Repositoryの診断名とも高い一致率 (>93%) を示している[引用]。”

バリデーション研究の活用例（2）

Use of haloperidol versus atypical antipsychotics and risk of in-hospital death in patients with acute myocardial infarction.

BMJ 2018;360:k1218

- Database　米国Premier Research Database
- Patient　急性心筋梗塞で入院　（＋せん妄あり）
- Exposure　ハロペリドール
- Control　非定型抗精神病薬
- Outcome　7日以内の在院死亡

バリデーション研究の活用例（2）

Use of haloperidol versus atypical antipsychotics and risk of in-hospital death in patients with acute myocardial infarction.

BMJ 2018;360:k1218

- Method内の記載
  “せん妄の病名が記録されていることは選択基準に含めなかった。せん妄があっても病名が記録されることが少なく、ICD-9コードによるせん妄の定義の感度は3%と低いことが示されているためである[引用]。代わりに、事前に精神疾患の病名がない患者において抗精神病薬が開始されたことをせん妄と定義した。”

バリデーション研究の種類

目的

- Target:　疾患・処置・薬剤・アウトカム
- 求める指標：　感度／特異度
  陽性的中率／陰性的中率

方法

- カルテレビュー
- より信頼性が高いデータとの結合
  例：院内がん登録・症例データベース
- 他の情報からの“間接証拠”
  例：先行研究の発生率

バリデーション研究の例

Validity of diagnoses, procedures, and laboratory data in Japanese administrative data.

J Epidemiol 2017;27:476-82.

概要

- Target：　DPCデータの疾患・処置
- 求める指標：　感度／特異度、陽性的中率／陰性的中率
- 方法：　カルテレビュー
  小規模（N=315）

Methods (1)

対象

- 国立病院機構の 4 病院
- 2014年度入退院、入院時18歳以上、DPC入院
- 各施設100入院をランダムに抽出

内容

- 病名： Charlson Comorbidity Indexの基準となる17疾患
  （主or入院時併存として有無）
- 処置： 10種類の処置 （入院日の実施有無）

Methods (2)

カルテレビュー方法

- 2名（看護師+医師）が独立に判断、相違があれば協議

データベースからの抽出方法

- 病名： 様式1 主傷病, 入院の契機となった傷病, 入院時併存病名×4 （疑い除く）
- 処置： EFファイル 入院日の実施有無

解析

- Validity： 感度・特異度、陽性・陰性的中率

Results (1)

病名の妥当性

疾患	カルテ上の病名出現頻度 n	%	DPCデータの感度(%)	DPCデータの特異度(%)	DPCデータの陽性的中率(%)	DPCデータの陰性的中率(%)
心筋梗塞	23	7.3	52.2	99.7	92.3	96.4
うっ血性心不全	32	10.2	68.8	97.5	75.9	96.5
末梢血管疾患	29	9.2	34.5	99.3	83.3	93.7
脳血管疾患	38	12.1	50.0	98.9	86.4	93.5
糖尿病(合併症なし)	46	14.6	52.2	96.7	72.7	92.2
糖尿病(合併症あり)	17	5.4	29.4	99.7	83.3	96.1
悪性腫瘍	97	30.8	83.5	97.7	94.2	93.0

Results (2)

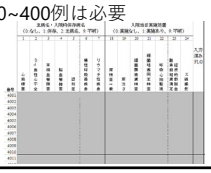
処置の妥当性

処置	カルテ上の処置実施頻度 n	%	DPCデータの感度(%)	DPCデータの特異度(%)	DPCデータの陽性的中率(%)	DPCデータの陰性的中率(%)
尿検査	74	23.5	98.6	98.3	94.8	99.6
細菌鏡見	35	11.1	91.4	100	100	98.9
細菌培養	35	11.1	97.1	100	100	99.5
呼吸心拍監視	30	12.8	66.7	92.2	55.6	95.0
X線撮影	161	51.1	97.5	99.4	99.4	97.5
CT	93	29.5	100	99.5	98.9	100
尿路カテーテル挿入	29	9.2	65.5	97.2	70.4	96.5

バリデーション研究の実際 (1)

事前準備

- 症例数
  10%の有病率、感度50%の場合：
  200例では真の陽性は10人のみ。300~400例は必要
- カルテレビューにかかる時間・費用
  10分×100例＝16.7時間
- 病院への計画説明＆協力依頼
- レビュー用シート作成



バリデーション研究の実際 (2)

カルテレビューの実施

- 日程
  - 連続する 2 日間で実施
  - 共同研究者・施設側との日程調整
- その他
  - 電子カルテの使用権限、環境の確保
  - 慣れない電子カルテの利用
  - 限られた時間　再訪問は困難
  - レビュー結果データのやりとり

バリデーション研究の実際（3）

トラブル

- 予定していた症例全ては終わらなかった
  Reviewerに“why?”と指摘され、論文に“due to time constraints”と追加
- ある病院では、鼻腔培養MRSAがほとんどの患者で実施されていた
- 別の病院では、過去の温度板が参照できなかった
  電子カルテシステム更新の影響

バリデーション研究の課題（1）

バリデーション研究は実施が難しい

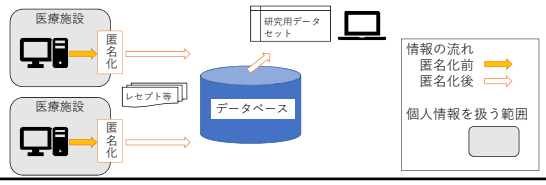
van Walraven et al. Administrative database research infrequently used validated diagnostic or procedural codes. *J Clin Epidemiol* 2011より：

- Measuring code accuracy can be time consuming and expensive.
- Securing grant support for such a study is difficult because a code-validation study may hold little appeal for many granting agencies.
- Finally, journals may decide to remove data regarding code validation in an effort to save space.

バリデーション研究の課題（2）

データ収集と匿名化の壁

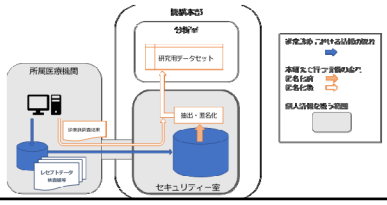
- 各施設から匿名化データを収集して活用する場合、カルテに戻ってのバリデーションは難しい



バリデーション研究の課題（2）

データ収集と匿名化の壁

- 国立病院機構におけるバリデーション研究の場合



バリデーション研究の現状

日本のバリデーション研究は増加傾向

- 報告数
  2017年11月～2022年3月：20編超の英語論文が発表
  2018: 1 → 2019: 2 → 2020: 1 → 2021: 13
- 領域　　：循環器、がん、糖尿病が多い
- 規模　　：単施設・小規模が多い
- 至適基準：カルテ調査が多いが、レジストリ等も活用

バリデーション研究の現状

例

筆頭著者、掲載誌、年	データ	対象	施設数	至適基準
Imai et al, J Med Virol 2019	医科 レセプト	B型肝炎ウイルスの既感染	4	カルテ調査
Shigemi et al, Cancer Epidemiol 2021	DPC データ	がん （15種類）	31	がん登録
Ono et al, BMC Health Serv Res 2021	歯科 レセプト	各種診断名、歯科処置等	1	カルテ調査

バリデーション研究の現状

例

筆頭著者、掲載雑誌、年	データ	対象	施設数	至適基準
Yamana et al, Pharmacoepidemiol Drug Saf 2022	DPC データ	術後感染症	4	カルテ調査
Konishi et al, Surg Today 2022	DPC データ	手術情報 （術式等）	1	カルテ調査
Fujita et al, J Glaucoma (in press)	医科 レセプト	緑内障	1	カルテ調査

データベース研究におけるバリデーション研究：まとめ

- 保健医療介護ビッグデータの懸念：研究に必要な情報が正確に記録されていないかもしれない
- バリデーション研究：データベースのうちの一部について、より正確な別の情報（至適基準）と比較して情報の妥当性を検証する研究
- データベースを活用する研究の質を向上させることができる
- 日本においても増加傾向