

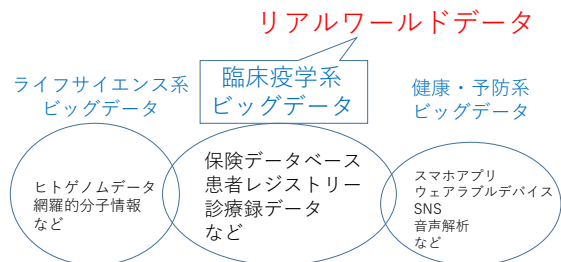
NDB・DPC等公的データベースの動向

東京大学大学院医学系研究科臨床疫学・経済学
康永秀生

1. リアルワールドデータの概要
2. 保険データベースの概要
3. NDBの直近の動向
4. 公的DB間の連携

1. リアルワールドデータの概要

様々な医療ビッグデータ



さまざまなRWD

1. 保険データベース (Administrative claims data)

保険者などから収集した診療報酬請求データを研究に二次利用。

- (1) 公的
匿名医療保険等関連情報データベース (NDB)
匿名診療等関連情報データベース (DPCDB)
要介護認定情報・介護レセプト等情報 (介護DB)、等
- (2) 民間
JMDC、DeSC、MDV、等

さまざまなRWD

2. 患者レジストリー (patient registry)

政府や学術団体などが特定の疾患・治療に関する患者データを各医療機関等から収集。

- (1) 公的レジストリ
全国がん登録データベース、匿名感染症関連情報データベース、指定難病患者データベース、小児慢性特定疾病児童等データベース、など
- (2) アカデミックレジストリー
National Clinical Database (NCD)、日本整形外科学会症例レジストリー (JOANR)、日本ICU患者データベース (JIPAD)、日本外傷データベース、等

さまざまなRWD

3 診療録データ (electronic medical records)

電子カルテ、検査データ、など

Medical Information Database Network (MID-NET)
国立病院機構データベース
徳洲会メディカルデータベース
など

2. 保険データベースの概要

保険データベースの類型

- (1) レセプトデータベース
 - 厚生労働省のNDB、介護DB
 - 各保険者のレセプトデータベース：国保データベース (KDB) など
 - 研究者が自治体などと連携して構築したレセプトデータベース
 - 民間企業が提供するレセプトデータ：JMDC、DeSC、など
- (ii) DPCデータベース
 - 厚生労働省のDPCDB
 - 研究者がDPC病院の協力を得て構築したDPCデータベース：JROAD、厚労科研DPCデータ調査研究班、など
 - 民間企業が提供するDPCデータ：MDV、など

公的医療保険の保険者

保険の種類別		保険者数	被保険者数 (万人)	被扶養者数 (万人)	給付費計 (億円)
健康保険	協会けんぽ	1	2,508	1,520	66,798
	組合管掌健康保険 (組合健保)	1,388	1,641	1,197	42,396
船員保険		1	6	6	215
共済組合	国家公務員共済	20	110	101	2,764
	地方公務員共済	64	306	257	9,032
	私学共済	1	61	34	1,623
国民健康保険	市町村国保	1,716	2,537	・	87,107
	国民健康保険組合	161	268	・	4,642
後期高齢者医療制度		47	1,843	・	157,556
合計		3,399	9,280	3,115	372,133

出典：厚生労働省、第24回医療経済実態調査（保険者調査）

本日4限「DPCデータ研究のデザインと実践」
本日5限「DeSCデータ研究のデザインと実践」
本日6限「JMDCデータ研究のデザインと実践」

レセプト：医科レセ・DPCレセ・歯科レセ・調剤レセ

レセプトのレコード：医療機関情報レコード、レセプト共通レコード、保険者レコード、傷病名レコード、診療行為レコード、医薬品レコード、特定器材レコード、など

レセプト電算マスター：傷病名マスター、診療行為マスター、医薬品マスター、特定器材マスター、など

レセプトデータに含まれる項目：i 診療開始日・診療実日数、ii 初診・再診、時間外等、iii 医学管理（医師の指導料等）、iv 疾病名、v 投薬、vi 注射、vii 処置、viii 手術、ix 検査、x 画像診断、xi 請求点数

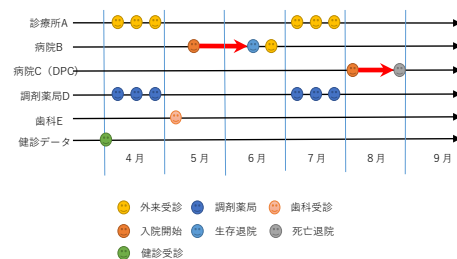
NDB

National Database of Health Insurance Claims and Health Checkup
レセプト情報・特定健診等情報データベース
匿名医療保険等関連情報データベース

匿名医療保険等関連情報データベースの利用に関するホームページ
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryoku/iryohoken/reseputo/index.html

本日2限の「NDBのデータ申請とオンサイトセンターの利用」をお聴きください。

匿名化された個人IDで紐付けが必要



社会保険診療報酬支払基金のホームページ

医科・DPC・調剤・歯科の「電子レセプト作成の手引き」
https://www.ssk.or.jp/seikyushiharai/rezept/iryokikan/iryokikan_02.html

「診療報酬情報提供サービス」のホームページ
<http://www.iryohoken.go.jp/shinryohoshu/>
各種マスターをダウンロードできる

介護DB

匿名介護情報等の提供について

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

- 提供形式：特別抽出、サンプリングデータセット、集計表情報
- NDBオープンデータと同様、介護DBオープンデータも公開されている。
- 2020年より、NDBとの連結データの第三者提供も開始された。
- 医療・介護データ等解析基盤（Healthcare Intelligence Cloud, HIC）の構築も進められている。

本日3限の「NDB研究のデザインと実践」をお聴きください。

介護DBに含まれる情報

介護レセプト情報

①施設サービス

特別養護老人ホーム（特養）、介護老人保健施設（老健）、介護療養型医療施設、介護医療院

②居宅サービス

訪問介護（ホームヘルプ）、訪問入浴介護、訪問看護、訪問リハビリテーション、通所介護（デイサービス）、通所リハビリテーション（デイケア）、居宅療養管理指導、短期入所生活介護・短期入所療養介護（ショートステイ）、特定施設入所者生活介護（有料老人ホームなどにおける介護）、福祉用具貸与、特定福祉用具販売、住宅改修

③地域密着型サービス

定期巡回・随時対応型訪問介護看護、夜間対応型訪問介護、小規模多機能型居宅介護、看護小規模多機能型居宅介護、認知症対応型通所介護、地域密着型通所介護（小規模デイサービス）、認知症対応型共同生活介護（グループホーム）、地域密着型特定施設入居者生活介護、地域密着型介護老人福祉施設入所者生活介護

要介護認定情報

申請時点の年齢区分、性別、申請区分（新規、更新など）、最終的に判定された要介護認定の区分に加えて、以下が含まれる。

要介護認定の一次判定における情報：

- (i) 「心身の状況に関する調査」における74項目の基本調査
- (ii) 主治医意見書のうち、短期記憶、認知能力、伝達能力、食事行為、認知症高齢者の日常生活自立度の項目
- (iii) 要介護認定等基準時間
- (iv) 一次判定結果

要介護認定の二次判定における情報：

認定有効期間および二次判定結果

NDB 収載データの拡充

	主なデータ	収載・提供開始（予定） <法令改正>
死亡情報	・ 死亡年月日時分 ・ 死亡の原因、種類、場所の種別	R6秋収載・順次提供 <R5.3公布・R5.4施行>
訪問看護レセプト	・ 訪問看護療養費明細書情報 ※訪問看護療養費コード、指示区分、指示期間、訪問場所・回数・職種等の情報	R6.8～収載 ・ 順次提供（検討中） <改正予定>
事業者健診	・ 事業者健康診査情報（40歳未満） ※情報項目は特定健診情報と同じ。	R7～収載・提供 <R6.12公布・施行>
医療扶助健診	・ 被保護者の健康診査情報 ※情報項目は特定健診情報と同じ。	R7～収載・提供 <R6.12公布・施行>
医療扶助保健指導	・ 被保護者の保健指導情報	R8～収載・提供 <R6.3下旬公布・R6.4施行予定>

3. NDBの直近の動向

令和6年6月12日 第21回匿名医療情報等の提供に関する専門委員会 資料1
<https://www.mhlw.go.jp/content/12400000/001262621.pdf>

NDBにおける新たな迅速提供の形態について（案）

新たな提供形態の概要

- 希少疾患や薬剤安全性の網羅的調査等の利用ニーズに応えるため、前回の専門委員会ですした探索的解析環境（通年パネルデータセット等）に加え、サンプリングを行わないデータ（NDBβ）も7日で利用可能とすることとしてはどうか。
- NDBβについては、簡易な審査による迅速な利用開始を可能とするため、ブラックリスト方式で個人特定につながりうる項目（日、地域、機関、数量に係る情報等）をマスクして提供することで、個人特定しえないデータとする。

LIFE情報

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、栄養・摂食嚥下情報、栄養ケア計画等情報、口腔衛生管理情報、口腔機能向上サービス管理情報、興味関心チェック情報、生活機能チェック情報、個別機能訓練計画情報、リハビリテーション計画書、リハビリテーション会議録、リハビリテーションマネジメントにおけるプロセス管理票、生活行為向上リハビリテーション実施計画書、褥瘡マネジメント情報、排せつ支援情報、自立支援促進情報、薬剤変更情報、ADL維持等情報、など。

令和6年6月12日
第21回匿名医療情報等の提供に関する専門委員会 資料2
<https://www.mhlw.go.jp/content/12400000/001262622.pdf>

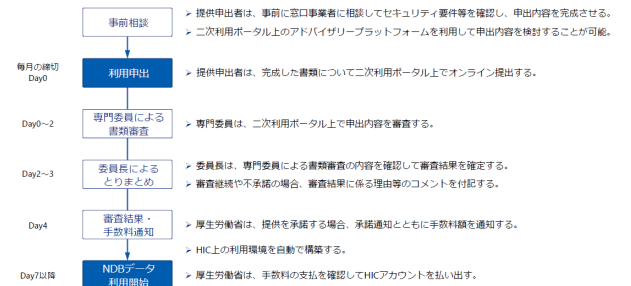
NDBに収載する死亡情報の概要

死亡情報の内容	詳細項目
死亡年月日時分	死亡年月日時分
死亡したところの種別	死亡場所（病院／診療所／自宅等）
死亡の原因と種類	原因、死亡の種類（病死／交通／溺水等）
外因死の追加事項	場所の種別（家／施設／街路等）、傷害が発生した年月、路上交通事故の有無
生後1年未満で病死した場合の追加事項	出生時体重、妊娠週数、回胎の数、母の情報（病歴、年齢）
その他	解剖の有無、関連する手術の情報（有無、年月日）、配偶者関係（既婚／未婚／死別等）

※下線で示した項目は、既に収載されている項目の審査基準に照らして、機微な項目として取り扱い、研究における必要性を特に慎重に審査する。
※赤字で示した項目は、提供の可否を慎重に審査するだけでなく、公表時には原則としてグルーピングすることを求める。

NDBデータを迅速に利用しうる審査の流れ

HIC上で迅速提供する個人特定性リスクの低いNDBデータについて、申出から実際に利用を開始しうるまで原則7日（※）を実現するため、二次利用ポータルを用いて以下のようなフローで審査を行うてはどうか。



- ※1 研究者等側が都合に要した期間は除く。
- ※2 HICでの処理が困難な場合は特別抽出して媒体提供する。
- ※3 研究のニーズに応じて取組環境。

※研究者等が都合に要した期間は除く。

NDBの利用に関するガイドラインの改正（概要）

趣旨
NDBについては、利便性・価値向上を図っていくため、二次利用ポータル及び医療・介護データ等解析基盤（HIC）を用いた提供の開始、死亡情報などの掲載情報の追加などの取り組みを順次進めており、利用に関するガイドラインの必要な見直しを行う。

- 主な改正内容
- ①二次利用ポータルを用いた申出手続開始への対応（11月末から順次対応予定）
 - ・二次利用ポータルに係る記載を追加する。
 - ・書類提出を二次利用ポータルにおいて行えるよう修正する。
 - ・二次利用ポータル上の入力や審査を容易にするため、提供申出書類における研究計画等の項目を整理する。
 - ②HICによるデータ提供の拡充への対応（11月末から順次対応予定）
 - ・通年パネルデータセット、NDB-β、リモート用全量NDBに係る記載を追加する。
 - ・承諾後の迅速なデータ提供を可能とするため、承諾後の提出を求めている書類について、提供申出時に提出するよう変更する。

- ③手数料体系変更の反映（11月1日施行予定）
 - ・政令改正による手数料体系変更を反映して、減免対象や積算方法等に係る記載を修正する。
- ④公表物の満たすべき基準の修正・追記
 - ・規制改革実施計画（令和5年6月16日閣議決定）に基づくオープンデータのマスキングルール変更を踏まえ、医療機関数3未満の場合のマスキングの記載を修正する。
 - ・死亡情報の掲載にあたって、第21回専門委員会の審議内容を踏まえ、死亡情報の公表に係る最小集計単位の原則の記載を追加する。
- ⑤成果の公表に係る注意喚起の追加
 - ・成果の公表に係る注意事項を明確化する。
 - ・NDBデータの利用場所において公表前確認の徹底を促す注意喚起用の書類の掲示を求めることを追加する。
 - ・公表前確認に係る記載を整理し、不適切利用の対象となることを強調する。
- ⑥その他
 - ・利用者からの質問の頻度が高い運用を明確化するなど、利用者の利便性向上のための見直しを行う。

28

29

30

令和6年3月14日

第176回社会保障審議会医療保険部会

資料2

<https://www.mhlw.go.jp/content/12401000/001225211.pdf>

他の医療・介護データ等との連結解析に向けて

識別子はID4、ID5

区分	DB名	主なデータ	NDBとの連結の意義・必要性	連結の検討状況等
公的	介護DB	・介護レセプト ・要介護認定情報	要介護者の治療前後の医療・介護サービスの利用状況の把握・分析に資する。	令和2年10月開始
	DPCDB	・DPCデータ （診療情報、請求情報）	急性期病院の入院患者の状態の把握が可能となり、急性期医療の診療実態の分析に資する。	令和4年4月開始
	療養福祉DB	・給付費等明確書情報 ・療養支援区分認定情報	障害者の治療前後の医療・療養福祉サービスの利用状況の把握・分析に資する。	連結解析を可能とする法案が成立、施行に向けて検討中。
	予防接種DB	・予防接種記録 ・副反応疑い報告	予防接種の有無を比較した、ワクチンの有効性・安全性に関する調査・分析に資する。	連結解析を可能とする法案が成立、施行に向けて検討中。
	感染症DB	・発生届情報	感染症の診療実態と予後の把握・分析に資する。	連結解析を可能とする法案が成立、令和6年4月施行予定。
	臨病DB	・臨床調査個人票	網羅的・経時的な診療情報を得ることが可能となり、より詳細な診療実態の把握・分析に資する。	連結解析を可能とする法案が成立、連結に向けて検討中。
	小児DB	・医療意見書		
民間	全国がん登録DB	・届出対象情報 ・死亡情報	各種がんの各ステージ分類等による診療実態と予後の把握・分析に資する。	引き続き検討中
	次世代DB	・医療機関の診療情報	医療情報と連結・分析を可能にするにより医療分類の研究開発を促進する。	連結解析を可能とする法案が成立、令和6年4月施行予定。

31

4. 公的DB間の連携

匿名感染症関連情報データベース
（感染症DB、iDB）

感染症法に基づく全数報告対象の感染症
→1類～4類感染症、および5類感染症の一部

各感染症の発生届に共通する項目：
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・感染原因など

COVID-19は当初2類相当→2023年5月8日に5類感染症

33

- ・改正感染症法の一部が2024年4月から施行
- ・匿名感染症関連情報データベース(感染症DB、iDB)が構築され、データの第三者提供が開始された。
- ・ガイドラインに沿って、第三者提供の申請を行う。
- ・提供されるデータについて、2024年4月時点ではCOVID-19の発生届情報に限られた。2020年2月3日から2023年5月7日までの報告年月日のCOVID-19患者の発生届データが対象である。
- ・iDBは、NDB、介護DB、DPCDBといった公的な保険データベースと連結したデータ申請も可能である。

匿名感染症関連情報の第三者提供の利用に関するホームページ
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryoku/kenkou/kekaku-kansenshou/idb_index.html

34

他の医療・介護データ等との連結解析に係る案件の審査体制（案）

第15回匿名医療・介護情報等の提供に関する委員会（令和6年2月9日）資料1

対応方針（案）
新たにNDB等と連結可能となる感染症DB、次世代DBの第三者提供については、法令上諮るべき審議体が異なることから、当面は、連結先のDBから提供されるデータの概要を踏まえつつ、それぞれのDBの審査委員会において第三者提供に係る審議を行うこととしてはどうか。

※NDB、DPCDB及び介護DBの連結解析に係る案件については、引き続き、匿名医療・介護情報等の提供に関する委員会において一体的に調査審議を行う。

なお、研究者の利便性を向上するための一元的な利用申請の受付・審査体制のあり方等については、医療等情報の二次利用に関するWG等において今後検討が行われる予定。
(<https://www.mhlw.go.jp/content/12301000/001206825.pdf>)

35

指定難病患者データベース
小児慢性特定疾病児童等データベース

難病＝発症の機序が明らかでなく、治療方法が確立しておらず、長期の療養を必要とする希少な疾病。

「指定難病」＝患者の置かれている状況を勘案して良質かつ適切な医療の確保を図る必要性が高い難病であり、医療費助成の対象となっている。現在、300疾患以上。
＜例＞筋萎縮性側索硬化症、パーキンソン病、重症筋無力症、多発性硬化症／視神経脊髄炎、ライソゾーム病、もやもや病、天疱瘡、スティーヴンス・ジョンソン症候群、バージャー病、全身性エリテマトーデス、ベーチェット病、特発性拡張型心筋症、再生不良性貧血、IgA腎症、特発性大腿骨頭壊死症、サルコイドーシス、特発性間質性肺炎、原発性胆汁性胆管炎、クローン病、潰瘍性大腸炎、ファロー四徴症、フェニルケトン尿症、ヒルシュスプルング病、など

36

小児慢性特定疾病

児童期に発症する疾病であって、慢性に経過し、長期にわたって生命を脅かし、長期にわたる症状・治療が生活の質を低下させ、高額な医療費の負担が続くもの。

16の対象疾患群：
悪性新生物、慢性腎疾患、慢性呼吸器疾患、慢性心疾患、内分泌疾患、膠原病、糖尿病、先天性代謝異常、血液疾患、免疫疾患、神経・筋疾患、慢性消化器疾患、染色体又は遺伝子に変化を伴う症候群、皮膚疾患、骨系統疾患、脈管系疾患

37

- ・ 難病・小慢データベースのデータ登録に当たっては、**患者からの同意**を取得している。
- ・ 2024年4月から、難病・小慢データベースの**法的根拠が明確化**。
- ・ **NDB等との連結解析**も可能に。（実際に申請可能となるのはまだ先）
- ・ 第三者提供の対象者は、それまで国・自治体、および厚労科研費や文部科研費の補助を受けている研究者に限定されていたが、2024年4からは製薬企業等の**民間企業にも提供可能**となった。
- ・ それまで研究者によるデータ活用は疾患の記述疫学データ（患者数、発病年齢、男女比、症状の分布、診療実態など）の把握にとどまっていたが、難病・小慢の患者に対する**医療・福祉分野の研究開発に資する分析**等にも活用可能となった。例えば、創業において、開発したい治療薬の対象患者の概要把握（治験の実行可能性等）、治験で使用するアウトカム指標の検討などに活用可能となった。

40

まとめ

- ・ リアルワールドデータ、特に保険データベースに関する概略をまとめた。
- ・ ここ最近、NDBに関してガイドライン改正など多くの変化が見られる。厚労省のNDBのホームページを要チェック。
- ・ 「医療等情報の二次利用に関するWG」も動いており、今後法改正の動きなども含めて、要注意。
（https://www.mhlw.go.jp/stf/newpage_36181.html）

43

指定難病患者データベース

医療費助成の申請時に提出される**臨床調査個人票(臨個票)**をベースに構築されている。臨個票には下記のデータが含まれる。

- 1. 基本情報**
生年月、性別、家族歴、発症年月、社会保障（介護認定）、生活状況（移動の程度/身の回りの管理/ふだんの活動/痛み・不快感/不安・ふさぎ込み）、等
- 2. 医療費支給審査項目**
①診断基準：病型分類、臨床所見、検査所見、遺伝学的検査、鑑別診断、診断のカテゴリー（判定基準）
②重症度分類：軽症/中等症/重症、疾患特異的重症度分類
- 3. その他の事項**
臨床所見、検査所見、発症と経過、治療履歴（薬物療法、外科的治療等）
- 4. 人工呼吸器装着の有無**

38

がん登録

がん登録推進法：「がん登録等の情報について、民間を含めがんに係る調査研究に活用、その成果を国民に還元する」

研究者等への提供

（国立がん研究センターがん対策研究所がん登録センター）
https://ganjoho.jp/reg_stat/can_reg/national/datause/researcher.html

41

小児慢性疾病児童等データベース（小慢データベース）

医療費助成の申請時に提出された医療意見書をベースに構築されている。
年齢・性別などの基礎的情報のほか、臨床所見、検査所見、経過、今後の治療方針等の情報が含まれる。

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第三者提供に関するホームページ
https://www.mhlw.go.jp/stf/nanbyou_teikyo.html

39

全国がん登録に含まれる情報

基本情報	①病院等の名称、②診療録番号、③カナ氏名、④氏名、⑤性別、⑥生年月日、⑦診断時住所
腫瘍の種類	⑧側性、⑨原発部位、⑩病理診断
診断情報	⑪診断施設、⑫治療施設、⑬診断根拠、⑭診断日、⑮発見経緯 ⑯進展度・治療前、⑰進展度・術後病理学的
初回治療 観血的治療	⑱外科的、⑲鏡視下、⑳内視鏡的、㉑観血的治療の範囲
初回治療 その他治療	㉒放射線療法、㉓化学療法、㉔内分泌療法、㉕その他治療
生存確認情報	㉖死亡日

42

ご清聴ありがとうございました

44

NDBのデータ申請とオンサイトセンターの利用

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
木村 悠哉

NDBの概要を知るには（重要）

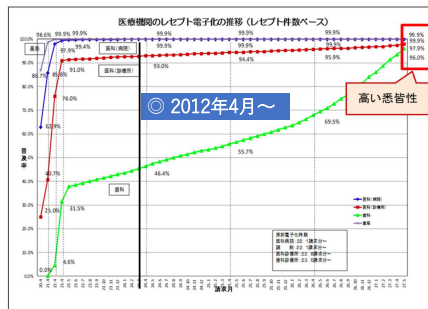
(厚生労働省) [ホーム](#) > [政策について](#) > [分野別の政策一覧](#) > [健康・医療](#) > [医療保険](#) >

【NDB】匿名レセプト情報・匿名特定健診等情報の提供に関するホームページ
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryuu/iryuuoken/reseputo/index.html

こちらに以下の情報が掲載されている

- ・利用を検討している方々へのマニュアル
- ・第三者提供の審査スケジュール
- ・動画による説明
 - 申請手続きについて
 - 参考となるデータベース使用経験

レセプトの電子化率



【図表2】医療機関のレセプト電子化の推移（社会保険診療報酬支払基金調べ）

Agenda

-
- The diagram illustrates a timeline for the NDB process, divided into two main phases by a horizontal dashed line. A vertical arrow on the right indicates the progression of time.
- Before NDB Commitment (厚労省承諾前):** This phase includes the first three items:
 1. NDBの概要
 2. 厚労省への申請 & 倫理申請
 3. 厚労省からの承諾 & データを待つ
 The duration for these steps is indicated as 2~3ヶ月 (2-3 months) for item 2 and 2~6ヶ月 (2-6 months) for item 3.
 - After NDB Commitment (厚労省承諾後):** This phase includes the remaining items:
 4. 抽出データの解析
 5. 厚労省へ公表申請
 6. Q & A
 The duration for these steps is indicated as 3~12ヶ月 (3-12 months) for item 4 and 2~3ヶ月 (2-3 months) for item 5.
- A final summary at the bottom right states: 計 1~2年 (Total 1-2 years).

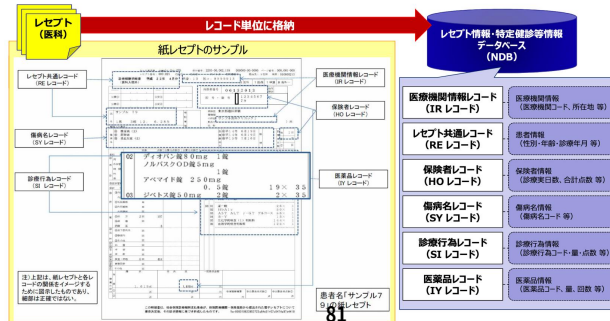
NDBの概要

- 正式名称：レセプト情報・特定健診等情報データベース
(The national database of health insurance claims and specific health checkups in Japan)
- 診療報酬データベース (administrative claims database) の1つ。
医療費を適正化することが本来の目的
(高齢者の医療の確保に関する法律)。
- 2011年から研究利用のための第三者提供が開始した。
- 2009年4月～のレセプトデータが含まれる。



レセプト構造

- レセプト情報は各種レコードに分割され、レセプトはレコード単位でNDBIに格納されています。



Agenda

-
- The diagram illustrates a timeline for a project involving NDB (New Development Bank) and the Thick Labor Province (厚労省). A vertical arrow points downwards, indicating the progression of time. The timeline is divided into two main sections by a horizontal dashed line.
- Before Thick Labor Province Promise (厚労省承諾前):** This section includes the first three items:
 1. NDBの概要 (Overview of NDB)
 2. 厚労省への申請 & 倫理申請 (Application to Thick Labor Province & Ethics Application)
 3. 厚労省からの承諾 & データを待つ (Approval from Thick Labor Province & Waiting for Data)
 The estimated duration for these steps is 2~3ヶ月 (2-3 months) for item 2 and 2~6ヶ月 (2-6 months) for item 3.
 - After Thick Labor Province Promise (厚労省承諾後):** This section includes the remaining three items:
 4. 抽出データの解析 (Analysis of Extracted Data)
 5. 厚労省へ公表申請 (Application for Publication to Thick Labor Province)
 6. Q & A
 The estimated duration for these steps is 3~12ヶ月 (3-12 months) for item 4 and 2~3ヶ月 (2-3 months) for item 5.
- The total estimated duration for the entire process is 計 1~2年 (Total 1-2 years).

NDBに含まれるデータ

- NDBには「電子化された匿名レセプト情報」が格納されています。紙レセプトの情報は格納されていません。
- ○はNDBに格納・提供されている情報です。

[illegible]

NDBの利用を検討している方へのマニュアル 2023/3/16更新 (厚生労働省保健局)

含まれる具体的な項目

- 匿名化個人ID (ID1, ID1N, ID2, ID4, ID5)
- 性別
- 年齢 (5歳刻み)
- 傷病名: 診療開始日, 「疑い」か否か, 「主傷病」か否か
- 診療行為: 行為名 (手術・処置・検査), 数量, 点数, 回数
- 薬剤: 医薬品名, 処方日, 調剤日, 使用量 (1日量), 処方日数
- 特定器材: 特定器材名, 使用量, 点数, 回数
- 死亡: 死亡転帰 (死亡時に何らかの医療行為が行われている場合), 死亡日
- 医療機関: 都道府県, 二次医療圏, 市町村

含まれないデータ

- **自費診療**（妊婦健診・経膈分娩、ワクチン、多焦点眼内レンズ、美容、歯科インプラント）
- **健康保険以外の保険などで行われた診療**（生活保護、労災、交通事故）
- **紙レセプト**（訪問看護、臓器移植）
- **老健や特養の介護施設での診療行為は一部算定不可**
（老健での初・再診料、在宅～料、リハビリ）
- **特定の短期入院は特定入院料に丸められるため診療行為拾えない**
- NDBと介護DBは連結した解析が可能となった（特別抽出のみ）

10

NDBデータの提供形式

1. **集計表情報**
2. サンプルングデータセット
3. 特別抽出
4. オンサイトセンター
5. NDBオープンデータ

2. サンプルングデータセット

- 1ヶ月分の抽出データ。
- 2011年1月以降の1月、4月、7月、10月
- 医科入院レセ、医科入院外レセ、調剤レセ、DPCレセの4種から、一定の割合でサンプリングされる。
- データを抽出後、研究者自ら記述統計や横断分析を行うことになる。

例：2012年4月診療分の**医科(入院外)**と**調剤**のデータセット抽出を依頼する。
自ら「**2012年4月の降圧薬の処方実態**」を解析する。

NDBの研究利用

- 記述統計研究
 - 薬剤の使用状況、疾病の分布
- 政策評価研究
 - 政策導入前後での受療行動の変化
- Ecological研究
 - 地域間で行われている医療行為の差異
- 治療やリスクファクターとアウトカムの研究（因果推論）
 - 傾向スコア、操作変数法
- 予測モデル
 - 機械学習

1. 集計表情報

- 提供依頼申出者の抽出依頼に応じて、集計表が提供される。
- 複雑な集計には対応できない。
- 抽出条件は、傷病名コード・診療行為コード・医薬品コード・特定機材コード等に限定し、それらに対して性別、年齢階級別、都道府県別等、原則3次元まで認められる。
- 集計表のイメージ図の提出が必要。
- データ提供されるまで時間がかかる（6ヶ月以上？）
「手続き終了後、本来業務など他の作業の合間に行われるため、遅延の可能性がある」

例：気管支鏡下レーザー腫瘍焼灼術（K510-3）を行った患者を抽出
⇒ 性別、年齢階級別、都道府県別の3次元で集計表の作成を依頼。

サンプルングデータセットのメリット

- 患者単位で処理可能
- 代表性に優れている
- 依頼者が指定した項目だけ抽出される「特別抽出」と異なり、多くの項目が含まれているため、疑問点の検証が簡単
- ハイスペックPCは不要
- 提供されるまでの時間が短い

NDBデータの提供形式

1. 集計表情報
2. サンプルングデータセット
3. 特別抽出
4. オンサイトセンター
5. NDBオープンデータ

NDBデータの提供形式

1. 集計表情報
2. **サンプルングデータセット**
3. 特別抽出
4. オンサイトセンター
5. NDBオープンデータ

サンプルングデータセットのデメリット

- 高額なレセは除外される
 - ✓ 70万点以上の医科入院レセは除かれる
- 希少な傷病名（**≦0.1%**）・診療行為（**<0.01%**）等は匿名化される。
- 出現回数少ないコードから順に、総出現回数の0.1%に達するまで匿名化。
- 1ヶ月分のみのデータであるため、時系列分析はできない。
 - ✓ 季節性変化は解析できない

NDBデータの提供形式

- 1. 集計表情報
- 2. サンプルングデータセット
- 3. 特別抽出
- 4. オンサイトセンター
- 5. NDBオープンデータ

特別抽出のデメリット

- データ抽出条件の指定が難しい
 - ✕「高齢者医療に興味がある。とりあえず65歳以上の患者レセに含まれる傷病名，診療行為，医薬品情報をすべて抽出したい」
 - 「**担癌患者における術前の口腔ケアは術後合併症を減らせるか**」
⇒ 担癌患者を抽出し，さらに医科レセ，歯科レセ，調剤レセ，DPCレセを依頼
- データ抽出後の解析が難しい
 - ✓ 月単位・医療機関単位で発行されるレセ情報の統合，患者IDの紐付け
 - ✓ 医科外来は月8000万件が格納される→ ビッグデータを扱える環境や人材
- データ提供までに時間かかる（6ヶ月以上？）
「手続き終了後、本来業務など他の作業の合間に行われる ため、遅延の可能性がある」

4. オンサイトセンター

- 厚生労働省、東京大学、京都大学の3拠点に存在する。
（2024年9月現在）
- 研究者自身がオンサイトセンターで解析を行う。
- NDBの全データを利用可能。
- 利用期間は6ヶ月（申請により延長可能）
- オンサイトリサーチセンターの取扱者のうち一人以上は、
第三者提供の個票抽出の利用経験を有する者
かつ
SQLの知識を持ち、CSVファイルをEXCELファイル等で加工できる者

3. 特別抽出

- 依頼者の申出内容に応じて抽出・提供される。
- ほぼすべての項目を利用（抽出）することができる。
- 個票データである。
- 以下の事項について専門委員会で審査される
 - ✓研究内容が公益性を備えたものか
 - ✓研究内容を達成するにあたりレセ情報が適切か
 - ✓抽出条件が，研究内容の達成のために適切か
 - ✓データの漏洩や紛失が発生しないよう，セキュリティは万全か

厳しいセキュリティ要件（物理的環境）

- データ利用者は利用申請者のみに限定される。
- 運用管理規定，内部監査規定
- 個人情報保護に関する方策を策定し，公開する。
- データ分析専用室（施設可能な物理的スペース）
- 専用室への入退室管理（台帳など）
- データ分析専用端末
- データ解析端末のインターネット接続は禁止
- 中間生成物の持ち出しは原則禁止。



データベースの整備・保守管理その他（オープンデータ等の扱い）， 厚労省より

特別抽出のメリット

- レセプト情報・特定健診等情報に含まれる，ほぼすべての項目を扱える。

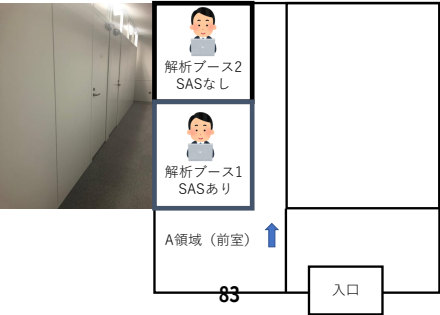
NDBデータの提供形式

- 1. 集計表情報
- 2. サンプルングデータセット
- 3. 特別抽出
- 4. オンサイトセンター
- 5. NDBオープンデータ

NDBデータの提供形式

- 1. 集計表情報
- 2. サンプルングデータセット
- 3. 特別抽出
- 4. オンサイトセンター
- 5. NDBオープンデータ
<https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000177182.html>
 - NDBデータから汎用性の高い基礎的な集計表を作成し、「NDBオープンデータ」として公表
 - Excelの集計データをダウンロードできる

オンサイトセンター（東京大学，教育研究棟13F）

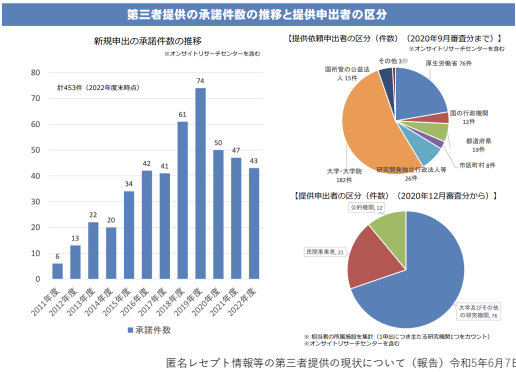


NDBオープンデータ分析サイト



https://www.mhlw.go.jp/ndb/opendatasite/

提供の現状



匿名レセプト情報等の第三者提供の現状について（報告）令和5年6月7日

NDBデータの提供形式

- 1. 集計表情報
- 2. サンプルングデータセット
- 3. 特別抽出
- 4. オンサイトセンター

上記4種のメリット、デメリットを理解したうえで、厚生労働省へのデータ利用申請をすすめていく。
今回は特別抽出の申請方法について、主に説明していく。
(オンサイトセンターの書類は特別抽出と類似)

Agenda

- 1. NDBの概要
 - 2. 厚生労働省への申請 & 倫理申請
 - 3. 厚生労働省からの承諾 & データを待つ
 - 4. 抽出データの解析
 - 5. 厚生労働省へ公表申請
 - 6. Q & A
- 2〜3ヶ月
2〜6ヶ月
3〜12ヶ月
2〜3ヶ月
計 1〜2年

厚生労働省承諾前

厚生労働省承諾後

厚生労働省窓口（重要）

問い合わせ

厚生労働省では、事務処理を円滑に行うため匿名レセプト情報等の申請者の事前相談や申請書の受付窓口を設けております。第三者提供に関するお問い合わせにつきましてはこちらの窓口をご利用くださいますようお願いいたします。

※ お問い合わせはE-mailにてお願いしております。

株式会社NTTデータ 第二公共事業本部 デジタルウェルフェア事業部 匿名レセプト情報等第三者提供窓口 宛

住所 〒135-8671 東京都江東区豊洲3-3-9 豊洲センタービルアネックス19階

E-mail teikyo_rezept@kits.nttdata.co.jp

- ・申請書類作成「前」に問い合わせる。
- ・厚生労働省窓口とやり取りを重ねることで、申請書類が完成する。

申請に必要な書類



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

様式1（研究計画書的なもの）

申請に必要な書類

通常の申請用

- 初回申請時に必要な書類
- ▶ X 様式1 匿名レセプト情報等の提供に関する申請書 [Excel: 75KB]
- ▶ X 様式1 記入例 [Excel: 101KB]
- ▶ P 別添2-1 運用フロー図 記入例 (外部委託無し) [PowerPoint: 272KB]
- ▶ P 別添2-1 運用フロー図 記入例 (外部委託有り) [PowerPoint: 266KB]
- ▶ X 別添2-2 リスク分析・対応表 記入例 [Excel: 34KB]
- ▶ W 別添2-3 運用管理規程 記入例 [Word: 35KB]
- ▶ W 別添2-4 (参考例) 自己点検規程 [Word: 28KB]
- ▶ X 別添8 申出依頼テンプレート (集計) [Excel: 954KB]
- ▶ X 別添8 申出依頼テンプレート (抽出) [Excel: 946KB]
- ▶ X 別添8 申出依頼テンプレートの注意点 [Excel: 638KB]



様式1



各シートを埋めていく

- ・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研究内容は認められません。
- ・他の研究や政策利用等を阻害するような特許を取得するものについても認められません。

- ・利用場所、保管場所の名称は、具体的な場所を特定可能な記載にしてください。
- ・別添2系と整合性がとれるように記載をお願いします。特に場所名等の名称は統一してください。

抽出期間は以下の範囲に基づいて指定してください。

自：レセプトは2009年4月診療分～、特定健診は2008年度実施分～

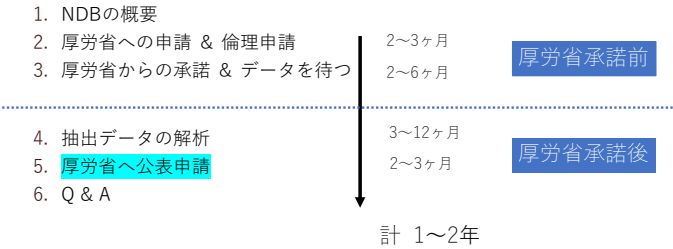
至：レセプトは診療月の6ヶ月後の抽出を推奨します。特定健診は実施年度の翌年度2月頃にNDBIに格納されます。

患者ID 1 : 保険者番号等を元に生成したID
患者ID 2 : 氏名等を元に生成したID

抽出データの解析

- 特別抽出
 - ✓DVDもしくはHDDでデータが届く
 - ✓セキュリティ要件を満たす部屋にデータを保存し、解析を進める
- オンサイトセンター
 - ✓オンサイトセンターの解析部屋で解析を進める

Agenda



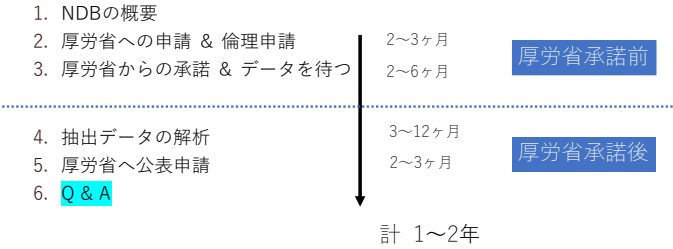
公表申請

- 解析結果は、厚生労働省の公表申請を経なければならない
 - 公表申請で許可が降りて初めて学会・抄録・論文で数字を出して良い
 - 許可が出るまで1~2ヶ月かかる
- ★公表申請前のデータを外部に発表してしまった研究班が、NDB無期限使用禁止になった事例がある。

公表申請 -予め知っておくべきこと-

- 最小集計単位
 - ✓10未満の患者数やレセプト数などは公表不可。
 - ✓医療機関や保険者なども、対象機関が3未満は公表不可。
- 年齢区分
 - ✓5歳毎にグループ化して集計し、85歳以上の場合は同一グループとする

Agenda



NDB Q&A

- 処方日や診療行為実施日ってわかる？
 - 2012年4月以降ならわかるため、この時期のデータ利用をおすすめ。
- 病名のついた日はわかる？
 - 医科レセで診療開始日がわかる。
- 病名の日本語検索できる？
 - NDBでは日本語コメントが削除されているため、不可能。

NDB Q&A

- 家族の連結は可能か？
 - できない。
- 医療機関の特性はどの程度わかるか？
 - ✓匿名化されたID（医療機関コード）のみわかる
 - ✓病院かクリニックかを完璧に分類することはできない
 - ✓A100一般病棟入院基本料、A101療養病棟入院基本料、A104特定機能病院入院基本料、A105専門病院入院基本料 などを用いて、ある程度は把握できる可能性はある。

最後に

- NDB研究は時間がかかるため、長期的視野が必要。
- 他のデータベースの使用も検討する。

2024年9月25日（水）
11:00-11:50

NDB・DPC等データベース研究人材育成Webinar

NDB研究のデザインと実践

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
木村 悠哉

NDB研究のデザインの傾向

- NDB, DPC, JMDCにおける研究デザインごとの割合は下記の通り.

記述研究, 因果推論研究(探索的), 因果推論研究(検証的), その他

- NDB: **62.7%** (42/67), 9.0% (6/67), 28.4% (19/67), 0.0% (0/67)
- DPC: 13.6% (35/257), 11.3% (29/257), 70.8% (182/257), 4.3% (11/257)
- JMDC: 41.7% (73/175), 12.6% (22/175), 40.6% (71/175), 5.1% (9/175)

Jun Fujinaga et al. *Drugs Real World Outcomes*. 2022; 9(4): 543-550.

NDB研究のデザインとしては記述研究が多く、因果推論研究が少ない。
主に以下2点が原因と考えられる。
①データハンドリングの負荷が大きいNDBで因果推論研究を実施することが時間的・技術的に厳しい
②因果推論研究で調整すべき変数がNDBでは取得ができない

NDBの長所・短所

- 長所：**悉皆性の高さ**
→記述研究、因果推論研究のいずれにしろ一般化可能性が高い。
→記述研究としてrare diseaseやrare eventの記述も可能。
→因果推論研究ではサンプル数を稼ぐことができる。

- 短所：**交絡因子として採用できる変数が少ない**

- その他の短所は諸々ある
 - 要求されるレセプトへの理解, データハンドリング能力
 - 利用申請の煩雑さ
 - など

Agenda

- NDB研究のデザインの傾向
- NDBの長所・短所
- NDB研究のデザインで意識すべきこと
- 実研究の紹介

NDB研究のデザインの傾向

- NDB, DPC, JMDCにおける因果推論研究の細かいデザインは下記の通り.

cohort, cross-sectional, case-control, quasi-experimental, self-controlled

- NDB: **17.9%** (12/67), 7.5% (5/67), 1.5% (1/67), 7.5% (5/67), 1.5% (1/67)
- DPC: **45.9%** (118/257), 0.8% (2/257), 2.0% (5/257), **37.4%** (96/257), 0.0% (0/257)
- JMDC: **32.6%** (57/175), 2.9% (5/175), 6.3% (11/175), 10.9% (19/175), 2.3% (4/175)

Jun Fujinaga et al. *Drugs Real World Outcomes*. 2022; 9(4): 543-550.

Agenda

- NDB研究のデザインの傾向
- NDBの長所・短所
- NDB研究のデザインで意識すべきこと**
- 実研究の紹介

Agenda

- NDB研究のデザインの傾向**
- NDBの長所・短所
- NDB研究のデザインで意識すべきこと
- 実研究の紹介

Agenda

- NDB研究のデザインの傾向
- NDBの長所・短所**
- NDB研究のデザインで意識すべきこと
- 実研究の紹介

NDB研究のデザインで意識すべきこと

- 長所は活かし、短所は補う！**他のDBではできない研究をする！！

- 記述研究
希少疾患や希少イベントの記述が可能。
ほぼ全国民を対象とした巨大な記述が可能（ほぼすべての医療機関, 年齢層, 地域, 入院外来）。

- 因果推論研究
考慮すべき交絡因子の情報をNDBから取得できそうであれば因果推論におけるデザインは自由。
もし取得できなさそうであればself-controlledなど、考慮すべき交絡因子の情報を取得できない状況下でも因果推論できそうなデザインを採用。

健診データを使用することでDPC様式1にも含まれない様な情報を交絡因子として利用可能。健診データがある人は限られるが(2-4割?), NDBではN数が膨大なので、それでもなお健診データがあるN数は多いはず。

健診データに含まれる情報

- 喫煙歴
 - BMI
 - 腹囲
 - 血圧
 - 健康関連行動(飲酒歴, 運動習慣, 歩行速度など)
 - 採血データ(HbA1c, 中性脂肪, LDL-C, Creなど)
- など

リアルワールドデータのレビュー研究紹介
NDB含めたRWDでどのような研究ができるか

Kimura Y *et al.* Clinical research using real-world data: a narrative review. *Respiratory Investigation*. 2024; 62(6): 929-934.

NDB, DPCなどの各種データベースの違いやそれぞれにおいてどのような研究ができるかを紹介しました。各種データベースに馴染みが薄い方には参考になるかもしれません。

Agenda

- NDB研究のデザインの傾向
- NDBの長所・短所
- NDB研究のデザインで意識すべきこと
- 実研究の紹介

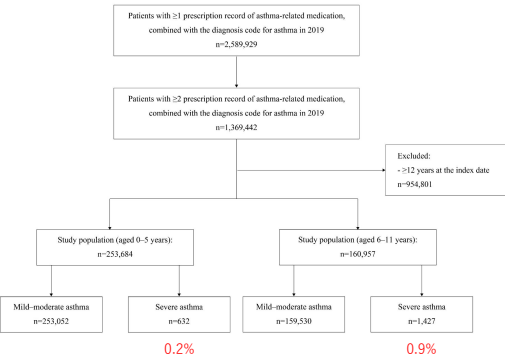
NDB研究紹介①(記述)
小児喘息のうち難治性の割合はどの程度か

Kimura Y *et al.* Epidemiology of severe childhood asthma in Japan: A nationwide descriptive study. *Allergy*. 2024;79(6):1598-1602.

- 難治性喘息の定義は2010年代に確立し, 喘息全体における難治性喘息の割合は5-10%と推定された。しかしながら, 小児においてその推定の根拠となる論文は, 1990年代のサンプルサイズの小さなbirth cohortのみである。
- そこで, 2010年代の定義に基づき, 小児喘息全体における難治性喘息の割合を記述することを目的とした。

研究デザイン

- year Y (2013, 2015, 2017, 2019年)におけるcross-sectionalデザイン。
- 喘息患者は喘息病名コードと喘息治療薬剤の組み合わせをyear Yに複数回有する症例と定義した。
- 難治性喘息の定義は2010年代に確立したものにに基づき, 喘息治療の強度が一定以上高いものと定義した。
- year Yの人口データから非難治性喘息, 難治性喘息の有病割合を算出した。
- 0-5歳と6-11歳では喘息の治療方針が若干異なるため, 別個に解析をした。



Kimura Yuya *et al.* *Allergy*. 2024;79(6):1598-1602 > Figure S1

Children aged 0-5 years	Prevalence of mild-to-moderate asthma ¹				Prevalence of severe asthma ¹			
	2013	2015	2017	2019	2013	2015	2017	2019
Total	5351.9 (5334.3-5369.45)	5434.4 (5416.3-5452.5)	4751.0 (4733.9-4768.1)	4413.2 (4396.4-4430.0)	19.7 (18.6-20.8)	16.6 (15.6-17.7)	11.9 (11.1-12.9)	11.0 (10.2-11.9)
Sex								
Males	5850.8 (5825.3-5876.5)	5929.8 (5903.5-5956.1)	5187.8 (5162.9-5212.8)	4788.1 (4763.7-4812.6)	22.0 (20.4-23.7)	18.4 (16.9-20.0)	14.0 (12.8-15.4)	12.7 (11.4-14.0)
Females	4829.5 (4805.6-4853.5)	4916.9 (4892.3-4941.6)	4291.5 (4268.2-4314.9)	4021.2 (3998.3-4044.3)	17.3 (15.8-18.8)	14.8 (13.5-16.2)	9.7 (8.7-11.0)	9.3 (8.2-10.5)
Age group								
0-1	3925.1 (3898.8-3951.5)	4129.0 (4091.3-4147.1)	3670.5 (3644.3-3696.9)	3458.5 (3432.1-3485.0)	8.4 (7.3-9.8)	9.5 (8.2-11.0)	7.3 (6.2-8.6)	7.1 (6.0-8.5)
2-3	5745.1 (5713.8-5776.6)	5841.9 (5809.7-5874.3)	5222.7 (5191.4-5254.1)	4674.9 (4645.4-4704.5)	14.4 (12.9-16.2)	11.3 (9.9-12.9)	9.4 (8.1-10.8)	8.8 (7.6-10.2)
4-5	6364.3 (6331.5-6397.3)	6256.0 (6223.3-6288.9)	5346.7 (5315.8-5377.7)	5052.7 (5022.0-5083.7)	35.9 (33.5-38.4)	28.4 (26.2-30.8)	18.9 (17.1-20.9)	16.9 (15.2-18.9)

Children aged 6-11 years	Prevalence of mild-to-moderate asthma ¹				Prevalence of severe asthma ¹			
	2013	2015	2017	2019	2013	2015	2017	2019
Total	2829.5 (2816.8-2842.2)	2779.4 (2766.3-2792.5)	2479.6 (2479.5-2503.8)	2545.2 (2532.9-2557.5)	44.4 (42.8-46.1)	38.3 (36.8-39.8)	25.7 (24.5-27.0)	22.8 (21.6-24.0)
Sex								
Males	3225.2 (3204.4-3244.2)	3358.0 (3338.4-3377.5)	2800.1 (2782.3-2818.1)	2831.3 (2813.2-2849.5)	52.3 (49.9-54.8)	44.8 (42.6-47.2)	29.7 (27.9-31.6)	25.7 (24.0-27.5)
Females	2413.5 (2396.8-2430.4)	2583.1 (2565.6-2600.7)	2166.8 (2150.7-2183.3)	2244.8 (2228.2-2261.4)	36.2 (34.2-38.3)	31.4 (29.5-33.4)	21.5 (19.9-23.2)	19.7 (18.1-21.3)
Age group								
6-7	3809.2 (3783.5-3834.9)	3942.9 (3916.9-3968.1)	3295.8 (3271.8-3320.1)	3349.6 (3324.9-3374.6)	47.4 (44.5-50.4)	39.9 (37.4-42.7)	27.2 (25.1-29.5)	23.3 (21.2-25.4)
8-9	2763.3 (2738.7-2788.2)	2896.8 (2874.3-2919.4)	2421.6 (2411.1-2432.4)	2496.9 (2475.9-2518.2)	45.5 (42.7-48.4)	41.8 (39.1-44.4)	26.7 (24.6-29.0)	24.1 (22.1-26.3)
10-11	1974.6 (1956.6-1992.8)	2110.5 (2091.5-2129.7)	1763.6 (1746.1-1781.4)	1829.9 (1812.1-1848.0)	40.7 (38.1-43.4)	33.1 (30.8-35.6)	23.2 (21.2-25.3)	21.0 (19.1-23.0)

Kimura Yuya *et al.* *Allergy*. 2024;79(6):1598-1602 > Table 1

NDB研究紹介②(記述+cohort)
慢性肺アスペルギルス症の治療実態と予後はいかに

Kimura Y *et al.* Epidemiology of chronic pulmonary aspergillosis: a nationwide descriptive study. *Respiratory Investigation*. In Press.

- 慢性肺アスペルギルス症は肺疾患を有する患者に生じる慢性感染症である。ガイドラインでは, 抗真菌薬の6ヶ月以上の継続が推奨されている。また, 近年のRCTでは6ヶ月治療と比較して12ヶ月治療を受けた群で予後が良かったことが示された。しかしながら, その治療実態を記述した研究は限られる。
- そこで, 同疾患に対する抗真菌薬治療の実態を記述することを目的とした。さらに, 背景肺疾患により予後がどう異なるかを明らかにすることを目的とした。

研究デザイン（記述部分）

- year Y (2016, 2018, 2020, 2022年)におけるcross-sectionalデザイン。
- 慢性肺アスペルギルス症患者は同疾患の病名コードと抗真菌薬の組み合わせを201401-202212に少なくとも1回有する患者とした。そのうちyear Yに少なくとも1回の病名コードを有する患者を"prevalent case in year Y"とし、そのうちyear (Y-1)以前に病名コードを有さない患者を"incident case in year Y"と定義した。
- Incident cases in 2016における抗真菌薬の継続状況を記述した。

19

研究デザイン（cohort部分）

- Incident case in 2016-2022の生存率をKaplan-Meier曲線で提示した。患者のフォローアップ開始は診断日とし、フォローアップ終了は死亡または最終のレセプト発生月月末とした。
- 以下の変数を説明変数に組み込んだ多変量Cox回帰モデルでそれぞれの変数のハザード比を算出した。
 - 年齢
 - 性別
 - 背景肺疾患(6種類)
 - 全身性ステロイドの使用

20

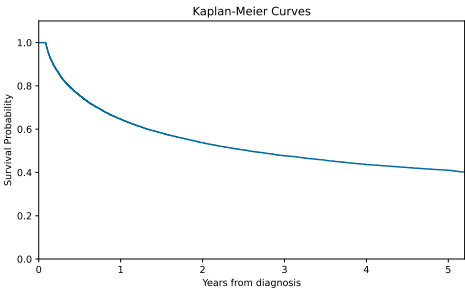
Incident cases in 2016	
全体	4490 (100%)
ITCZを開始した患者	1645 (36.6%)
VRCZを開始した患者	1911 (42.6%)
抗真菌薬を開始しなかった患者	934 (20.8%)
ITCZを開始した患者	
全体	1645 (100%)
6ヶ月以上ITCZを継続した患者	459 (27.9% ^a ; 35.3% ^b)
6ヶ月以上抗真菌薬を継続した患者	576 (35.0% ^a ; 44.2% ^b)
12ヶ月以上ITCZを継続した患者	247 (15.0% ^a ; 22.4% ^b)
12ヶ月以上抗真菌薬を継続した患者	324 (19.7% ^a ; 29.3% ^b)
VRCZを開始した患者	
全体	1911 (100%)
6ヶ月以上VRCZを継続した患者	533 (27.9% ^a ; 35.6% ^b)
6ヶ月以上抗真菌薬を継続した患者	648 (33.9% ^a ; 43.2% ^b)
12ヶ月以上VRCZを継続した患者	259 (13.6% ^a ; 20.6% ^b)
12ヶ月以上抗真菌薬を継続した患者	334 (17.5% ^a ; 26.5% ^b)

a. 全体における割合
b. 生存者における割合

Kimura Yuya et al. Respiratory Investig. In press
> Table 3を改変

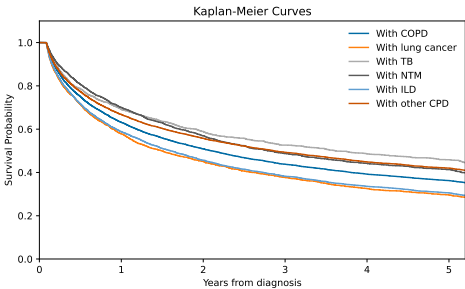
21

KM曲線（全体）



Kimura Yuya et al. Respiratory Investig. In press > Figure 3より抜粋

KM曲線（背景肺疾患ごと）



Kimura Yuya et al. Respiratory Investig. In press > Figure 3より抜粋

多変量Cox回帰モデルの結果

	患者数	ハザード比（95%信頼区間）
全体	27888 (100%)	
65歳以上	19869 (71.2%)	2.65 (2.54-2.77)
男性	18626 (66.8%)	1.24 (1.20-1.29)
COPD	11871 (42.6%)	1.05 (1.02-1.09)
肺がん	2701 (9.7%)	1.12 (1.06-1.18)
肺結核	2144 (7.7%)	0.81 (0.76-0.86)
NTM	2860 (10.3%)	0.91 (0.86-0.96)
間質性肺疾患	6496 (23.3%)	1.19 (1.14-1.24)
その他の肺疾患	10020 (35.9%)	0.89 (0.85-0.92)
全身性ステロイドの使用	9624 (34.5%)	1.13 (1.09-1.17)

Kimura Yuya et al. Respiratory Investig. In press > Table 2を改変

24

NDB研究紹介③(因果推論)

県レベルのICUベッド数と人工呼吸器使用患者の死亡率の相関はあるか

Ohbe H et al. Regional Critical Care Bed Capacity and Incidence and Mortality of Mechanical Ventilation in Japan. *Am J Respir Crit Care Med.* 2024;210(3):358-361.

- 人工呼吸器使用患者の管理はICUで行うことがガイドラインで推奨されている。しかし、ICUベッド数が少ない地域ではICU外でその管理が行われることがある。よって、地域のICUベッド数と人工呼吸器使用患者の死亡率の相関に興味を持たれるが、これを探索した研究はない。
- そこで、県レベルのICUベッド数と人工呼吸器症患者の死亡率に相関があるかを確認することを目的とした。

25

研究デザイン

- cohortデザイン。
- 201904-202003に人工呼吸器を使用した患者（ただし、手術やリハビリに関連したものは除く）を対象とした。
- 県レベルのICUベッド数についてのデータは『令和元(2019)年医療施設(動態)調査・病院報告』より取得した。
- 47都道府県について、人工呼吸器使用後1年における調整済み死亡率を算出した。調整は、年齢、性別、Charlson comorbidity index、主病名を説明変数に組み込んだ多変量ロジスティック回帰モデルに基づいた。

26

メインの結果

- 100,000人あたり1つのICUベッドが増えるごとに、調整済み1年死亡率は0.78%(-1.11% to -0.45%; p<0.001)低下した。
- 100,000人あたり1つのHCUベッドが増えるごとに、調整済み1年死亡率は0.23%(-0.46% to 0.01%; p=0.061)低下した。

27

NDB研究紹介④(因果推論)

新規糖尿病治療薬はメトホルミンと比較して喘息増悪を減少させるか

Kimura Y *et al.* Association of Novel Antihyperglycemic Drugs Versus Metformin With a Decrease in Asthma Exacerbations. *J Allergy Clin Immunol Pract.* 2024; S2213-2198(24)00467-7.

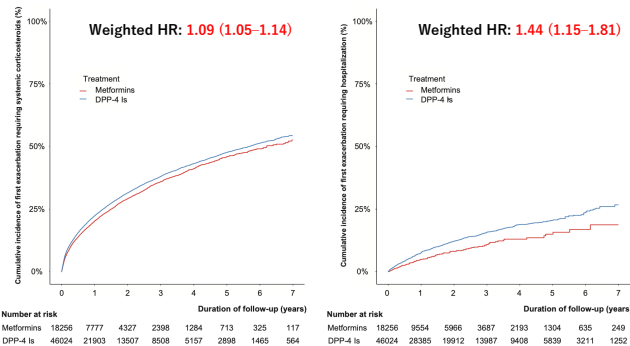
- かねてより欧米で糖尿病治療の第一選択薬であったメトホルミンには抗炎症効果があるとされている。過去10年間ほどで出現したDPP-4 Is, GLP-1 RAs, SGLT-2 Isにも同様の抗炎症効果があることを示唆する研究がある。抗炎症効果により喘息増悪が抑制される可能性があるが、メトホルミンと比較して新規糖尿病治療薬が喘息増悪を減少させるか探索した研究はない。
- そこで、メトホルミンと比較して、DPP-4 Is, GLP-1 RAs, SGLT-2 Isそれぞれが喘息増悪を減少させるのかを検証することを目的とした。

研究デザイン

- cohortデザイン。DPP-4 Is, GLP-1 RAs, SGLT-2 Isのそれぞれについてcohortを作成した。
- 201401-202212にメトホルミンまたは新規糖尿病治療薬を新規に開始した喘息患者を対象とした。
- 主要アウトカムを『全身性ステロイド投与を要する喘息増悪』, 副次的アウトカムを『入院を要する喘息増悪』とした。
- フォローアップ期間の開始をメトホルミンまたは新規糖尿病治療薬とし、フォローアップ期間の終了をアウトカム発生, 薬の中止または変更, 死亡, 最終のレセプト発生月月末, の最も早いものとした。

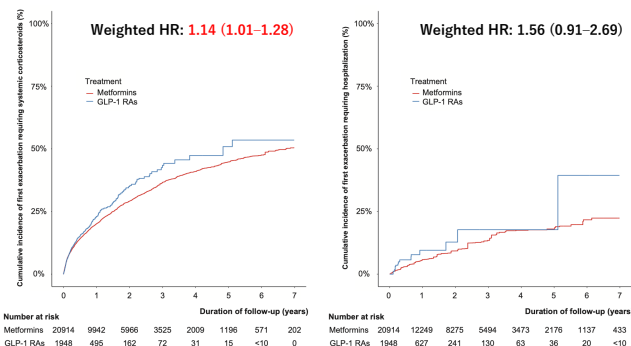
研究デザイン

- propensity scoreに基づく重みづけによって患者背景の調整を行った上で、Cox 比例ハザードモデルにより新規糖尿病治療薬 (vs メトホルミン) のハザード比を算出した。なお、propensity scoreの算出は下記共変量を説明変数にしたロジスティック回帰モデルに基づいた。
- 共変量
 - 年齢
 - 性別
 - 合併症
 - 薬
 - healthcare resource use
 - 健康関連行動
 - 採血データ



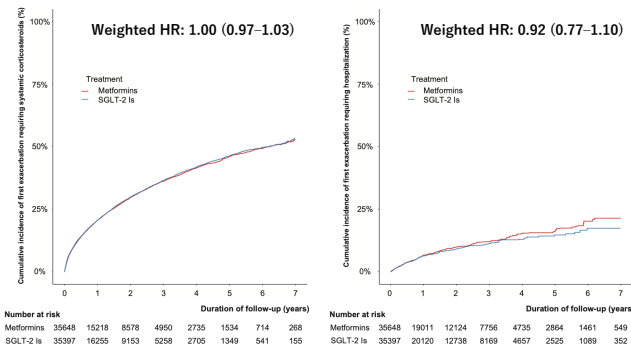
DPP-4 Is cohort

Kimura Yuya *et al.* *J Allergy Clin Immunol Pract.* 2024; S2213-2198(24)00467-7. > Figure 2より抜粋



GLP-1 RAs cohort

Kimura Yuya *et al.* *J Allergy Clin Immunol Pract.* 2024; S2213-2198(24)00467-7. > Figure 2より抜粋



SGLT-2 Is cohort

Kimura Yuya *et al.* *J Allergy Clin Immunol Pract.* 2024; S2213-2198(24)00467-7. > Figure 2より抜粋

研究のまとめ

- メトホルミンと比較して、
 - DPP-4 Isは全身性ステロイドの投与を要する喘息増悪, 入院を要する喘息増悪のいずれも増やした。
 - GLP-1 RAsは全身性ステロイドの投与を要する喘息増悪を増やしたが, 入院を要する喘息増悪については有意には増やさなかった。
 - SLGT-2 Isは全身性ステロイドの投与を要する喘息増悪, 入院を要する喘息増悪のいずれも同等であった。

メトホルミン ≒ SLGT-2 Is > GLP-1 RAs ≒ (?) DPP-4 Is

まとめ

- NDBの長所は**悉皆性の高さ**である。記述研究では希少疾患・希少イベントの記述が可能であり、因果推論研究では膨大なサンプル数を稼ぐことができる。
- NDBの短所の1つは**交絡因子として採用できる変数が少ない**ことである。考慮すべき交絡因子の情報を取得できない場合にはself-controlledデザインなどの研究デザインの工夫を考える。また、健診データに含まれる変数の利用を考える。

NDBの病名コードの妥当性は？

- NDBの病名コードを直接的に検証した研究はこれまでにない。

NDB研究紹介（補足）

NDBにおける疾患定義の妥当性を『間接的に』検証できないか

Kimura Y *et al.* Screening rates for HIV and diabetes in patients with active TB: results of a nationwide survey in Japan. *ITJLD Open*. 2024; 1(7): 326-328.

	2016	2017	2018	2019	2020	2021
研究で同定された患者数	19775	18537	17181	15975	13370	12196
国から報告された患者数	17625	16789	15590	14460	12739	11519
差分の絶対値(%)	12.2%	10.4%	10.2%	10.5%	5.0%	5.9%

- 結核患者を病名コード+薬剤コードで定義
- 『国から報告された患者数』は全数把握に基づくものであり, ground truthと見なせる.

ずれは小さい（最大10%程度）.
ただし, 病名コードのみで定義した場合には, その差は非常に大きい. (data not shown)

個人的な考え

- Reviewer, Editorに納得してもらえる対象患者の定義を定める必要がある.
例. 病名コード+薬剤コード（@NDB研究紹介①②④）
- NDBにおいて病名コードのみで対象患者を定義することは（避けられる場合には）避けた方が無難.
- 先行研究で使われた定義や臨床経験を元に妥当な定義を模索することで, Reviewer, Editorに納得してもらえる様努力する.

DPCデータ研究のデザインと実践

東京大学大学院医学系研究科臨床疫学・経済学
康永秀生

1

Diagnosis Procedure Combination (DPC) とは

- わが国で独自に開発された診断群分類。
- 患者集団を診断 (Diagnosis) と処置 (Procedure) を用いて分類するシステム。
- 一日あたり包括支払システム (DPC/PDPS) にも利用されている。
- DPCを採用する病院は大学病院をはじめ比較的規模の大きい急性期病院。

4

(v) 診療情報

- ①入院時・退院時の ADL スコア
- ②がんの初発、再発、UICC 病期分類、がんの Stage 分類、化学療法の有無
- ③FIM (Functional Independence Measure) ④入院時・退院時JCS (Japan Coma Scale)
- ⑤脳卒中患者の発症前および退院時mRS (modified Rankin Scale) 、脳卒中の発症時期
- ⑦ Hugh-Jones 分類、肺炎の重症度分類 (A-DROP) 、医療介護関連肺炎に該当の有無
- ⑧救急受診時の P/F 比、救急受診時の酸素投与の有無、救急受診時の FiO2、救急受診時の呼吸補助の有無、治療室又は病棟入室時の P/F 比、治療室又は病棟入室時の酸素投与の有無、治療室又は病棟入室時の FiO2、治療室又は病棟入室時の呼吸補助の有無
- ⑨心不全患者の救急受診時の NYHA、治療室又は病棟入室時の NYHA、入院時BNP・NTproBNP
- ⑩狭心症の CCS 分類、急性心筋梗塞の Killip 分類、急性心筋梗塞の発症時期
- ⑪救急受診時の収縮期血圧、救急受診時の循環作動薬の使用、治療室又は病棟入室時の収縮期血圧、治療室又は病棟入室時の循環作動薬の使用
- ⑫ 解離性大動脈瘤Stanford A/B型 ⑬肝硬変の Child-Pugh 分類 ⑭急性肺炎の重症度分類
- ⑮入院周辺の分娩の有無、分娩時出血量 ⑯川崎病患者へのガンマグロブリンの追加治療の有無
- ⑰ Burn Index
- ⑱入院時 GAF 尺度、精神保健福祉法における入院形態、精神保健福祉法に基づく隔離日数、精神保健福祉法に基づく身体拘束日数
- ⑲退院に向けた会議の開催状況、個別支援の実施状況、外出又は外泊の実施状況、障害福祉サービス等の連携に関する情報
- ⑳ SOFAS コア、pSOFAS コア
- など

7

Topics

- DPCデータの概略
- DPCデータを用いた研究の実例

2

さまざまなDPCデータベース

(1)厚生労働省のDPCデータベース (DPCDB)

平成29年度から厚生労働省が保有する全DPC病院のデータの集計データを提供。2022年4月から個票データの提供も開始。

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

(2)学会・学術団体が構築するDPCデータベース

厚生労働科学研究・DPCデータ調査研究班
日本循環器学会・循環器疾患診療実態調査(JROAD)
など

(3)民間企業が提供する商用のDPCデータベース

Medical Data Visionなど

5

E F ファイルから得られるデータ

(i)薬剤・特定保険医療材料の名称・使用日・使用量

(ii)検査・処置の実施日・回数

麻酔、手術、リハビリテーション、気管内挿管、人工呼吸、血液浄化などの、個別の医療行為の実施。
術後在院日数、集中治療室の滞在日数、人工呼吸の期間、胸腔ドレナージの期間、なども算出可能。

(iii)医療費

出来高換算された診療費のデータを用いて、入院基本料、手術・麻酔、投薬、注射、検査、放射線治療、食事料などの費用データも算出可能。

93

8

1. DPCデータの概略

DPCデータ 様式1の項目

(i) 入院情報

施設コード、診療科コード、入院年月日、退院年月日、入院経路、他院よりの紹介の有無、自院の外来からの入院、予定・救急医療入院の別、救急車による搬送、退院先、退院時転帰、24時間以内の死亡の有無、入院前・退院後の在宅医療の有無、前回退院年月日、前回同一疾病で自院入院の有無、再入院種別、再転棟種別、自傷行為・自殺企図の有無など

(ii) 患者背景

生年月日、性別、患者住所地の郵便番号、身長、入院時・退院時体重、喫煙指数、入院時・退院時の褥瘡の有無、入院中の褥瘡の最大深度等、現在の妊娠の有無、入院時の妊娠週数、出生時体重、出生時妊娠週数、認知症高齢者の日常生活自立度判定基準、要介護度、低栄養の有無、摂食・嚥下機能障害の有無、経管・経静脈栄養の状況、入院後48時間以内の栄養アセスメントの実施、転倒・転落回数、インシデント影響度分類レベル3b以上の転倒・転落、身体的拘束日数、など

(iii) 診断名

主傷病名、入院の契機となった病名、医療資源を最も消費した病名、医療資源を2番目に消費した病名、入院時併存症、入院後合併症

(iv) 手術情報

手術日、点数表コード、手術回数、手術側数、麻酔、予防的抗菌薬投与、手術名

6

DPCデータの利点

- 実施された処置・手術、投与された薬剤、在院日数、退院時転帰、入院医療費などのデータは正確
- 「様式1」にいくつかの臨床的指標が含まれる

対象のfocusをしぼってうまくデザインを組めば、かなり有用な研究が可能

9

DPCデータベース研究の限界

- 1. 患者が他の病院に移ると追跡できない。
長期生存の追跡には向いていない。
(5年生存率など)
→NDBと連結すれば追跡可能に
- 2. 検査データが無い。
→電子カルテデータと連結すれば追跡可能に

10

DPC研究におけるリスク調整

- (1) 様式1の臨床指標
NYHA、急性心筋梗塞killip分類、虚血性心疾患CCS分類、
がんTNM分類・がんステージ、A-DROP、Hugh-Johns、modified
Rankin Scale、Japan Coma Scale、Barthel Index、Child-Pughなど
- (2) Charlson Comorbidity Index
病名(ICD-10コード)をベースにした併存症指数

11

- (3) 処方・処置・手術などのデータの活用
 - ・「高血圧」という病名の感度は低い。降圧薬の処方データは感度も特異度も高い。「高血圧」の存在は、病名ではなく処方データで同定する。
 - ・「急性心不全」という病名の感度は低い。カテコラミンの使用量は処方データから分かる。IABP、PCPSなどを要する重症の心不全は、処置データで同定できる。
 - ・「急性呼吸不全」という病名の感度は低い。「気管内挿管」「人工呼吸器管理」といった処置データは感度も特異度も高い。軽度の呼吸不全は同定できないが、気管内挿管や人工呼吸器管理を要する重症の呼吸不全は、処置データで同定できる。
 - ・「腎不全」という病名の感度は低い。腎代替療法を要する重症の腎不全は、処置データで同定できる。
 - ・手術時間のデータは存在しない。しかし麻酔時間のデータは存在する。なぜなら、麻酔の保険点数は30分ごとに加算されるからだ。
 - ・出血量のデータは存在しない。しかし輸液量や輸血量は処方データから分かる。

12

(3) 処方・処置・手術などのデータの活用（つづき）

- ・術後にどの抗生物質をどれだけの容量で何日間やったか、処方データから分かる。
- ・消化管穿孔による腹腔内の汚染の程度は分からない。しかし手術中に腹腔内の洗浄のために使用した生理食塩水の量は、処方データから分かる。
- ・リハビリテーションを毎日何分やったかが分かる。なぜならリハビリの点数は20分ごとに加算されるからだ。
- ・胸腔ドレーンの留置期間は分かる。なぜなら持続吸引を行うドレーン法は1日あたり50点の保険点数がつけられ、ドレーンを抜去するとつかなくなるからだ。
- ・肺血栓塞栓予防のための弾性ストッキングまたは間歇的空気圧迫装置を使用したかどうかは、肺血栓塞栓症予防管理料が算定されているかどうかを見れば分かる。

13

2. DPCデータを用いた研究の実例

【方法】2010年7月－2013年3月、n=6342
肺炎・敗血症性DIC患者6342名
トロンボモジュリン投与群(n=1280)および非投与群(n=5062)
主な統計解析：1:1傾向スコアマッチング

【結果】
28日以内死亡率：投与群37.6%、非投与群37.0%
絶対リスク差0.6% (95%信頼区間 -3.4 to 4.6)

【結論】
投与群・非投与群間で28日以内死亡率に有意差なし

16

2019年、多国籍大規模RCT(n=800)

Effect of a Recombinant Human Soluble Thrombomodulin on Mortality in Patients With Sepsis-Associated Coagulopathy-The SCARLET Randomized Clinical Trial. JAMA. 2019;321(20):1993-2002

【方法】並行群間二重盲検ランダム化比較試験
26か国159施設のICU、2012年10月－2018年3月
循環呼吸不全を伴う敗血症DICの成人
トロンボモジュリン投与群395人
プラセボ投与群405人

【結果】28日以内死亡率
トロンボモジュリン群26.8% vs プラセボ群29.4% (P=0.32),
絶対リスク差2.55% (95%信頼区間, -3.68% ~8.77%).

【結論】
投与群・非投与群間で28日以内死亡率に有意差なし

94

17

【研究例1】

敗血症性DICに対するトロンボモジュリンの効果

Recombinant human soluble thrombomodulin and mortality in severe pneumonia patients with sepsis-associated disseminated intravascular coagulation: an observational nationwide study. J Thrombosis Haemost 2015;13(1):31-40.

【背景】

2007年、トロンボモジュリンの臨床試験(n=232) (J Thromb Haemost. 2007;5(1):31-41.)

対象:造血管悪性腫瘍・感染症を基礎疾患とするDIC患者
⇒heterogeneousな対象集団

エンドポイント：DIC離脱率
トロンボモジュリン群66.1%、ヘパリン群49.9%、リスク差16.2% (3.3%~29.1%)
true endpointではなく、surrogate endpoint
⇒死亡率減少効果は不明

15

【研究例2】

食道がん手術患者におけるステロイド術前予防投与と術後呼吸不全および在院死亡の関連

Impact of prophylactic corticosteroid use on in-hospital mortality and respiratory failure after esophagectomy for esophageal cancer: Nationwide inpatient data study in Japan. Annals of Surgery 2023;277(6):e1247-e1253.

食道がん手術は呼吸不全などの術後合併症のリスクが高い。
手術直前にステロイドを投与することで、術後合併症のリスクを軽減できるかどうかは不明。

17

18

データソース：DPCデータ

期間：2010年7月から2019年3月

N＝35,501

手術日にコルチコステロイドを使用した症例は22,620例（63.7％）

主な統計解析：傾向スコア逆確率重み付け法

コルチコステロイド非使用群に対する使用群のオッズ比

在院死亡：OR 0.80（95％CI 0.69–0.93, P＝0.003）

呼吸不全：OR 0.84（95％CI 0.79–0.90, P<0.001）

重症呼吸不全：OR 0.87（95％CI 0.80–0.95, P＝0.002）

食道がん術前のステロイド投与は術後の呼吸不全や在院死亡の減少と関連する。

19

【研究例3】

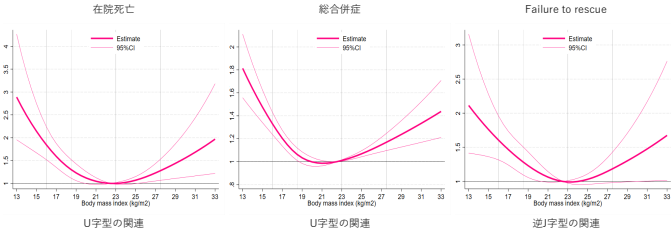
BMIと食道癌術後アウトカムの関連

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. *Annals of Surgery* 2023;277(4):e785-e792.

DPCデータベース2010/7-2019/3

食道癌手術例 (n=39,406)

Restricted cubic spline



低BMIでは、合併症が多いだけでなく、合併症が発生した後の死亡 (failure to rescue) も多かった。

20

21

考察

- BMIは食道癌の重要な術前評価項目となりうる
 - ✓ 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

- Limitationとして以下の情報が不明
 - ✓ 食道癌の組織型
 - ✓ 術前体重減少の有無

22

【研究例4】

90歳以上の心不全患者における急性期リハビリテーション早期開始とアウトカムの関連

Association of early acute-phase rehabilitation initiation on outcomes among patients aged≥90 years with acute heart failure. *Journal of the American Geriatrics Society*. 2023;71:1840-50

Diagnosis Procedure Combinationデータベース（2010年1月～2018年3月）

Patients：心不全で入院した90歳以上、NYHAクラスII以上の患者 (N=41,896)

Exposure：入院後2日以内のリハビリテーション開始あり

Control：入院後2日以内のリハビリテーション開始なし

Outcome：在院死亡、在院日数、30日再入院、ADLの改善

統計解析 1:1傾向スコアマッチング (8587ペア)

23

【研究例5】

重症肺炎患者の人工呼吸器管理を集中治療室（ICU）またはハイケアユニット（HCU）で実施した場合の予後の比較

Intensive care unit versus high-dependency care unit for mechanically ventilated patients with pneumonia: a nationwide comparative effectiveness study. *The Lancet Regional Health - Western Pacific*. 2021;13:100185.

背景

HCUは中等度レベルの医療を比較的低コストで提供できるが、人工呼吸患者はICUとHCUのどちらで治療すべきかについてのエビデンスは限られている。

方法

DPCデータベース 2014年4月-2019年3月

対象：入院当日にICUまたはHCUに入院し人工呼吸器を装着された肺炎患者

主要アウトカム：30日院内死亡率

1：1傾向スコアマッチング、操作変数法

25

結果

N＝14,859、ICU群7,528例（51％）、HCU群7,331例（49％）

傾向スコアマッチング後

30日院内死亡率：ICU群24.0％、HCU群31.2％

（差、-7.2％；95％信頼区間 -10.0％～-4.4％）

操作変数でも同様の結果

結論

人工呼吸管理を要する肺炎患者に対するICUでのクリティカルケアは、HCUでのケアと比較して、30日院内死亡率の7.2％低下と関連していた。

26

【研究例6】

ネオアジュバント化学療法の併用有無による乳癌術後短期成績の比較

Short-term outcomes following breast cancer surgery with and without neoadjuvant chemotherapy: a nationwide administrative database study in Japan. *Annals of Surgical Oncology*. 2022;29:3000-3010.

背景

乳癌に対するネオアジュバント化学療法（NAC）⇒短期手術成績への影響、化学療法と手術の間隔は不明。

方法

期間：DPCデータベース 2010年7月～2017年3月

対象：0～III期乳がんの手術を受けた患者

NACあり11,722例、NACなし120,538例

年齢、施設、入院年度をマッチさせた1:4マッチドペアコホート

アウトカム：術後合併症、麻酔時間、総入院費用

感度分析：NAC投与から手術までの間隔が短い患者、特定のNACレジメンを受けた患者、特定の外科手術を受けた患者

27

結果

術後合併症の発生率 6.0%
NACから手術までの期間の中央値は31日（24～39日）
合併症発生率は両群で有意差なし
（オッズ比、0.95；95%信頼区間、0.88-1.04）
NACは、麻酔期間の短縮および入院総費用の減少と有意に関連していた。感度分析でも同様の結果が示された。

結論

乳がんに対するNACの有無にかかわらず、術後合併症に有意差は認められなかった。患者は、間隔を長くしなくても安全に手術とNACを受けることができる。

28

【研究例8】

重症熱性血小板減少症候群に対する抗菌薬の有効性

(Kutsuna S, et al. Exploring the Efficacy of Routine Antimicrobial Therapy in Severe Fever with Thrombocytopenia Syndrome: overlap weighting analysis using a nationwide inpatient database. Journal of Infection and Chemotherapy. 2024 in press)

【背景】

重症熱性血小板減少症候群（Severe fever with thrombocytopenia syndrome, SFTS）は、確立された治療戦略がない。
本研究は、日本におけるSFTSと診断された患者に対するルーチンの抗菌薬療法の効果の評価することを目的とした。
【方法】DPCデータベース 2010年7月～2021年3月
対象：成人SFTS患者
曝露：入院初期2日以内の抗菌薬治療あり
対照：入院初期2日以内の抗菌薬治療なし
アウトカム：院内死亡率、入院、入院期間
統計解析 傾向スコア・オーバーラップ重み付け

31

【研究例7】

重症熱傷患者に対するアルブミン輸液と死亡率の関連

Nakamura K, et al. Effect of fluid resuscitation with albumin on mortality in patients with severe burns: A nationwide inpatient data analysis. Burns. 2024 in press

【目的】

重症熱傷患者の死亡率に対するアルブミン投与の効果を評価する。

【方法】

DPCデータ2014年4月から2021年3月まで
対象：重症熱傷（熱傷指数≧15）で入院した15歳以上の患者
曝露：入院後2日以内のアルブミン投与あり
対照：入院後2日以内のアルブミン投与なし
アウトカム：28日死亡率

統計解析： 1:1傾向スコアマッチング

29

【研究例9】

甲状腺クリーゼに対するヨウ化カリウムの効果

(Matsuo Y, et al. Potassium Iodide Use and Patient Outcomes for Thyroid Storm: An Observational Study. Journal of Clinical Endocrinology & Metabolism. 2024 in press)

【目的】

ガイドラインでは、抗甲状腺薬と組み合わせたヨードが甲状腺クリーゼの初期薬物療法として推奨されている。しかし、甲状腺クリーゼの管理におけるヨードの臨床的有効性はまだ未解明である。
【方法】
DPCデータベース 2010年7月から2022年3月まで
対象：甲状腺クリーゼで入院した患者
曝露：入院後2日以内にKIを投与あり（KI群）
対照：入院後2日以内にKIを投与なし（非KI群）
アウトカム：院内死亡率など
統計解析 一般化推定方程式によるロジスティック回帰
欠損値には多重代入法を適用

30

【研究例10】

胃十二指腸潰瘍穿孔の術後死亡率を
予測する機械学習スコアリングシステム

New machine learning scoring system for predicting postoperative mortality in gastroduodenal ulcer perforation. *Surgery*. 2022;171(4):1036-42.

背景

胃十二指腸潰瘍穿孔の術後死亡リスクを推定する従来の予測モデルは、予測性能が不十分であった。
術後死亡率を予測するための新しい機械学習モデルと、それを基にした整数スコアを開発し、検証することを目的とした。

方法

データソース：DPCデータベース
対象：外科的修復を受けた胃十二指腸潰瘍穿孔の患者

derivation cohort：2010年7月から2016年3月までのコホート
パラメータ：45の予測因子
モデル：
✓Lasso回帰
✓Lasso回帰で得られた重要変数に基づく整数スコア
✓Extreme Gradient Boosting (XGBoost)

validation cohort：2016年4月から2017年3月までのコホート
上記のモデルと、従来の米国麻酔科学会（ASA）リスクスコアの間でC統計量を比較。

33

35

34

36

結果

25,886人の対象患者のうち、術後死亡は1,176人（4.5%）
Lasso回帰では45因子のうち、19因子が選択された。
19因子のodds ratioに基づき、1-4の整数スコアを割り当てた。
重要な13の因子、特に重要な6因子を用いた整数スコアモデルを構築した。

37

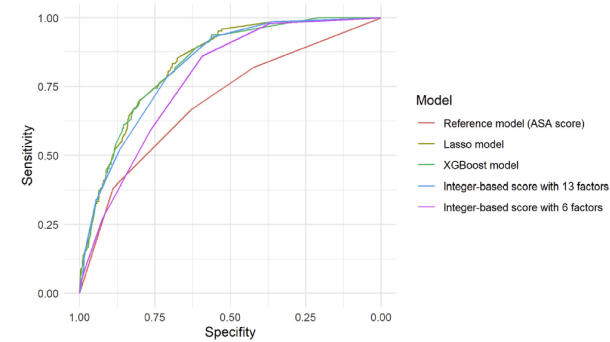
Predictors identified by Lasso regression and their assigned weights for the integer score

Identified predictors	OR	95% CI	Assigned weights
Age of 60–69 years	3.0	2.2–4.1	1
Chronic renal failure	3.0	2.1–4.1	
Disseminated intravascular coagulation	2.9	2.0–4.2	
Metastasis	2.5	1.7–3.7	2
Sepsis	2.3	1.8–2.9	
Age of 70–79 years	5.2	3.8–7.0	
Acute renal failure	7.4	5.3–10.3	
Body mass index ≥40 kg/m ²	7.4	1.7–28.6	
Bowel ischemia	6.9	1.7–23.9	3
Coma	5.9	4.3–8.2	
Malignant solid tumor*	4.2	3.3–5.2	
III or IV in NYHA classification	11.7	3.4–37.5	
Age of 80–89 years	9.7	7.3–13.0	4
Malignant lymphoma	10.8	5.3–21.0	
Severe liver disease	9.6	4.4–20.0	
Age of ≥90 years old	18.7	13.3–26.5	

38

整数スコア	死亡率（%）
0	0.1
1-2	2.3
3-4	9.3
≥5	29.0

39



40

LassoとXGBoostはASAスコアよりも有意に高い予測性能を示した

13因子による整数スコアは、LassoおよびXGBoostと同程度の予測能力を有していた（C統計量0.83；95%信頼区間0.81-0.86）。

41

ご清聴ありがとうございました

42

DeSCデータ研究のデザインと実践

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
笹刈 裕介

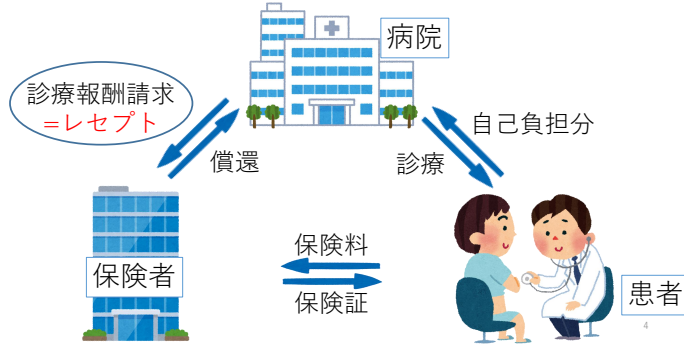
Agenda

1. DeSCデータベースとは
2. DeSCデータベースを用いた研究

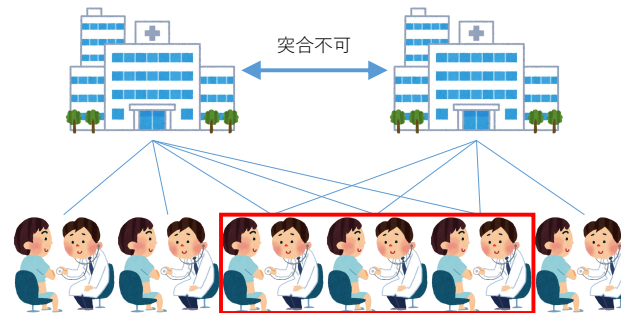
保険データの種類

- ▶医療機関由来（DPCデータ, MID-NET等）
受診した患者のレセプトデータ
様式1 情報
その他医療機関で得られる情報（検査値等）
- ▶保険者由来（レセプトデータ）
加入者のレセプトデータ（医科・DPC・調剤・歯科）
健診情報

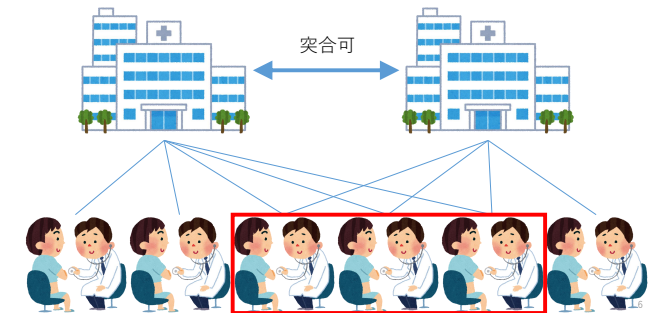
レセプトデータ



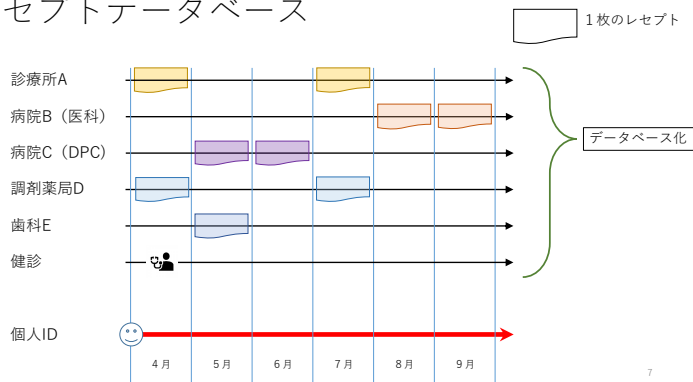
医療機関から収集



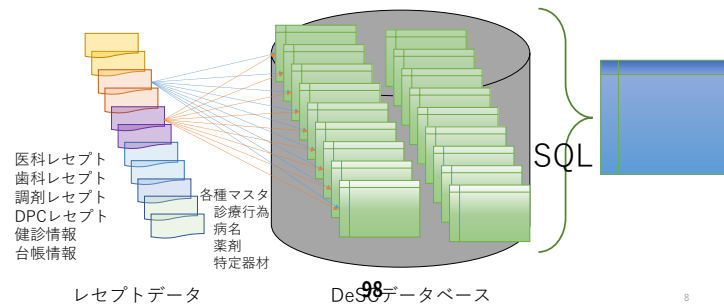
保険者から収集



レセプトデータベース



DeSCデータベース



DeSCデータベース

保険請求データ（2014~2021年, 1200万人）
国民健康保険
健康保険組合
後期高齢者医療制度
健康診断データ
≥65歳（約161万人）・≥75歳（約66万人）
加入者数1200万人
匿名加工情報
国保・後期では一部台帳情報あり
死亡を100%追跡可能

データ項目

診断

国際疾病分類第10改訂（ICD-10）コード
処方薬

解剖学的治療化学分類（ATC分類）

検査・処置・手術・麻酔・特定器材
レセプト電算コード

健診

特定健診・高齢者健診

費用

後期高齢者健診
質問票

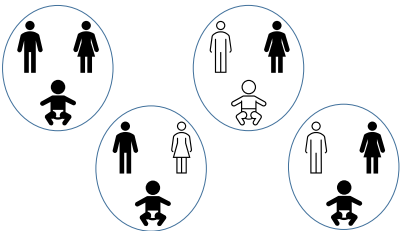
＜質問票の内容について＞			
類型名	No	質問文	回答
健康状態	1	あなたの現在の健康状態はいかがですか	①よい ②まあよい ③ふつう ④まあよい ⑤よくない
	2	毎日の生活に満足していますか	①満足 ②やや満足 ③やや不満 ④不満
食習慣	3	1日3食きちんと食べていますか	①はい ②いいえ
	4	半年前に比べて強いものが食べにくくなったか ※ささいな、たみなど	①はい ②いいえ
口腔機能	5	お茶や汁物等でむせることがありますか	①はい ②いいえ
	6	6か月間で2〜3kg以上の体重減少がありましたか	①はい ②いいえ
運動・転倒	7	以前に比べて歩く速度が遅くなってきたと思いませんか	①はい ②いいえ
	8	この1年間に転んだことがありますか	①はい ②いいえ
認知機能	9	ウォーキング等の運動を週に1回以上していますか	①はい ②いいえ
	10	周知の人から1つを同じことを聞くなどの 物忘れがあると思いませんか	①はい ②いいえ
視覚	11	今日が何月何日かわからない時がありますか	①はい ②いいえ
	12	あなたはたばこを吸いますか	①吸っている ②吸っていない ③やめた
社会参加	13	週に1回以上は外出していますか	①はい ②いいえ
	14	ふだんから家族や友人と付き合いがありますか	①はい ②いいえ
ソーシャルサポート	15	体調が悪いときに、身近に相談できる人がいますか	①はい ②いいえ

特定健診
質問票

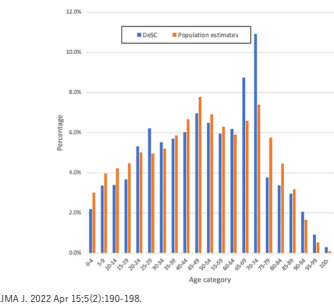
質問項目		回答
1-3	現在、aからcの薬の使用の有無 ^(*)	
1	a. 血圧を下げる薬	①はい ②いいえ
2	b. インスリン注射又は血糖を下げる薬	①はい ②いいえ
3	c. コレステロール ^(*) を下げる薬	①はい ②いいえ
4	糖尿病、脂質異常症、高血圧、脳梗塞等にかかっているといわれたり、治療を受けたことがありますか	①はい ②いいえ
5	糖尿病、心臓病（狭心症、心筋梗塞等）にかかっているといわれたり、治療を受けたことがありますか	①はい ②いいえ
6	医師から、飲性の腎不全にかかっているといわれたり、治療（人工透析）を受けたことがありますか	①はい ②いいえ
7	医師から、貧血といわれたことがありますか	①はい ②いいえ
8	現在、たばこを習慣的に吸っている ^(*)	①はい ②いいえ
9	1日1回以上の喫煙量（平均喫煙量）は、1年未満（平均喫煙量）は、1年未満	①はい ②いいえ
10	1日1回以上の喫煙量（平均喫煙量）は、1年未満（平均喫煙量）は、1年未満	①はい ②いいえ
11	日常生活において歩行又は同等の身体活動を1日1時間以上実施	①はい ②いいえ
12	1日1回以上の歩行又は同等の身体活動を1日1時間以上実施	①はい ②いいえ
13	この1年間で歩行又は同等の身体活動を1日1時間以上実施	①はい ②いいえ
14	1日1回以上の歩行又は同等の身体活動を1日1時間以上実施	①はい ②いいえ
15	歩行又は同等の身体活動を1日1時間以上実施	①はい ②いいえ
16	夕食後に喫煙（食後の喫煙）量と同等の量に3回以上ある	①はい ②いいえ
17	夕食後に喫煙（食後の喫煙）量と同等の量に3回以上ある	①はい ②いいえ
18	喫煙量を喫煙量が1日1回以上ある	①はい ②いいえ
19	喫煙量を喫煙量が1日1回以上ある	①はい ②いいえ
20	喫煙量を喫煙量が1日1回以上ある	①はい ②いいえ
21	運動や食生活等の生活習慣を改善してみようと思いませんか	①はい ②いいえ
22	生活習慣の改善について保健指導を受ける機会があれば、利用しますか	①はい ②いいえ

家族ID, 親子ID

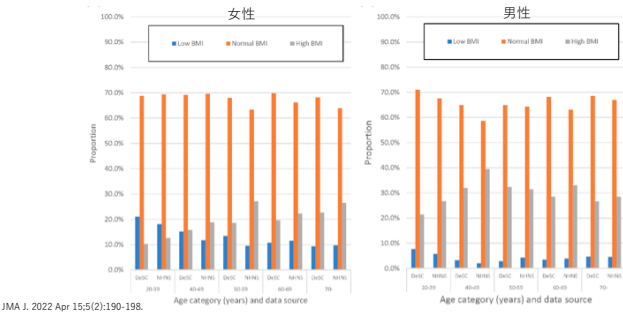
- 家族、親子を特定可能



DeSCデータベースの年齢構成



DeSCデータベースのBMIの分布



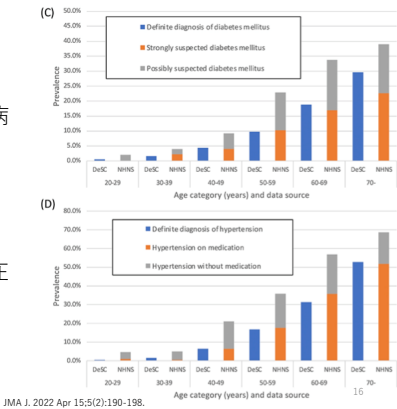
有病割合

DeSCデータベース
国民健康栄養調査

比較

糖尿病

高血圧



研究目的別に利用可能な医療情報 DB の例示

研究目的	研究タイプ・分析方法等	活用可能な医療情報 DB
有病率調査 処方パターンの実態把握	記述研究・横断研究など 記述研究	NDB/NHWS など NDB/JMDC Claims database/ MDV データ/DeSC データベース /MID-NET/RWD データベースなど
薬剤の有効性評価	コホート研究	NDB/JMDC Claims database/ MDV データ/DeSC データベース /MID-NET/RWD データベースなど
薬剤の安全性評価	症例対照研究／コホート研究／ Sequence Symmetry Analysis／ Self-Controlled Case Series など	NDB/JMDC Claims database/ MDV データ/DeSC データベース /MID-NET/RWD データベースなど

他部門での医療情報 DB の活用例

部門	目的	活用可能な医療情報 DB
PV 部門	製造販売後データベース調査など	NDB/JMDC Claims database/MDV データ/DeSC データベース/MID-NET/RWD データベースなど
Marketing 部門	処方実態把握/競合品の分析/患者背景分析など	JMDC Claims database/MDV データ/DeSC データベース/RWD データベース/NHWS など
R&D 部門	処方実態把握/患者背景分析など	JMDC Claims database/MDV データ/DeSC データベース/RWD データベース/NHWS など
HTA 対応部門	費用分析/競合品の分析など	JMDC Claims database/MDV データ/DeSC データベースなど

Agenda

- 1. DeSCデータベースとは
- 2. DeSCデータベースを用いた研究

診療実態調査

Postoperative adverse events and re-treatment among patients who have undergone laparoscopic and robotic sacrocolpopexy for pelvic organ prolapse in Japan

Int J Gynecol Obstet. 2023;161:114–119.

背景

- 骨盤臓器脱（POP）は女性に比較的良好に見られる
 - 腹腔鏡下／ロボット支援下仙骨腔固定術
- 目的
- POP患者に対する腹腔鏡下およびロボット支援下仙骨腔固定
 - 有害事象と再発に対する再治療を記述・比較

研究デザイン

Data：DeSCデータベース
2014年4月－2021年5月

P 骨盤臓器脱の診断を受けた女性患者
E 腹腔鏡下仙骨腔固定術
C ロボット支援下仙骨腔固定術
O 膣びらん/術後尿失禁/術後排尿困難
尿路傷害/腹部切開ヘルニアの複合アウトカム

結果

• POPと診断された女性患者28,748人
腹腔鏡下仙骨腔固定術：409人（1.4%）
ロボット支援下仙骨腔固定術：52人（0.2%）

	腹腔鏡 n = 409	ロボット n = 52	Pvalue
アウトカム			
術後合併症	85 (20.8)	7 (13.5)	0.270
再手術	0 (0)	1 (1.9)	0.113

Recent trends in anti-vascular endothelial growth factor intravitreal injections: a large claims database study in Japan

Jpn J Ophthalmol. 2023 Jan;67(1):109-118.

背景

- 抗血管内皮増殖因子（VEGF）の硝子体内注射
加齢黄斑変性症・網膜静脈閉塞症・糖尿病黄斑浮腫
近視性脈絡膜新生血管・血管新生緑内障・未熟児網膜症 等
海外では使用動向が報告されている
- 高齢者では加齢黄斑変性症や網膜静脈閉塞などの有病率が高い
日本で疾患と抗VEGF薬の使用傾向が異なる可能性

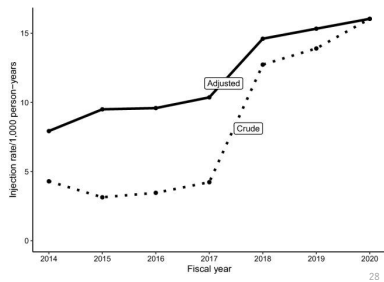
目的：日本における抗VEGF薬の使用動向を記述する

方法

- DeSCデータベース
- 2014年4月－2021年3月
- 対象疾患：加齢黄斑変性症・網膜静脈閉塞症・糖尿病黄斑浮腫
近視性脈絡膜新生血管・血管新生緑内障・未熟児網膜症
- 抗VEGF薬注射回数と観察期間から注射率（/1000人年）を算出
- 性別・年齢・疾患ごとに層別化

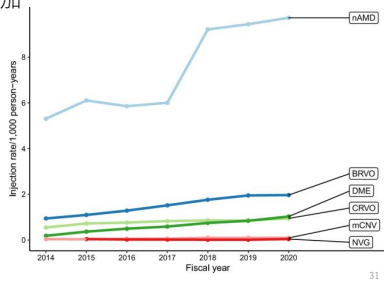
結果

- 抗VEGF薬投与率
- 2014年から2020年
約2倍に増加



結果

- 加齢黄斑変性で著明に増加

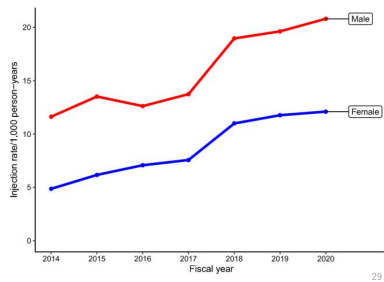


方法

- DeSCデータベース
- 2014年4月ー2021年3月
- 対象疾患：緑内障（40歳-99歳）
- 点眼薬投与量・手術頻度(/1000人年)を算出
- 費用の集計

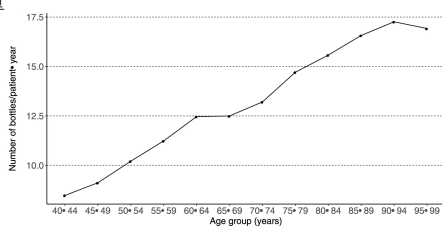
結果

- 男性で高い注射率



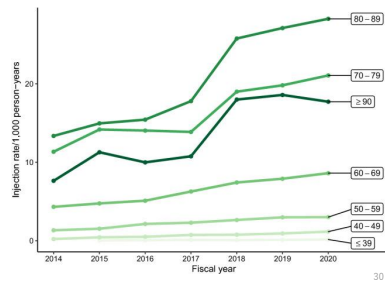
結果

- 点眼薬投与量



結果

- 高齢者で高い注射率



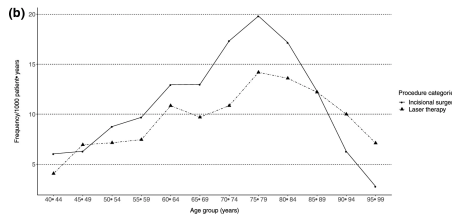
背景

緑内障
失明の主な原因
世界的に増加
治療：点眼薬、手術

治療にかかる費用
アメリカ：623-2511\$/年
イギリス：405£/年

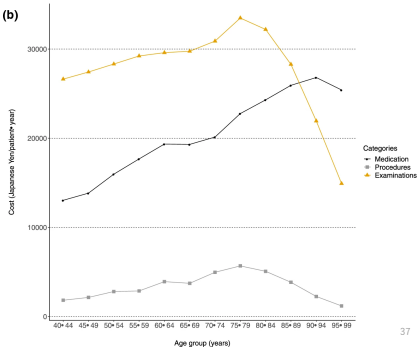
結果

- 手術頻度 (/1000人年)



結果

- 費用



コホート研究

Risk of post-extraction bleeding with direct oral anticoagulant compared with warfarin: Retrospective cohort study using large scale claims data in Japan

Thrombosis Research. 2023;222:24–30

背景

- 拔牙後の出血
 - 歯科診療でよくみられる合併症
 - 特に抗凝固薬を使用している患者
- ワーファリン内服中であっても拔牙は安全に実施可能
- DOACを服用患者における拔牙後の出血のリスクは不明
- 目的
 - ワーファリンとDOACの拔牙後出血リスクを比較

方法

- DeSCデータベース
- 2015年4月ー2020年11月
- 対象疾患：拔牙
- 曝露／対照：DOAC／ワーファリン
- アウトカム：7日以内の処置を要する拔牙後出血
当日の再受診・血栓塞栓症

結果

	Warfarin (n = 927.5)	DOACs (n = 927.5)	Adjusted OR (95 % CI)
Post-extraction bleeding	2.2%	1.9%	0.84 (0.54, 1.31)
Revisit	1.4%	1.2%	0.80 (0.46, 1.42)
Thromboembolism	0	0	—

Proteinuria screening and risk of bone fracture:a retrospective cohort study using a nationwide population-based database

Clin Kidney J. 2023;17:sfad302

背景

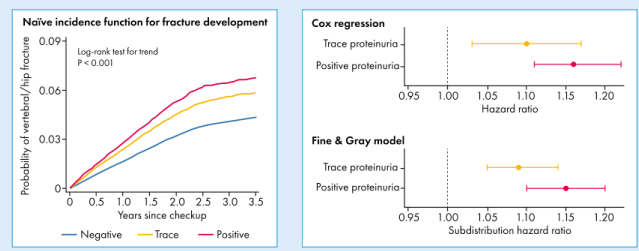
- 蛋白尿は慢性腎臓病（CKD）の主要な要素
心血管疾患や腎機能の悪化
がんや貧血など他の疾患の発症リスク
- CKD
 - 骨折のリスク因子
- 蛋白尿と骨折の関連
 - 一貫していない→本研究で検討

方法

Data：DeSCデータベース
2014年8月ー2021年2月

Participant	健診を受けた加入者
Predictor	尿蛋白定性検査(なし・弱陽性・陽性)
Outcome	股関節骨折/椎体骨折

結果



46

質問票の利用

Association between subjective physical function and occurrence of new fractures in older adults: A retrospective cohort study

Geriatr Gerontol Int 2024;24: 337–343.

47

48

背景

- 高年齢者は転倒リスク大
- 欧米
 - 12の質問を用いた転倒リスク評価
- 日本
 - 後期高齢者質問票→3つの質問で転倒リスク評価
 - 3つの質問と転倒リスクの関連は不明

49

方法

Data : DeSCデータベース
2014年4月－2021年8月

P 後期高齢者健診を受けた75歳以上の高齢者
E/C 歩行速度・転倒歴・運動習慣
O 骨折

高齢者健診

<質問票の内容について>			
質問票名	No	質問文	回答
健康状態	1	あなたの現在の健康状態はよいですか	①よい ②まあよい ③ふつふつ ④あまりよくない ⑤よくない
	2	毎日の生活に満足していますか	①満足 ②やや満足 ③やや不満 ④不満
食習慣	3	1日3食きちんと食べていますか	①はい ②いいえ
口腔機能	4	半年前に比べて固いものが食べにくくなりましたか	①はい ②いいえ
	5	米を噛み砕く、たろあんなど	①はい ②いいえ
体重変化	6	お茶や汁物等でむせることがありますか	①はい ②いいえ
	7	6か月間で2〜3kg以上の体重減少がありましたか	①はい ②いいえ
運動・転倒	8	以前に比べて歩く速度が遅くなったと思いますか	①はい ②いいえ
	9	この1年間に転んだことがありますか	①はい ②いいえ
	10	ウォーキング等の運動を週に1回以上していますか	①はい ②いいえ
認知機能	11	周りの人から「いつも同じことを聞く」などの物忘れがあると思われていますか	①はい ②いいえ
	12	今日が何月何日かわからない場合がありますか	①はい ②いいえ
視覚	13	あなたははたばたを履いていますか	①履いている ②履っていない ③やめた
	14	週に1回以上は外出していますか	①はい ②いいえ
社会参加	15	ふだんから家族や友人と付き合いがありますか	①はい ②いいえ
	16	体調が悪いときに、身近に相談できる人がいますか	①はい ②いいえ

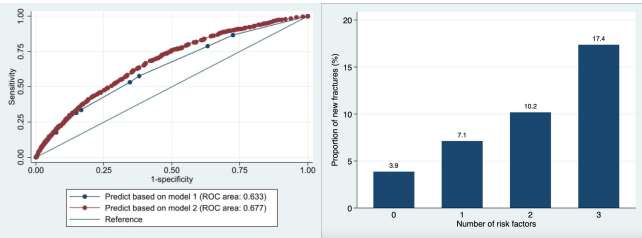
50

結果

	Model 1			Model 2		
	HR	95% CI	P-value	HR	95% CI	P-value
Physical function and falls section of the Late-Stage Elderly Questionnaire						
Decline in subjective gait speed	1.63	1.40–1.89	<0.001	1.50	1.29–1.74	<0.001
History of falls within the past year	2.03	1.77–2.33	<0.001	1.96	1.71–2.24	<0.001
Absence of exercise habits	1.29	1.13–1.47	<0.001	1.25	1.09–1.42	0.001
Age	NA			1.04	1.03–1.06	<0.001
Male	NA			0.48	0.42–0.56	<0.001

52

結果



103

53

Association between self-reported difficulty in chewing or swallowing and frailty in older adults: A retrospective cohort study

Geroscience. 2024. doi: 10.1007/s11357-024-01325-7.

54

背景

- 口腔虚弱
加齢に伴う口腔内の健康状態の変化に起因する脆弱性
食事機能の低下や身体的・精神的な障害を引き起こす
- 後期高齢者質問票
咀嚼と嚥下の困難に関する2つの項目
口腔虚弱をスクリーニング可能かどうか不明

方法

Data：DeSCデータベース
2014年4月－2022年11月

P 後期高齢者健診を受けた75歳以上の高齢者
E/C 自己申告による咀嚼・嚥下機能
O 主要評価項目：誤嚥性肺炎による入院
副次評価項目：全入院／全死亡

高齢者健診

<質問票の内容について>			
類型名	No	質問文	回答
健康状態	1	あなたの現在の健康状態はいかがですか	①よい ②まあよい ③ふつふ ④まあよくない ⑤よくない
	2	毎日の生活に満足していますか	①満足 ②やや満足 ③やや不満 ④不満
食習慣	3	1日3食きちんと食べていますか	①はい ②いいえ
口腔機能	4	半年前と比べて飲み込むのが食べにくくなりましたか	①はい ②いいえ
	5	お茶や汁物等でむせることがありますか	①はい ②いいえ
体重変化	6	6か月間で2～3kg以上の体重減少がありましたか	①はい ②いいえ
運動・転倒	7	以前に比べて歩く速度が遅くなったと感じますか	①はい ②いいえ
	8	この1年間に転んだことがありますか	①はい ②いいえ
	9	ウォーキング等の運動を週に1回以上していますか	①はい ②いいえ
認知機能	10	周りの人からいつも同じことを聞かれますか	①はい ②いいえ
	11	今日が何月何日かわからない場合がありますか	①はい ②いいえ
喫煙	12	あなたはたばこを吸いますか	①吸っている ②吸っていない ③やめた
社会参加	13	週に1回以上は外出していますか	①はい ②いいえ
	14	ふだんから家族や友人と付き合いがありますか	①はい ②いいえ
ソーシャルサポート	15	体調が悪いときに、身近に相談できる人がいますか	①はい ②いいえ7

結果

口腔機能の回答あり	359,111	
口腔機能低下なし	219,092	61.0%
咀嚼機能低下	67,367	18.8%
嚥下機能低下	42,011	11.7%
両機能低下	30,641	8.5%

結果

	ハザード比	95% 信頼区間	p値
誤嚥性肺炎による入院			
咀嚼機能低下	1.35	1.15–1.58	<0.001
嚥下機能低下	1.1	0.89–1.36	0.39
両機能低下	1.81	1.51–2.16	<0.001
全入院			
咀嚼機能低下	1.08	1.05–1.11	<0.001
嚥下機能低下	1.01	0.97–1.04	0.678
両機能低下	1.12	1.09–1.16	<0.001
全死亡			
咀嚼機能低下	1.28	1.14–1.44	<0.001
嚥下機能低下	1.02	0.87–1.20	0.777
両機能低下	1.75	1.53–1.99	<0.001

Association between COVID-19 pandemic and mental disorders in spouses of intensive care unit patients

Intensive Care Med. 2023;49:112–114.

背景

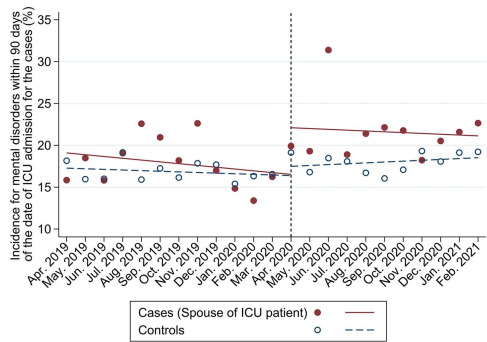
- COVID-19パンデミック
ICU患者の面会機会減少
患者家族の精神状態に影響を与える可能性
- 目的
COVID-19パンデミックがICU患者の配偶者の精神状態に及ぼす影響を検討すること

方法

- DeSCデータベース
- 2019年1月－2021年2月
- 分割時系列デザイン+マッチドペアコホート
2020年4月1日以前/以降
時点マッチング
- 対象疾患：結婚している夫婦
- 曝露／対照：配偶者のICU入室患者の有無
- アウトカム：ICU入室90日以内の精神疾患

家族IDの利用

結果



検診結果の利用

BP classification using the 2017 ACC/AHA BP guidelines with risk of cardiovascular events in older individuals

J Cardiol. 2024. Online ahead of print.

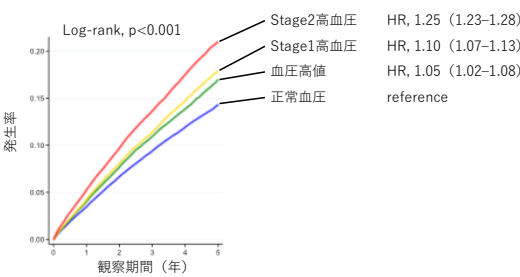
背景

- Stage 1, 2高血圧は心血管疾患のリスク
血圧高値 : sBP $\geq$ 120 & dBP<80 mmHg
Stage 1 高血圧 : $\geq$ 130/80 mmHg
Stage 2 高血圧 : $\geq$ 140/90 mmHg
- 目的
高齢者において血圧高値、Stage 1高血圧・血圧高値が心血管疾患の発症と関連するかを検討

方法

- DeSCデータベース
- 2019年1月ー2021年8月
- コホート研究
健診を受けた時点から観察開始
- 対象疾患 : 健診受診した $\geq$ 60歳の高齢者
- 曝露／対照 : 正常・血圧高値・Stage 1高血圧・Stage 2高血圧
- アウトカム : 心血管疾患 (心筋梗塞・狭心症・脳卒中・心不全)

結果



ご清聴ありがとうございました

JMDCデータ研究のデザインと実践

東京大学大学院医学系研究科イートロス医学

大野 幸子

目次

1. JMDCデータの概要
2. 解析で注意する点
3. 研究紹介

目次

1. JMDCデータの概要
2. 解析で注意する点
3. 研究紹介

JMDCデータ

4

JMDCデータ

5

JMDCデータ

6

株式会社JMDCが提供する、 保険者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）
- 被保険者: 約1400万人（2022年2月時点）
- 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)

株式会社JMDCが提供する、 病院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: 病院（主に大病院）
- 収集期間: 2005-現在（毎月更新）
- 被保険者: 約1400万人（2022年2月時点）
- 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
- 保険者データに加え検査値を含む病院データ提供開始

JMDCの特徴

- 20-65 歳のデータが最も多い（男＞女）
- 子供のデータもある
- 家族IDがある
- 65歳以上のデータはほとんどない
- 数年の経過を追跡可能（平均3.5年）
- 保険から離脱しない限り複数の医療機関をまたいで追跡可能

JMDCデータ

7

他データベースとの比較

8

JMDC vs. 他のRWD

9

健診データを含む

- 問診
- 身長、体重、BMI、腹囲（内臓脂肪面積）
- 血圧
- 脂質異常検査（中性脂肪、HDL-C、LDL-C）
- 肝機能検査（AST、ALT、 γ GTP）
- 血糖検査（空腹時血糖またはHbA1c）
- 尿検査（尿糖、尿蛋白）
- 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。

	NDB	DPC	JMDC	MID-NET	電子カルテ
N	◎	◎	○	△	X
外来	○	X	○	△	△
調剤薬局	○	X	○	X	X
病院間追跡	○	X	○	X	X
重症度	X	○	X	X	○
検査値	X	X	△	○	○
家族ID	X	X	○	X	X
地域ID	○	○	X	○	○
病院ID	△	○	△	○	○
高齢者	○	○	X	○	○
保険者台帳	X	X	106	X	X

JMDC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（重症度指標がない）
- 通常診療の検査値がない
- 高齢者が少ない
- 社会経済的状態（SES）の高い集団に偏っている

目次

1. JMDCデータの概要
2. 解析で注意する点
3. 研究紹介

JMDCデータ解析で注意する点

- 1.データハンドリング
- 2.リスク調整
- 3.適応交絡
- 4.欠損値の取り扱い
- 5.Dropout

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	

2 リスク調整

- 処方・処置・手術：保険算定に関連してデータが充実
- 傷病名：妥当性はvalidation研究を参照
- 個々の疾患の重症度や進行度：直接的な指標なし(JMDCが提供するDPCデータには一部存在)
- 検査値：健診時のみ存在 (JMDCが提供する病院データには一部通常診療の検査値が存在)

傷病名の妥当性

- 病名のみ

➢ 高血圧：感度80.7% 特異度95.3%

➢ 糖尿病：感度91.4% 特異度92.5%

➢ 脂質異常：感度48.9% 特異度90.4%
- 病名+薬剤

➢ 高血圧：感度74.5% 特異度98.2%

➢ 糖尿病：感度78.6% 特異度99.6%

➢ 脂質異常：感度34.5% 特異度97.2%

2 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

2 リスク調整 (対応策続き)

- ICDコードや処置コードを用いた重症度指標

● Charlson Comorbidity Index

Quan H, et al. Am J Epidemiol. 2011;173: 676-682.

-ICD-10コードを用いた併存症指数

● Yamana Severity Index

Yamana H, et al. BMC Health Serv Res. 2015;15: 261.

-処置データを用いた重症度指数

● Claims-based frailty index

Nakatsuka K, et al. J Epidemiol. 2024;34: 112-118.

Kim DH, et al. J Gerontol A Biol Sci Med Sci. 2018;73: 980-987.

-ICD-10コードでフレイルを評価するスコア

3 適応交絡

- 統計的対処

● 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)

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

5 Dropout

- JMDCは、以下の場合は追跡不能
 - 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
- 特に問題となるのが病気の重症化による退職
→ censoring weight（打ち切りの重み）などで対処

JMDCデータはどのような研究に向いているか

- RCTの限界
 - Too few（被験者数）
 - Too simple（併存症、併用療法）
 - Too median-aged（高齢者、小児を除外）
 - Too narrow（治療方法を限定）
 - Too brief（投薬・追跡期間）
- 上記を補完するような研究がGood

JMDCデータでの特性を活かす

- 保険者台帳の利用
- 健診データの利用
- 家族IDの利用（妊娠中のXXが子に与える影響, 家族内集積に興味がある疾患）
- 比較的稀な疾患・状態
- 小児の疾患
- 成人の疾患
- 曝露～5年程度でアウトカムが発生する疾患

目次

- JMDCデータの概要
- 解析で注意する点
- 研究紹介

National dataと組み合わせてJMDCを利用した研究

Long-term autoimmune inflammatory rheumatic outcomes of COVID-19 : A binational cohort study: A binational cohort study

- COVID-19が自己免疫性リウマチ疾患(AIRD)の発症リスクを増加させる可能性があることが示唆されているが、既存研究には方法論的な限界があった
 - 対照群が非感染者（health care seeking behaviorが違う、感染者とAIRDのリスクが違う）
 - ワクチンの影響を考慮していなかった
 - その他の交絡調整が不十分

COVID-19感染後に自己免疫性炎症性リウマチ疾患の発症が増加するか

対象者	20歳以上の韓国人10,027,506人と日本人12,218,680人
曝露・対照	COVID-19 vs インフルエンザ、COVID-19 vs 感染なし
アウトカム	自己免疫性炎症性リウマチ疾患 (AIRD) （ICD-10コード）
デザイン	傾向スコアマッチング→生存時間解析

データソース

韓国 (K-COV-Nコホート)

- 健康診断結果
- 外来および入院データ
- 医薬品情報
- 死亡記録
- COVID-19ワクチン接種情報
- COVID-19検査結果
- 世帯収入
- 地域

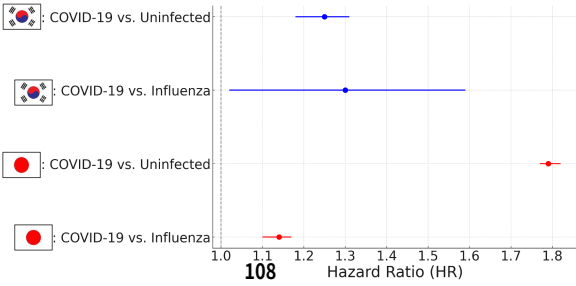


日本 (JMDC)

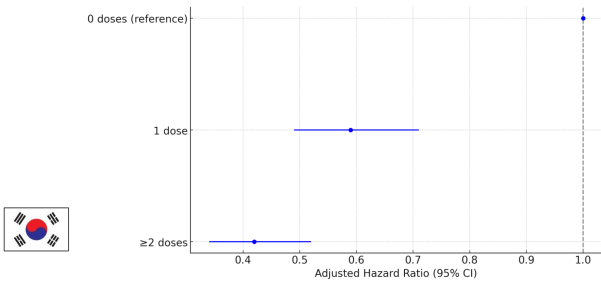
- 健康診断結果（血液検査含む）
- 外来および入院データ
- 医薬品情報
- 死亡記録（資格喪失事由/転帰）
- 保険状況（被扶養者等）



結果 自己免疫性リウマチ疾患発症



結果 ワクチン接種回数別自己免疫性リウマチ疾患発症



結論

- COVID-19感染は、韓国と日本の両方でAIRDsの新規発症リスクを増加
- 韓国と日本での結果は概ね一致
- COVID-19ワクチン接種はAIRDsの発症リスクを低減させる
- 2回以上のワクチン接種が最もAIRDsの発症リスクを低減させる
- COVID-19感染後のAIRDs発症に注意

研究としてのポイント

- RCTが困難なテーマの実施（曝露をランダム化できない、イベント数が必要）
- 2国のデータを利用して一般化可能性を高め五大誌のひとつに掲載
- JMDCの規模は韓国のNational dataに匹敵

歯科レセプト、医科レセプト、検診情報を利用した研究

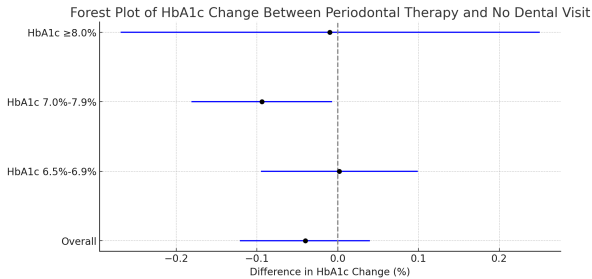
Effect of Periodontal Therapy on Glycaemic Control in Type 2 Diabetes

- 歯周病と糖尿病は双方向の関連性が報告されている
- 歯周病治療は短期的（3～4ヶ月後）にHbA1cを0.29%低下させる
- ベースライン時HbA1cが高いほど歯周病治療の効果が高い可能性があるが、研究は不足

歯周病治療は血糖コントロールを改善するか

対象者	40歳以上の2型糖尿病患者のうち咀嚼に問題あり
曝露・対照	歯周病治療有無
アウトカム	HbA1c変化
デザイン	傾向スコア逆確率重み付け：iptw & ipcw（治療、打ち切り）

結果 歯周病治療あり vs なし（対照）



結論

- 歯周病治療は血糖コントロールが悪い集団で効果がある可能性
- 糖尿病診療には歯科との連携も重要
- HbA1c8%以上の集団は検出力が不足するためさらなる研究が必要
- 研究実施時点での歯科レセプトのvalidationは1件のみ

研究としてのポイント

- RCTが困難なテーマの実施（コントロール不良の人の組入れ）
- 脱落してアウトカムが取れない対象者の存在に対し統計的対処を実施
- 世界でも珍しい歯科レセプトデータを利用して歯科のleading journalに掲載

健診データを利用した研究

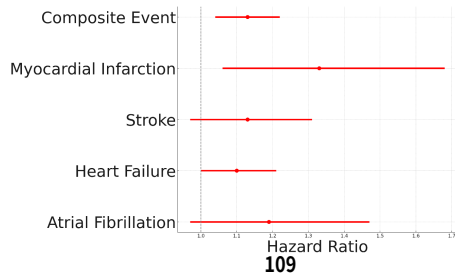
Unawareness of being prescribed medications for diabetes and incident cardiovascular disease

- 糖尿病患者の中には糖尿病治療薬の処方を認識していない患者が存在
- 糖尿病の病識が心血管疾患（CVD）の発生リスクに与える影響を調べた研究は存在しない

糖尿病治療薬の処方認識の欠如は心血管疾患リスクを増加させるか

対象者	糖尿病治療薬を処方されている94,048人の糖尿病患者
曝露・対照	処方認識の有無
アウトカム	心筋梗塞（MI）、脳卒中、心不全（HF）、心房細動（AF）の複合エンドポイント
デザイン	Cox回帰

結果 処方認識なし vs あり（対照）



結果・結論

- 7779/94,048 (8.3 %)に処方認識なし
- 若年でコントロール良好な場合も処方認識なしで心血管疾患のリスク上昇
- 適切な患者説明・教育による意識向上が必要

家族IDを利用した研究

Trends in antihypertensive prescription for pregnant women with hypertension and their peripartum outcomes before and after label and guideline revisions in Japan

- ~2011年まで一部の降圧薬 (labetalol と nifedipine) は添付文書に妊婦への使用が禁忌と記載
- ~2014年まではガイドラインでも禁忌の記載があった
- 添付文書の変更、ガイドラインの変更が降圧薬の処方、周産期アウトカムに与える影響を評価

家族IDを利用した研究

対象者	高血圧の妊婦 (出産関連の保険処置または新生児と家族IDが同一で判定)
曝露・対照	2011年の添付文書変更、2014年のガイドライン変更
アウトカム	降圧薬の処方パターン、周産期アウトカム（早産、帝王切開、緊急帝王切開、HELLP症候群）
デザイン	コクランアーミテージ傾向検定、ロジスティック回帰

結果・結論

- 1,739/13,797（12.6%）が妊娠中に経口降圧剤を使用
- 添付文書およびガイドライン改訂前に最も頻繁に処方された降圧剤はメチルドパであったが、改訂後はニフェジピンが最も頻繁に処方された
- ニフェジピンの処方割合は増加し（ $P < 0.001$ ）、ヒドラルジンの処方割合は減少した（ $P < 0.001$ ）が、メチルドパとラベタロールの処方割合には有意な傾向は見られなかった
- 4つの周産期アウトカム（早産、帝王切開、緊急帝王切開、HELLP症候群）において、有意な差は見られなかった

薬の継続を見た研究

Treatment Persistence Between Long-Acting Injectable Versus Orally Administered Aripiprazole Among Patients with Schizophrenia in a Real-World Clinical Setting in Japan

- 統合失調症は薬物治療の継続が重要である
- 統合失調症治療薬アリピプラゾールには筋注(1回/月) と内服(毎日)等が存在する
- 薬の持続期間の違いにより治療継続が影響を受ける可能性がある

アリピプラゾール筋注は内服より治療中断が起こりにくい

対象者	統合失調症に対してアリピプラゾールを開始した患者
曝露・対照	アリピプラゾール筋注(1回/月) vs. 内服(毎日)
アウトカム	治療継続
デザイン	コホート、Cox回帰

結果・結論

- 198名の筋注群と1240名の内服群を同定した。
- Cox回帰で筋注群は治療中断が起こりにくかった。(HR 0.54; 95%CI 0.43-0.68)
- 統合失調症の薬物治療継続には毎日の内服よりも月一回の筋注が有利である可能性が示唆された

プラクティスパターンとコストの研究

風邪に対する抗菌薬処方ほどの程度医療費を増加させているか
Cost of inappropriate antimicrobial use for upper respiratory infection in Japan.

対象者	抗菌薬投与が不要な上気道炎(ICD-10で定義)
曝露・対照	抗菌薬投与
アウトカム	上記不適切投与による追加医療費
デザイン	コホート研究, 年齢分布から日本全体の追加医療費を推定

- 抗菌薬不適切投与にかかる追加医療費は2013年に424億円、2014年に341億円、2015年に350億円、2016年に297億円
- 追加医療費の9割を第3世代セフェム、マクロライド、フルオロキノロンが占めていた

- JMDCは健診と各種レセプトデータを含む比較的扱いやすく実績豊富な本邦最大規模の医療リアルワールドデータ
- 特に小児～成人の研究に向く
- 重症度や詳細な検査値が必須の研究には向かない
- RCTで実現できないEvidence gapを埋める可能性
- データベースの特性を理解した上での研究立案、データハンドリング、統計解析、結果の解釈が重要

便利な資料

- 診療点数早見表 医学通信社
- しろぼんねっと (<http://shirobon.net>)
- 今日の治療薬 南江堂
- 疾病、傷害および死因統計分類提要ICD-10準拠 厚生省大臣官房統計情報部
- 病名くん (<http://www.byomei.org/byomei-kun.2.0/>)
- 朝井淳 (2017). SQLポケットリファレンス 技術評論社
- 本橋智光 (2018). 前処理大全 技術評論社
- 奥村泰之 (2017). レセプトの落とし穴 Monthly IHEP, 268, 16-25

疫学・統計の応用(1) 傾向スコア分析の応用

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
笹刈 裕介

1

今日の目標

- 傾向スコア の概念を理解する
 - ✓ 傾向スコアの基本的な使い方を理解する
 - ✓ 傾向スコアの利点と限界を理解する
- 傾向スコアを使用した論文を読むことができる
 - ✓ Methodsに書かれていることが理解できる
 - ✓ Resultsを自分で解釈できる
- 傾向スコアを自分で使って分析を行う手順を知る
 - ✓ 自分で傾向スコアを使った分析を行う

2

Agenda

- はじめに
- 傾向スコアとは
- 傾向スコアを使った分析方法
- 傾向スコアの利点と限界
- 傾向スコアを利用した臨床疫学研究
- 高次元傾向スコア

3

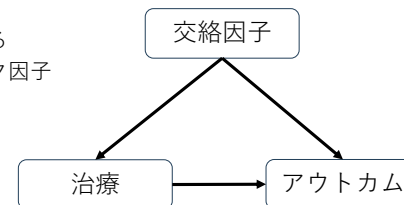
ランダム化比較試験と観察研究

- ランダム化比較試験
治療の割り当ては無作為（ランダム）
患者背景の分布や構成比は似通ってくる
- 観察研究
患者の治療選択に対して介入しない
治療の選択に影響を与える要因（年齢・重症度・検査値など）
患者背景の分布や構成比は異なる

7

交絡因子

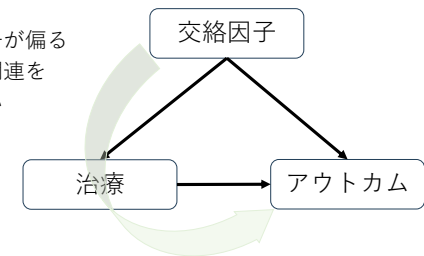
- 交絡因子**
1. 治療選択に影響する
 2. アウトカムのリスク因子
 3. 中間因子でない



8

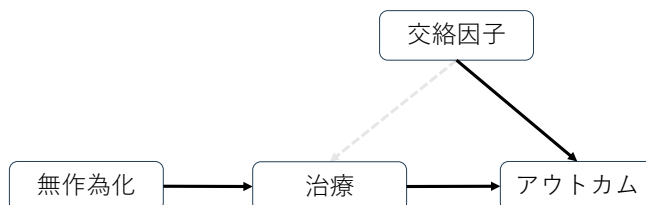
観察研究

- 交絡因子**
- 治療によって交絡因子が偏る
治療とアウトカムの関連を
正しく見積もれない



9

ランダム化比較試験



10

疑似ランダム化

バランスが取れていない観察データ

↓
バランスをとる
Stratification/Matching

交絡因子が増えると

Stratification

- 層が増える

Matching

- 一致する人がいなくなる

→傾向スコア

12

13

傾向スコアに対する過剰な期待

- 傾向スコアはRCTの代わりになる
観察研究でもRCTのように
交絡の影響を取り除くことができる
- 傾向スコアを利用しない研究 = 質が低い
回帰分析は傾向スコアより質が低い

14

傾向スコアに対する過剰な期待

- 傾向スコアはRCTの代わりになる
観察研究でもRCTのように
交絡の影響を取り除くことができる
- 傾向スコアを利用しない研究 = 質が低い
回帰分析は傾向スコアより質が低い

傾向スコアを正しく理解しましょう

15

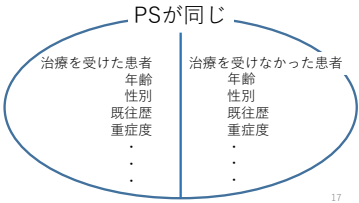
Agenda

- はじめに
- 傾向スコアとは
- 傾向スコアを使った分析方法
- 傾向スコアの利点と限界
- 傾向スコアを利用した臨床疫学研究
- 傾向スコアの発展的利用

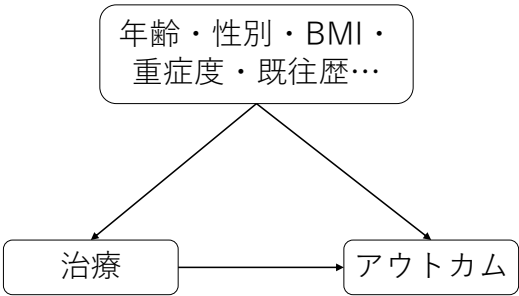
16

バランシングスコア

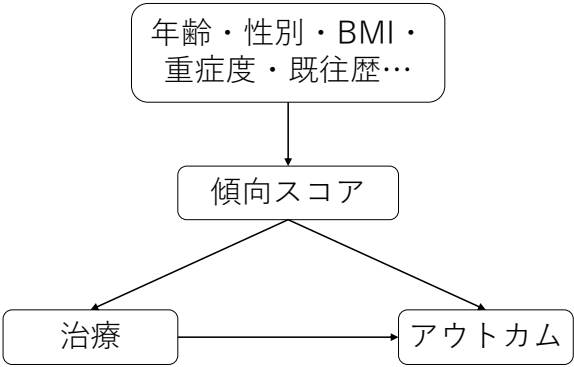
- 傾向スコアとは
- 各患者が治療を受ける傾向をスコア化
 - 治療を受ける確率 (0~1)
 - **Balancing score**



17



18



19

前提条件

未測定の交絡因子が存在しない

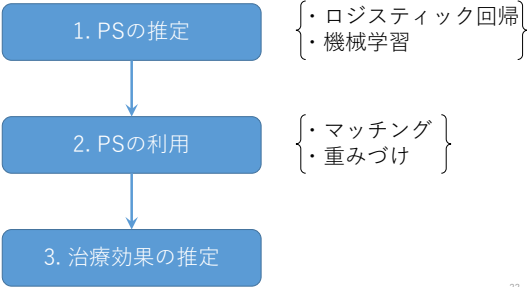
20

Agenda

- はじめに
- 傾向スコア分析理解のための基礎知識
- 傾向スコアとは
- 傾向スコアを使った分析方法
- 傾向スコアの利点と限界
- 傾向スコアを利用した臨床疫学研究
- 高次元傾向スコアの発展的利用

21

傾向スコア (PS) の利用



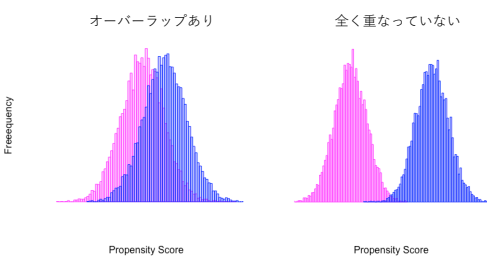
22

傾向スコアの推定

- 真の傾向スコアはわからない
- ベースラインの共変量から推定する
治療を受けるよりも前に決定している要因
- ロジスティック回帰モデル
- 機械学習（ランダムフォレスト/ニューラルネットワーク等）

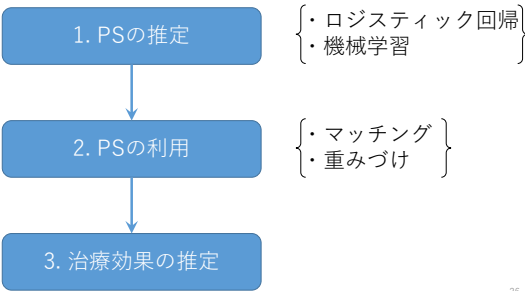
23

PSの重なりが存在（common support）



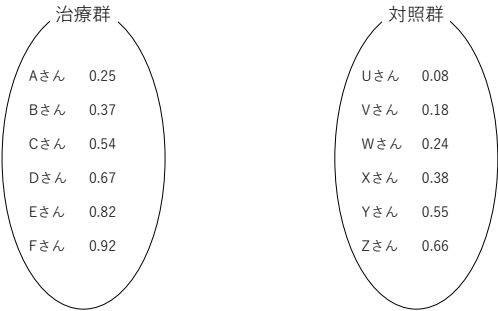
25

傾向スコア(PS) の利用



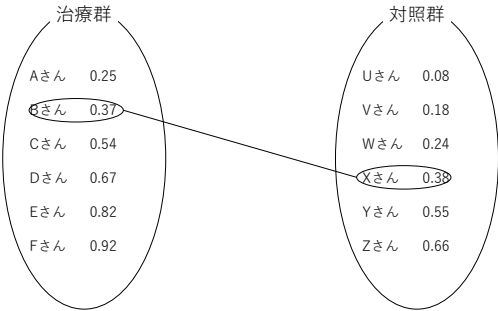
26

マッチング



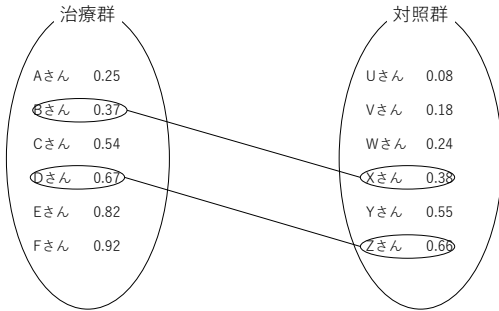
27

マッチング



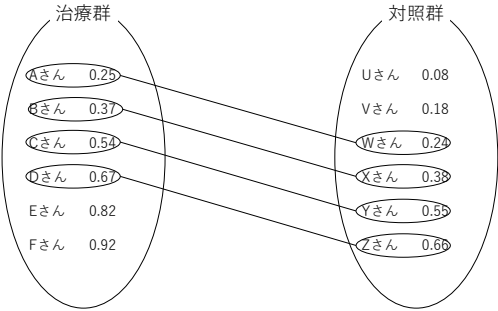
28

マッチング



29

マッチング



30

マッチングの方法

- | | |
|---|-----------------------|
| アルゴリズム | 対照の選択方法 |
| • Nearest neighbor matching without caliper | • without replacement |
| • Nearest neighbor matching with caliper | • with replacement |
| | 対照の人数 |
| | • 1 : n matching |

31

バイアス

	小さい	大きい
caliper	あり	なし
replacement	あり	なし
対照群の人数	1 : 1	1 : 複数

37

バランスの確認

- 治療群と対照群の共変量を比較
Standardized difference
- 治療群と対照群の傾向スコアの分布を比較
ヒストグラムなど

38

Standardized difference

連続変数

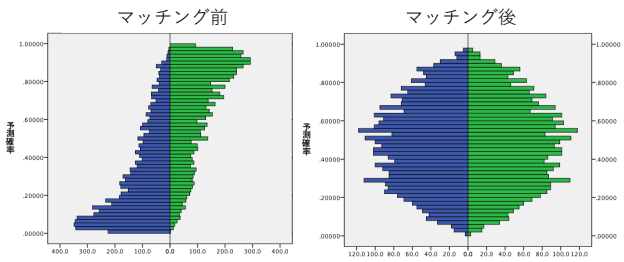
$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

カテゴリー変数

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1 - \hat{p}_{treatment}) + \hat{p}_{control}(1 - \hat{p}_{control})}{2}}}$$

39

傾向スコアの分布



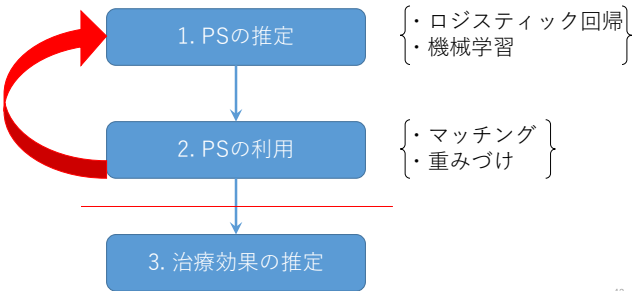
40

バランスが取れていない場合の対処

- 傾向スコアを作成するための変数を増やす
- 傾向スコアを作成するモデルへ投入する因子
連続変数の二乗項、三乗項を投入
連続変数をデータに合わせてカテゴリー化
様々な変数の交互作用項を投入
- 層別化と重み付けやマッチングを組み合わせる
- 傾向スコアマッチング後に回帰分析を行う

41

傾向スコア(PS) の利用



42

治療効果の推定

- 2群のアウトカムを比較
- カテゴリー変数：フィッシャー正確検定
 - 連続変数：t検定やWilcoxon順位和検定

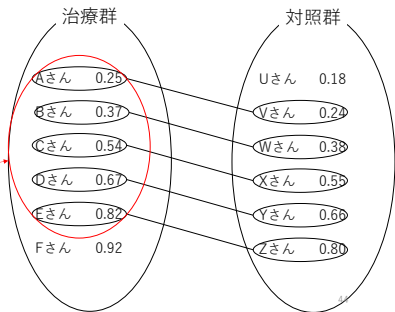
43

治療効果

治療群における処置効果

Average Treatment Effect for Treated (ATT)

治療を受けた人の治療効果



マッチングの利点と欠点

- 利点
- バイアスが少ない

- 欠点
- 解析対象が減る
 - 一般化可能性

重み付け

	治療群	対照群
IPTW	1/PS	1/(1-PS)
Overlap weighting	1-PS	PS

IPTW : Inverse Probability of Treatment Weight

重み付けの例 (IPTW)

治療群	PS	weight	対照群	PS	weight
A	0.1	10	V	0.1	1.1
B	0.3	3.3	W	0.3	1.4
C	0.5	2	X	0.5	2
D	0.7	1.4	Y	0.7	3.3
E	0.9	1.1	Z	0.9	10

47

重み付けの例 (Overlap weighting)

治療群	PS	weight	対照群	PS	weight
A	0.1	0.9	V	0.1	0.1
B	0.3	0.7	W	0.3	0.7
C	0.5	0.5	X	0.5	0.5
D	0.7	0.3	Y	0.7	0.3
E	0.9	0.1	Z	0.9	0.1

48

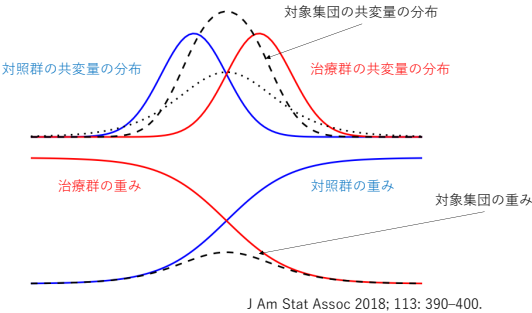
IPTWとOverlap weightingの比較

- IPTW
- Average treatment effect
 - PSが0又は1に近い患者の重みが大きくバイアスの原因
 - トリミングを行うことで多くのサンプルが除外される可能性

- Overlap weighting
- Average treatment effect for the overlap population
 - PSが0又は1に近い患者の重みは小さくなる
 - 正確に共変量のバランスが取れる

49

Overlap weightingにおける対象集団

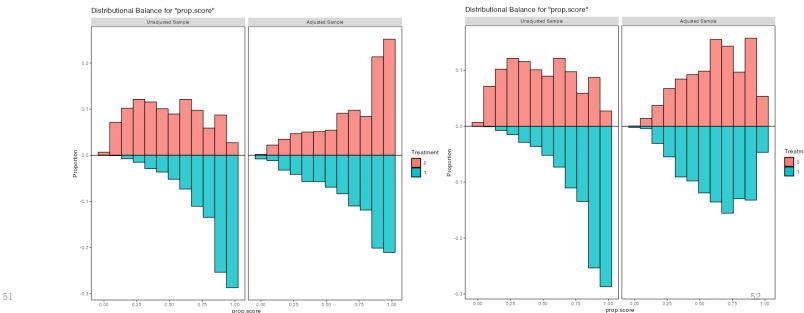


50

バランスの確認

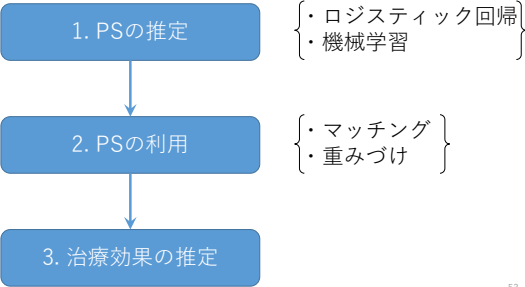
- 重み付けした後の治療群と対照群を比較
Standardized difference
- 治療群と対照群の傾向スコアの分布を比較
ヒストグラムなど

IPTW Overlap weighting



51

傾向スコア (PS) の利用



53

治療効果の推定

重み付けを考慮した回帰分析

Agenda

- はじめに
- 傾向スコア分析理解のための基礎知識
- 傾向スコアとは
- 傾向スコアを使った分析方法
- 傾向スコアの利点と限界
- 傾向スコアを利用した臨床疫学研究
- 傾向スコアの発展的利用

54

55

傾向スコアの利点

① モデルの評価が容易

- 傾向スコアモデル
 - 背景のバランスを評価する
 - 傾向スコアの分布を比較する

• 回帰モデル

治療や背景要因とアウトカムとの関係を正しく指定
評価は非常に困難
 R^2 は治療の有無による系統誤差を排除できない

② デザインと分析を分けることができる

- 傾向スコアモデル
 - RCT同様に治療効果の推定は研究終了後
 - PSの推定とバランシング評価後に治療効果の推定
- 回帰モデル
 - 常にアウトカムが先に見えてしまう
 - モデルの調整中に自分の期待する結果が出たらどうなる？

③ アウトカムが稀な場合にも柔軟に対応可能

- 傾向スコアモデル
 - 一般的によく行われる治療であれば特に問題なし
- 回帰モデル
 - 共変量の数はアウトカム/10以下 (binary outcome)
 - アウトカムが稀な場合、全ての共変量は調整できない

56

57

58

④ 傾向スコアのオーバーラップの程度が明確

- 傾向スコアモデル
 - 傾向スコアのオーバーラップがない
 - 少ないオーバーラップ部分の患者を対象にする
 - 比較可能性のない2群である
- 回帰モデル
 - オーバーラップは確認できない
 - 比較可能性のない2群でも何らかの数値は出てくる

傾向スコア分析の限界

未測定の交絡因子に対しては無力

- PSの計算に利用した変数
 - バランスが取れる
- 未測定 of 交絡因子
 - バランスは不明

59

60

61

傾向スコアマッチングの研究例

Risks and Benefits of Stress Ulcer Prophylaxis for Patients With Severe Sepsis

Crit Care Med. 2016;44:e464-9.

Agenda

- はじめに
- 傾向スコア分析理解のための基礎知識
- 傾向スコアとは
- 傾向スコアを使った分析方法
- 傾向スコアの利点と限界
- 傾向スコアを利用した臨床疫学研究
- 傾向スコアの発展的利用

背景

ストレス潰瘍予防

- 敗血症ガイドライン - 重症敗血症 + 出血リスク
- 推奨の根拠
 - ICU患者
 - 人工呼吸患者
- 院内肺炎、*C. Difficile*腸炎の増加

重症敗血症に対して利益が不利益を上回るのか？

62

63

64

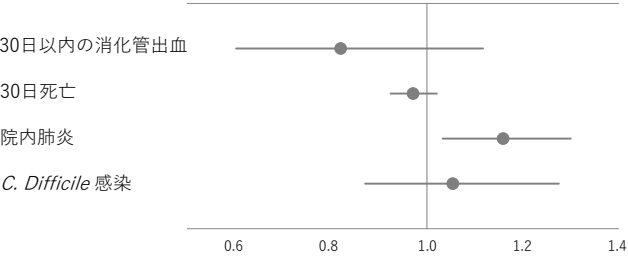
目的

- P：重症敗血症
- I：ストレス潰瘍予防
- C：ストレス潰瘍予防なし
- O：消化管出血/死亡
院内肺炎/*C. Difficile*腸炎

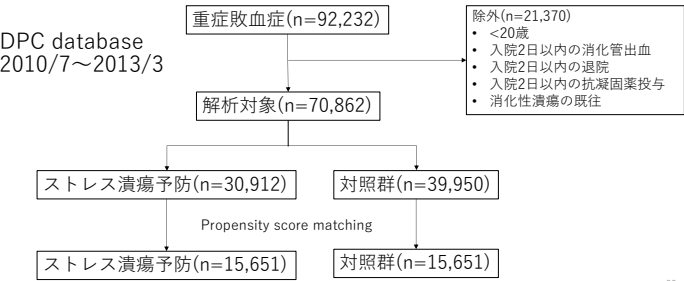
マッチング

- 1：1
- Nearest neighbor matching without replacement
- caliper：傾向スコアの標準偏差の0.2倍
- マッチングした2群のアウトカムを単純比較

治療効果の推定



方法



結果/患者背景

	Full cohort		PS matched	
	予防群	対照群	予防群	対照群
n	30,912	39,950	15,651	15,651
年齢, mean(SD)*	72.6 (14.3)	74.2 (14.8)	73.5 (14.3)	73.7 (14.8)
男性, (%)	55.4	55.0	55.1	55.5
障害臓器数, Median(IQR)	1 (1-2)	1 (1-1)	1 (1-1)	1 (1-1)
ICU, (%)	24.5	7.5	11.1	12.0
経腸栄養, (%)	7.4	2.4	3.6	3.4
人工呼吸, (%)	44.0	18.4	25.7	25.3
昇圧薬, (%)	61.2	35.5	46.3	45.2
腎代替療法, (%)	12.9	6.6	8.0	8.0

結語

重症敗血症に対するストレス潰瘍予防

死亡・出血は減らない
肺炎を増やす

臨床所見や検査値などは未測定

傾向スコアの推定

- 年齢／性別
- Charlson comorbidity index
- BMI
- 年度
- 大学関連病院
- 救急搬送の有無
- 施設別年間平均敗血症治療人数
- 臓器不全
- ICU入室の有無
- 経管栄養の有無
- 輸血の有無
- 昇圧薬使用の有無
- 腎代替療法の有無

結果/患者背景

	Full cohort		PS matched	
	予防群	対照群	予防群	対照群
n	30,912	39,950	15,651	15,651
年齢, mean(SD)*	72.6 (14.3)	74.2 (14.8)	73.5 (14.3)	73.7 (14.8)
男性, (%)	55.4	55.0	55.1	55.5
障害臓器数, Median(IQR)	1 (1-2)	1 (1-1)	1 (1-1)	1 (1-1)
ICU, (%)	24.5	7.5	11.1	12.0
経腸栄養, (%)	7.4	2.4	3.6	3.4
人工呼吸, (%)	44.0	18.4	25.7	25.3
昇圧薬, (%)	61.2	35.5	46.3	45.2
腎代替療法, (%)	12.9	6.6	8.0	8.0

Agenda

- はじめに
- 傾向スコア分析理解のための基礎知識
- 傾向スコアとは
- 傾向スコアを使った分析方法
- 傾向スコアの利点と限界
- 傾向スコアを利用した臨床疫学研究
- 傾向スコアの発展的利用

High dimensional propensity score (hd-PS)

- Schneeweiss (Schneeweiss S, et al. Epidemiology. 2009;20:512-22.)
- 傾向スコアは「未測定の交絡が存在しない」ことが前提
- 診療報酬請求データには臨床情報が含まれない

傾向スコアで大丈夫なのか？

診療報酬請求データ

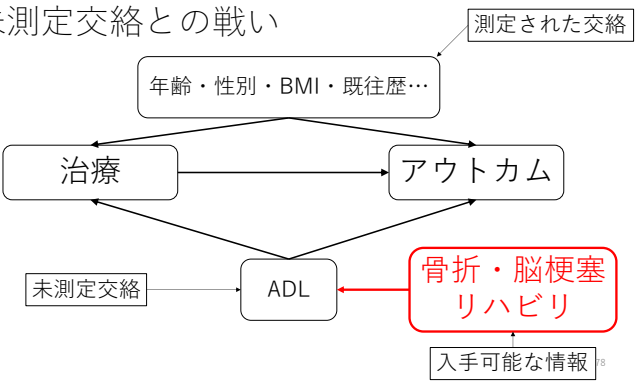
- 利点
- Primary dataの収集は高コスト
 - ランダム化が非倫理的
 - 圧倒的に規模が大きい
- 欠点
- 通常検査結果や画像診断等の臨床情報はない
 - コーディングエラー

診療報酬請求データ

- 利点
- Primary dataの収集は高コスト
 - ランダム化が非倫理的
 - 圧倒的に規模が大きい
- 欠点
- 通常検査結果や画像診断等の臨床情報はない
 - コーディングエラー

未測定の交絡

未測定交絡との戦い



診療報酬請求データ

- 利点
- Primary dataの収集は高コスト
 - ランダム化が非倫理的
 - 圧倒的に規模が大きい
- 欠点
- 通常検査結果や画像診断等の臨床情報はない
 - コーディングエラー

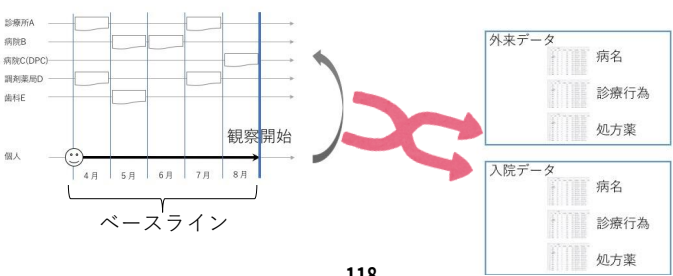
hd-PS

- 通常の分析と比較してRCTの結果に近づく
- 従来の傾向スコア分析よりも治療効果は期待値に近い
- 未測定の変数について二群でバランスが取れる

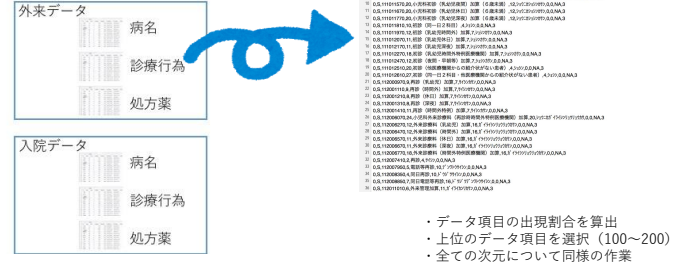
hd-PSの概略

1. 次元の決定 (データの種類)
2. 変数コードの抽出
3. 変数コードの出現回数による集計
4. 共変量のバイアス評価
5. 共変量の選択
6. 傾向スコアの推定

次元の決定



変数コードの抽出



- データ項目の出現割合を算出
- 上位のデータ項目を選択 (100~200)
- 全ての次元について同様の作業

変数コードの出現回数

	A	B	C	D	E	F	G	H	I	J	K
1 患者ID	診療行為A<1	診療行為A<median	診療行為A<75%	診療行為B<1	診療行為B<median	診療行為B<75%	診療行為C<1	診療行為C<median	診療行為C<75%	...	
2	1	1	1	0	1	1	1	1	0	0	...
3	2	0	0	0	1	0	0	0	0	0	...
4	3	1	1	1	1	1	1	1	1	0	...
5	4	0	0	0	0	0	0	0	0	0	...
6	5	1	0	0	0	0	0	1	1	1	...
7	6	1	1	1	1	0	0	0	0	0	...
8	7	1	1	0	1	1	0	1	1	0	...
9	8	1	1	1	1	1	1	1	1	1	...
10	9	1	0	0	0	0	0	0	0	0	...
11	10	1	1	1	1	1	0	1	0	0	...

- データ項目についてさらに集計
- ・ 1 回以上
 - ・ 中央値以上
 - ・ 75パーセンタイル以上

84

共変量のバイアス評価と選択

multiplicative bias (Bias_M)

$$Bias_M = \frac{Pc1(RRcd-1)+1}{Pc0(RRcd-1)+1}$$

- * Pc1とPc0:治療群と対照群における各変数の割合
- * RRcd:アウトカムに対する各変数の未調整リスク比

$$|\log(Bias_M)|$$

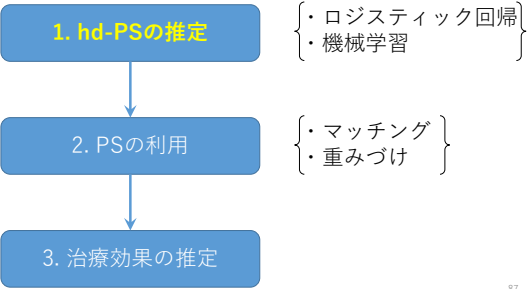
85

共変量の選択と傾向スコアの推定

- $|\log(Bias_M)|$ の大きいものから順に共変量を選択
一般的に50〜300程度の変数を選択する
- ロジスティック回帰モデルによって傾向スコアを推定
上記共変量
既知の重要な交絡因子

86

傾向スコア (PS) の利用



87

hd-PSの研究例

Association between perioperative oral care and postoperative pneumonia after cancer resection: conventional versus high-dimensional propensity score matching analysis

Clin Oral Investig. 2019;23:3581-3588.

88

背景

- ✓周術期の口腔ケアががん切除術後肺炎減少
これまでの研究は小規模
患者背景の厳密なコントロールが不充分
- ✓JMDCデータベースを利用
- ✓High-dimensional propensity score

90

デザイン

- Retrospective cohort study (JMDCデータベース)
- P : がん切除術を受けた患者
- E : 歯科医による術前口腔ケア
- C : 術前口腔ケアなし
- O : 術後肺炎

hd-PS

- ベースライン期間：365 day
- 次元数：5
- PSの推定：ロジスティック回帰モデル
- hd-PSの推定に用いた変数：500
+ その他の臨床的に重要な変数
- マッチングアルゴリズム：最近傍 1 : 1 マッチング
- 閾値：傾向スコアのlogitの0.2倍
- バランスの確認：Standardized mean differences

92

結果（患者背景）

	Unmatched		hd-PS-matched for postoperative pneumonia			
	With oral care % (n = 621)	Without oral care % (n = 4374)	ASD	With oral care % (n = 581)	Without oral care % (n = 581)	ASD
Age (years)						
≤ 50	30	48.9	0.395	31	31.2	0.004
51-60	37.7	29.3	0.179	36.8	36.8	0.000
≥ 61	32.4	21.9	0.238	32.2	32	0.004
Sex, female	50.6	47.3	0.066	49.1	50.9	0.038
Academic hospital	14.2	18.9	0.127	14.3	15.5	0.034
Hospital volume ≥ 12 cases/year	61.5	18.9	0.265	60.9	61.4	0.011
Cancer site						
Head and neck	2.4	1.8	0.046	2.2	2.6	0.022
Esophagus	4.2	1.3	0.177	3.1	3.4	0.019
Lung	19	28.5	0.224	19.1	20.1	0.026
Stomach	19.5	9.8	0.276	20.8	20.8	0.000
Colon and rectum	27.5	23.5	0.092	26.7	27.9	0.027
Breast	27.4	35.1	0.167	28.9	28.1	0.019
Disease history						
Hypertension	30.1	22	0.185	29.8	28.6	0.027
Diabetes mellitus	58.6	47	0.235	58	58.7	0.014
Renal failure	1.3	1.8	0.044	1.2	2.4	0.091
Ischemic heart disease	20.6	18.8	0.045	19.6	22.7	0.076
Chronic heart failure	15.9	14.6	0.038	15.8	14.5	0.038
Pneumonia	13.7	13.1	0.018	13.9	14.6	0.020
COPD	4.5	3.1	0.073	4.5	3.4	0.093
Asthma	10	12.5	0.081	10.2	10.8	0.022
Cerebrovascular disease	9.8	7	0.101	9.6	7.2	0.087

93

まとめ

- 傾向スコアの分析手順
PS推定→PS利用→効果推定
- 未測定 of 交絡には無力
- hd-PSは未測定 of 交絡もある程度調整できる可能性

96

結果（PSモデルに含めていない変数）

	Unmatched		hd-PS-matched for postoperative pneumonia			
	With oral care % (n = 621)	Without oral care % (n = 4374)	ASD	With oral care % (n = 581)	Without oral care % (n = 581)	ASD
Body mass index (kg/m ²)						
< 18.5	7.2	6.7	0.020	7.2	5.0	0.094
18.5-24.9	54.4	44.5	0.199	54.6	54.7	0.003
≥ 25.0	14.5	12.2	0.067	13.9	13.8	0.005
Missing data	23.1	36.5	0.279	24.3	26.5	0.051
Smoking						
Current or ex-smoker	21.3	14.7	0.171	21	19.8	0.030
Nonsmoker	53	47.3	0.114	52.7	52.3	0.007
Missing data	25.8	38.0	0.265	26.3	27.9	0.035
Alcohol consumption						
Drinker	41.1	32.1	0.186	40.4	40.3	0.004
Nondrinker	30.4	26.1	0.096	30.8	28.2	0.057
Missing data	28.5	41.5	0.275	28.7	31.5	0.060

94

結果

		Postoperative pneumonia	
		With oral care	Without oral care
Unmatched group	n	621	4374
	Event, n (%)	18 (2.9)	139 (3.2)
	Risk difference (%), (95% CI)	− 0.3 (− 1.7 to 1.1)	
hd-PS matched group	n	581	581
	Event, n (%)	17 (2.9)	19 (3.3)
	Risk difference (%), (95% CI)	− 0.3 (− 2.3 to 1.7)	

95

操作変数法・不連続回帰デザイン・ 差の差分析

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
麻生将太郎

本日の内容

- 操作変数法
- 不連続回帰デザイン
- 差の差分析

傾向スコア

- 交絡因子をできる限り調整して、傾向スコアを作成
- ランダム化比較試験と同等
- 測定していない交絡因子が存在したら？
- 傾向スコアで調整しても、真の結果とは異なる結果を生むかもしれない

未測定の交絡因子も調整できるのが
操作変数法

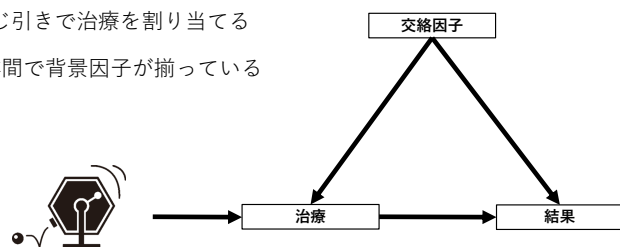
1

2

3

ランダム化比較試験

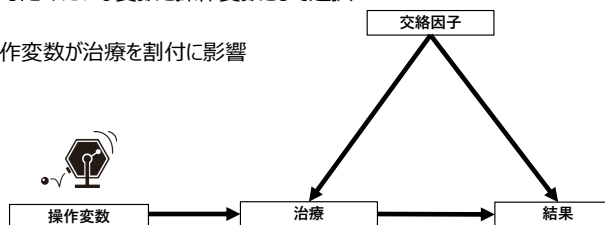
- くじ引きで治療を割り当てる
- 2群間で背景因子が揃っている



4

操作変数法

- くじ引きみたいな変数を操作変数として選択
- 操作変数が治療を割付に影響



5

操作変数の例

- 施設や医師の特定の治療の選好率(ある施設はA薬をよく処方する)
- 特定の治療を行う病院までの距離(内視鏡手術よく行う病院が近くにある)
- 日付や曜日(土日は休日なので、内科的治療を選びやすい)

適切な操作変数を見つけるのが難しい

6

操作変数法は万能か？

- 操作変数さえ見つけ出せば、未測定の交絡因子も調整できる
- 傾向スコアよりも優れている？
- 条件を満たしていないと操作変数法を適用できない

7

操作変数の条件

- ① 治療と強く関連
- ② アウトカムとは関連しない(治療を介してのみ関連)
- ③ 交絡因子と関連していない

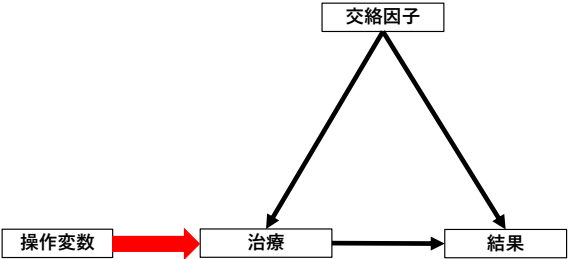
操作変数の条件

- ① 治療と強く関連
- ② アウトカムとは関連しない(治療を介してのみ関連)
- ③ 交絡因子と関連していない

8

9

① 治療と強く関連



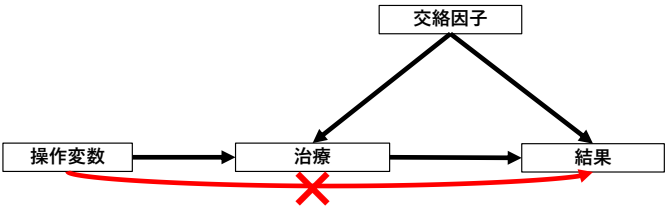
10

① 治療と強く関連

- 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値>10が目安)
- F値 ≤10: 弱い操作変数(weak instrument)で誤った結果を導く

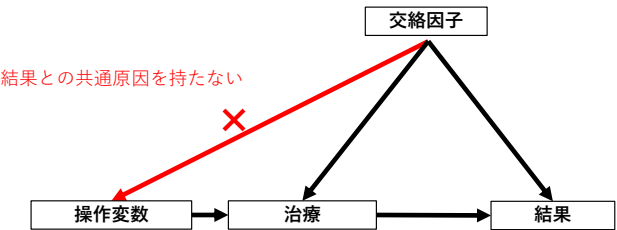
11

② アウトカムとは関連しない(治療を介してのみ関連)



12

③ 交絡因子とは関連しない



13

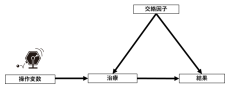
条件の評価

- ① 治療と強く関連 → F値>10
- ② アウトカムとは関連しない(治療を介してのみ関連)
- ③ 交絡因子と関連していない
→ 証明できない、臨床的に判断

14

単調性

- 操作変数法の仮定
- complierのみを対象(集団全体の効果を見ていない)
- cate (complier average treatment effect)



		操作変数(-)	
		治療を受ける	治療を受けない
操作変数(+)	治療を受ける	always-taker	complier
	治療を受けない	defier	never-taker

15

推定方法

- Wald推定量
- 2-stage least squares (2SLS)
- 2-stage residual inclusion (2SRI)

など

16

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
- 1段階目で、治療を操作変数や交絡因子で重回帰分析
- 回帰の結果で、治療の予測値を算出
- 2段階目で、アウトカムを治療の予測値と交絡因子で重回帰分析

17

2 SRI

- 非線形モデルを仮定
- アウトカムが二値の時に適用されることが多い
- 1段階目で、治療を操作変数や交絡因子で回帰分析
- 回帰の結果で、治療の実測値と予測値の差(応答残差)を算出
- 2段階目で、アウトカムを治療と交絡因子と応答残差で回帰分析

18

論文を書く時のChecklist

- ① CATEや単調性を明記
- ② 操作変数と治療の関連をF値を用いて明記
- ③ 交絡因子と操作変数や治療の関係を記載
- ④ 複数の操作変数を用いる場合は、過剰識別制約を明記
- ⑤ 二値のアウトカム、治療、操作変数の組み合わせを明記
- ⑥ 2値アウトカムでは、ロバスト分散がブートストラップ法を用いる

19

本日の内容

- 操作変数法
- 不連続回帰デザイン
- 差の差分析

22

限界

- 適切な操作変数を見つけることが難しい
- 操作変数と結果や交絡因子が関連がないことを確実に証明できない
- 仮定が多い
- 操作変数単独の研究は多くない(傾向スコアと併用)
- 傾向スコア分析と結果が異なることも

20

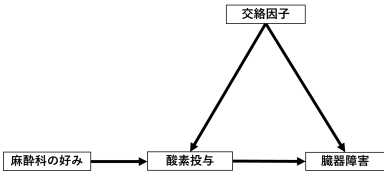
ある疑問

- 「あるウイルス性疾患で38℃以上の発熱に対して解熱薬を使用したら効果があるか？」を**観察研究**で実施したい
- 解熱剤使用群と未使用群では患者背景が異なる
- ベースラインの体温は使用群で高い
→傾向スコアを用いても調整は難

23

術中酸素投与と臓器障害の関連

- P : 術中SPO₂>92%の患者
- E : 酸素投与
- C : 酸素投与なし
- O : 術後臓器障害
- 操作変数 : 麻酔科医の酸素投与の好み(選好率)



21

不連続回帰デザイン

- ガイドライン上38℃を超えると解熱剤の推奨
- 37.9℃は解熱剤未使用で、38.0℃は使用することが多い
→38℃が解熱剤使用の閾値
→臨床的には37.9℃と38℃には大きな違い
- 一方で、単なる体温という数字であれば
- 37.9℃と38.0℃はほとんど同じ値で、差はない
38℃近傍の患者はほとんど背景が同じ

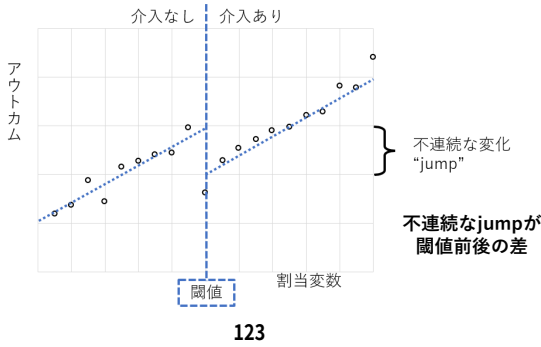
38℃未満 vs. 38℃以上で比較したら、RCTと類似

24

不連続回帰デザイン(Regression discontinuity)

- ある介入に用いられる割当値(連続)に閾値がある
- その閾値近傍で、閾値未満と以上でアウトカムを比較
- ランダム化比較試験と類似した状況を生み出す

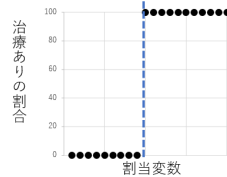
25



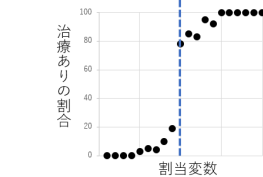
26

Sharp RDとFuzzy RD

- 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



27

Fuzzy RD

- 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
 - 背景と閾値は関連しない
- 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)

操作変数と同じ

28

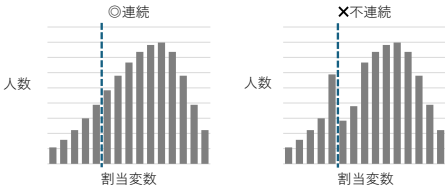
不連続回帰デザインの条件

1. 割当のルールと閾値が明確
 - 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する場合は×

29

不連続回帰デザインの条件

3. 割当変数の分布が閾値前後で不連続ではない (操作されていない)
 - 分布をチェック



30

メリットとデメリット

メリット

- 閾値近傍の局所的な無作為割当が可能 → RCTと同等の内的妥当性

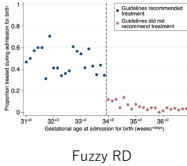
デメリット

- 比較しているのは閾値近傍 → 一般化可能性に乏しい
- 閾値近傍なので、サンプルサイズが小さくなる

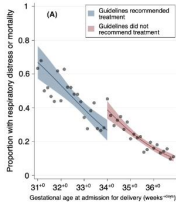
31

出生前ステロイドと新生児呼吸障害

- P : 新生児
- E : 出生前ステロイド
- C : 出生前ステロイドなし
- O : 新生児呼吸障害
- 割当変数 : 出生週数
- 介入 : ガイドラインのステロイド投与推奨(34週以下)



BJOG. 2024;131:1064–1071

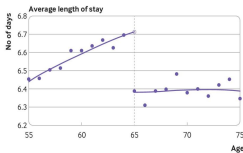


32

65歳のMedicare加入と外傷の入院アウトカム

BMJ 2023;382:e074289

- P : 外傷患者
- E : Medicare加入資格
- C : Medicare加入資格なし
- O : 在院日数、退院先など
- 割当変数 : 年齢
- 介入 : 65歳



33

本日の内容

- 操作変数法
- 不連続回帰デザイン
- 差の差分析

34

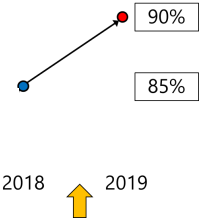
差の差分析を行うモチベーション

- ガイドラインが変わったことでアウトカムはどうなった？
- 政策や法律が施行された影響を知りたい
- イベントが起きた前後で比較する(前後比較)
- 前後比較は単に時間経過を見ているだけではないか？
- RDでは対照群がいないので、イベントの影響がどうかわからない

経時的な変化を評価

まずは2018年と2019年の介入前後で評価

- ただ単に経年変化を見てるだけかも
- 介入の効果がよくわからない

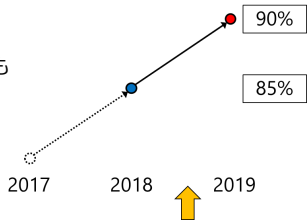


35

36

経時的な変化を評価

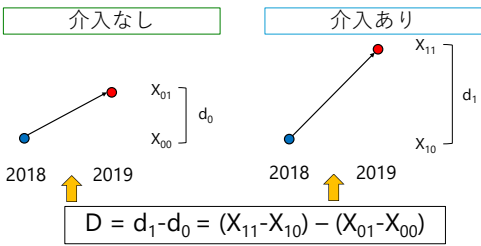
- 2017年も併せて経時変化を見たら
介入の効果がわかる？
- ただ単に年毎に改善しているだけかも



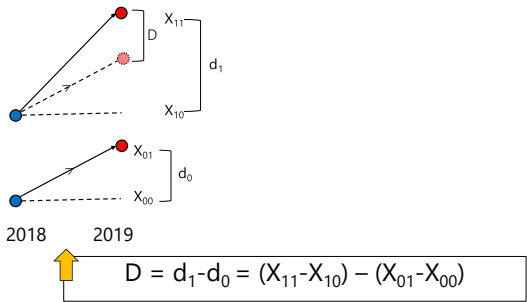
37

経時的な変化を評価

- 対照群を設ければ、介入の効果がわかる



38



39

交互作用を導入した回帰モデル

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	

40

Difference in differenceの仮定

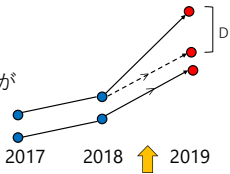
- ① 平行トレンド(parallel trends)
- ② 共通ショック(common shocks)

平行トレンド(parallel trends)

- 介入がなかった場合、介入群のアウトカムの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



42

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が発生した場合、外的要因の影響の程度が介入群と対照群で同じ
 - 天災、新興感染症
 - 効果を検証したい外的要因と同時に施行された、全く別の要因
2020年4月に施行された受動喫煙対策とCOVID-19
- 検証方法：なし

43

東日本大震災と避けられた入院

J Epidemiol Community Health 2017;71:248-252

- P：避けられた入院
- E：東北6県
- C：東北6県以外の都道府県
- O：避けられた入院
- 割当変数：年
- 介入：東日本大震災(2011年)

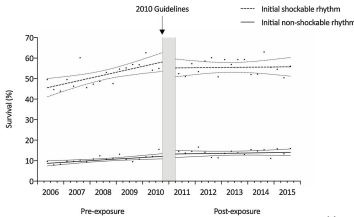
125

44

ショック非適用心肺停止とアトロピン

Resuscitation. 2019;137:69-77.

- P：心肺停止患者
- E：アトロピン投与なし
- C：アトロピン投与
- O：生存
- 割当変数：年
- 介入：2010年ガイドライン改訂



45

疫学・統計の応用(3) 多重代入法

東京大学大学院 医学系研究科 健康科学・看護学専攻
看護管理学／看護体系・機能学分野 講師
森田 光治良

Contents

1. 分析データの欠測の問題
2. 欠測のメカニズム
3. よくみられる対処方法
 - 完全ケース分析
 - 欠測を示す値の設定
 - 単一補完法
 - 多重代入法
4. 多重代入法の注意点やポイント
5. 報告のポイントと実際の研究例

Contents

1. 分析データの欠測の問題
2. 欠測のメカニズム
3. よくみられる対処方法
 - 完全ケース分析
 - 欠測を示す値の設定
 - 単一補完法
 - 多重代入法
4. 多重代入法の注意点やポイント
5. 報告のポイントと実際の研究例

欠測データとは

- 研究利用するはずの項目が、何かの理由によって完全には収集できず、データが測定できていないこと
 - 研究対象全員である項目をそもそも測定していない事ではない
- 解析に利用したい複数の変数で欠測が起こり得る
 - 共変量、アウトカム、またはその両方

NDB・DPCなどRWDの特徴

- 研究において予定していたデータを全て測定することは困難
- 特にRWDを用いた観察研究はデータ収集済みであり、予定された研究よりも欠測が起こりやすい

	RCT	RWD
設定	管理された特殊な環境	日常臨床
データの量	△～○ 参加者を集めるのは大変	○～◎ 集めやすい(まれな疾患、有害事象)
データの質	○～◎ 必要なデータ項目の測定・収集が可能 対象者の脱落が課題	△ 必要なデータがあるとは限らない 入力されたデータの質や妥当性の問題
アウトカム	△ 稀にしか発生しないアウトカム、潜伏期間 が長いアウトカムを追跡するには不向き	△～○ 集団の規模や特性が合えば、十分な数の アウトカムが見込める
時間・費用	△ フォローアップに難(時間・費用)	○ 既にデータ収集が完了していることが多く、 経費や時間が比較的少ない

本セミナーの対象範囲

- RWDを用いた観察研究のうち、調整解析がメインとなる際に問題となるような共変量や交絡変数で欠測が発生する場面に焦点
- 対象者のアウトカムを複数回経時的に測定する研究での欠測問題は取り扱わない

欠測データ発生メカニズム

- 欠測への対処を行うには、欠測が起きたメカニズムを深く考える必要がある
- メカニズム
 - Missing completely at random (MCAR)
 - Missing at random (MAR)
 - Missing not at random (MNAR)

Missing completely at random (MCAR)

- データの欠測が完全にランダムに発生
 - データの欠測は、欠測した変数自体やその他の変数にも依存しない
- 例
 - ある日、血圧計が突然壊れ、血圧に欠測値が生じた
 - 突然の積雪に伴う交通機関麻痺のためセンターに来院できず、その日の対象者では各種検査値が測定できなかった



Contents

1. 分析データの欠測の問題
2. 欠測のメカニズム
3. よくみられる対処方法
 - 完全ケース分析
 - 欠測を示す値の設定
 - 単一補完法
 - 多重代入法
4. 多重代入法の注意点やポイント
5. 報告のポイントと実際の研究例

Missing at random (MAR)

- データの欠測が完全にランダムに発生
 - データの欠測は、欠測した変数自体には関係なく発生 + その他の変数では説明できる状況 (欠測がおきる確率は他の変数と関連)

● 例

- 心血管系疾患の既往があると血圧がよく測定されているが、既往がない場合や若年者ではあまり測定されていない。



10

Missing not at random (MNAR)

- データの欠測がその変数自体の影響で発生
 - 欠測データの確率が変数自体の値に依存
- 例
 - 肥満の人ほど、体重を回答しなかったため欠測が起こった
 - 救急来院時にショック状態で血圧が低い人で血圧の欠測が起こった



11

欠測に対する主な対処方法

- データの欠測メカニズムや状況によって対処方法は異なる
 - 多くの対処方法は、欠測メカニズムとしてMARを前提
- MCAR: Little's MCAR testによって統計的有意差($p < 0.05$)を示せば、MCARであることは棄却され、MARかMNARとされる。検定方法自体の検出力は低く、Type II errorが起きやすいため「MCARである」ことを証明することは困難。

12

欠測に対する主な対処方法

- MARかMNARかを客観的に判断することは不可能
 - 臨床的な知識を用いて何が妥当かを判断しなければならない
 - 多くの欠測補完方法がMARの仮定を前提としている

Contents

13

- 分析データの欠測の問題
- 欠測のメカニズム
- よくみられる対処方法
 - 完全ケース分析
 - 欠測を示す値の設定
 - 単一補完法
 - 多重代入法
- 多重代入法の注意点やポイント
- 報告のポイントと実際の研究例

完全ケース分析(Complete case analysis)

- 欠測データがある対象者は除外、データが完全にそろっている対象者のみで解析を行う方法

ID	年齢	男性	BMI	収縮期血圧
1	50	1	27	150
2	55	0	.	135
3	60	0	30	.
4	65	1	22	.
5	70	1	19	145
6	75	0	.	125
7	80	1	29	160
8	85	0	16	100



ID	年齢	男性	BMI	収縮期血圧
1	50	1	27	150
5	70	1	19	145
7	80	1	29	160
8	85	0	16	100

14

完全ケース分析(Complete case analysis)

妥当な場面	- 欠測のメカニズムがMCARである場合 →ランダムに対象者を除外した場合と同じ - アウトカム変数の欠測がMARによって発生し、解析で用いる共変量でアウトカム欠測が説明でき、共変量には欠測がない場合
問題点	- 欠測データを持つ対象者が多い場合、解析対象者数が減少して統計的検出力が減少、推定値の標準誤差や信頼区間が広がってしまう。 (論文を読む際の注意点) ある1つの変数で欠測がある対象者割合が少なくても、いくつかの変数で欠測が存在する場合、結果的に多くの対象者が解析から除外される。

15

Contents

- 分析データの欠測の問題
- 欠測のメカニズム
- よくみられる対処方法
 - 完全ケース分析
 - 欠測を示す値の設定
 - 単一補完法
 - 多重代入法
- 多重代入法の注意点やポイント
- 報告のポイントと実際の研究例

127

インジケータ法(Missing indicator method)

- 欠測には"999"のような値(カテゴリの一つ)を与える操作的方法

ID	年齢	男性	BMI	収縮期血圧
1	50	1	27	150
2	55	0	.	135
3	60	0	30	.
4	65	1	22	.
5	70	1	19	145
6	75	0	.	125
7	80	1	29	160
8	85	0	16	100



ID	年齢	男性	BMI	肥満	収縮期血圧	高血圧
1	50	1	27	1	150	1
2	55	0	.	999	135	0
3	60	0	30	1	.	999
4	65	1	22	0	.	999
5	70	1	19	0	145	1
6	75	0	.	999	125	0
7	80	1	29	1	160	1
8	85	0	16	0	100	0

- 残差交絡の問題を増加させる可能性や解釈困難な結果(予測不可能なバイアス)を生むリスクが高い

Knol MJ, et al. J Clin Epidemiol. 2010
Vach W, et al. Am J Epidemiol. 1991

16

17

Contents

1. 分析データの欠測の問題
2. 欠測のメカニズム
3. よくみられる対処方法
 - 完全ケース分析
 - 欠測を示す値の設定
 - 単一補完法
 - 多重代入法
4. 多重代入法の注意点やポイント
5. 報告のポイントと実際の研究例

単一補完法(single imputation methods)

- 欠測データを一つの値で補完する手法の総称

平均値補完法 (Mean imputation)	データを収集できた集団から平均値を求め、欠測した部分を平均値で置き換える
回帰補完法 (Regression imputation)	他の変数を用いて、欠測値のある変数を説明する回帰モデル(線形回帰、またはロジステック回帰など)を作成し、欠測した部分に回帰分析にもとづく予測値を補完
LOCF法 (Last observation carried forward)	時間をおいて何度も測定する変数で欠測が起きた場合に、最後に観察された値で欠測した部分を補完

単一補完法(single imputation methods)

- 平均値補完法 (Mean imputation)の例

ID	年齢	男性	BMI	収縮期血圧
1	50	1	27	150
2	55	0	.	135
3	60	0	30	.
4	65	1	22	.
5	70	1	19	145
6	75	0	.	125
7	80	1	29	160
8	85	0	16	100



ID	年齢	男性	BMI	収縮期血圧
1	50	1	27	150
2	55	0	23.8	135
3	60	0	30	135.8
4	65	1	22	135.8
5	70	1	19	145
6	75	0	23.8	125
7	80	1	29	160
8	85	0	16	100

$$(27+30+22+19+29+16)/6=23.8$$

$$(150+135+145+125+160+100)/6=135.8$$

単一補完法(single imputation methods)

- 推測統計を行うような一般的な疫学研究では推奨されない
 - 欠測値を100%正しく予測することは不可能なため、欠測値の補完には必ず不確実性が生じる
 - 単一補完後のデータセットでは、観測値と補完値は区別されずに解析することになり、不確実性は考慮されていない
 - 最終的な解析結果が欠損補完の不確実性を無視、誤った結果を導く可能性大(統計的な誤差を過小評価→狭い信頼区間に)

Donders AR, et al. J Clin Epidemiol. 2006

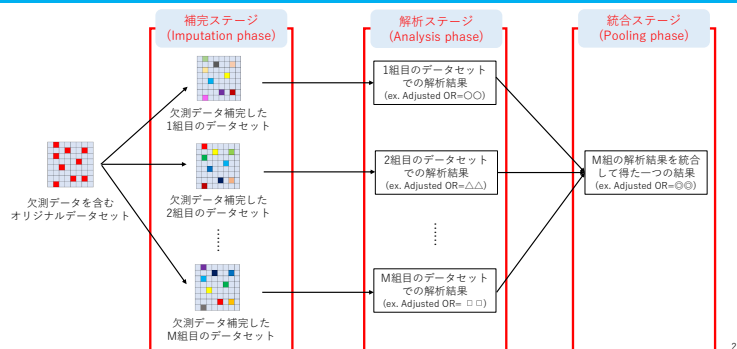
Contents

1. 分析データの欠測の問題
2. 欠測のメカニズム
3. よくみられる対処方法
 - 完全ケース分析
 - 欠測を示す値の設定
 - 単一補完法
 - 多重代入法
4. 多重代入法の注意点やポイント
5. 報告のポイントと実際の研究例

多重代入法(Multiple imputation)

- 欠測している変数の補完を同じ対象者に対して複数回行うことで、補完値の不確実性を考慮する方法
- 以下の3段階からなる解析方法
 1. 欠測値を補完する補完ステージ
 2. 研究目的に沿った解析を行う解析ステージ
 3. 最終結果を得る統合ステージ

Multiple imputationの流れ



補完ステージ

- 手元のデータセットのうち、欠測のない対象者のデータを用いて回帰モデルを作成、欠測変数の推定値と分散の情報を用いてランダムに補完値を生成
- 値を補完した擬似完全データセット作成を複数回 (M回) 繰り返し、M組の擬似完全データセットを作成。
 - 同じあるいは異なる値が補完されたデータセットが複数生成される

補完ステージ

- 連鎖方程式による多重補完法(Multivariate imputation by chained equations: MICE)
 - 最も一般的に利用される補完アルゴリズム「全ての変数が他の変数によってお互いに説明可能」という強い仮定を置くがよく臨床研究で利用される一標準的な統計ソフトで実装
 - 型が異なる複数変数(連続変数、二値変数、順序変数、名義変数など)での欠測、対象者ごとに異なる複数の変数で欠測があっても適用可能
 - 欠測変数ごとに補完を行う回帰モデルを指定

解析ステージ

- それぞれの擬似完全データセットそれぞれで個別に、元々の研究目的に沿った解析を実施
 - それぞれのデータセットで解析を行うため、解析結果はM個算出

統合ステージ

- 解析ステージで得られたM個の解析結果を統合して1つの最終結果を得る。
- Rubin's ruleとよばれる統合方法が用いられる
 - 点推定値($\hat{\theta}$) : M個の解析結果の平均 $\hat{\theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m$
 - 分散(T) : 「各補完データセットで得られた分散の平均(W)」と「補完データセット間での推定値のばらつき(B)」を用いて計算
$$T = W + \left(1 + \frac{1}{M}\right) B$$
 ここで、 $W = \frac{1}{M} \sum_{m=1}^M SE_m^2$, $B = \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_m - \hat{\theta})^2$

28

Stataでの解析例

喫煙とheart attackの関係を調査した症例対象研究デモデータ

Contains data from https://www.stata-press.com/data/r16/mheart0.dta				
obs:	154	154名	Fictional heart attack data; bmi missing	
vars:	9		19 Jun 2018 10:50	
size:	2,310			
variable name	storage type	display format	value label	variable label
attack	byte	%9.0g		Outcome (heart attack)
smokes	byte	%9.0g		Current smoker
age	float	%9.0g		Age, in years
bmi	float	%9.0g		Body Mass Index, kg/m^2
female	byte	%9.0g		Gender
hsgrad	byte	%9.0g		High school graduate
marstatus	byte	%9.0g	mar	Marital status: single, married, divorced
alcohol	byte	%24.0g	alc	Alcohol consumption: none, <2 drinks/day, >=2 drinks/day
hightar	byte	%9.0g		Smokes high tar cigarettes

29

Stataでの解析例

	attack	smokes	age	bmi	female	hsgrad	marstatus	alcohol	hightar
1	0	0	60.292	21.11455	0	0	Divorced	Three or more drinks/day	0
2	0	0	45.01435	24.8684	0	1	Married	Less than 3 drinks/day	0
3	0	0	61.95533	30.50274	0	0	Divorced	Three or more drinks/day	0
4	0	0	39.60295	24.62917	1	1	Single	Three or more drinks/day	0
5	0	0	45.47691	22.52744	1	1	Married	Less than 3 drinks/day	0
6	0	0	49.75778	21.87975	0	1	Married	Three or more drinks/day	0
7	0	1	65.39862	17.72857	0	0	Single	Less than 3 drinks/day	0
8	0	0	60.35888	.	0	0	Single	Do not drink	0
9	0	0	51.81486	23.47249	1	0	Divorced	Less than 3 drinks/day	0
10	1	0	61.33911	24.48016	0	1	Divorced	Less than 3 drinks/day	0
11	0	1	57.75695	.	1	1	Married	Less than 3 drinks/day	0

BMIで欠測が存在

Logistic回帰を行ってみると・・・

. logistic attack smokes age bmi hsgrad female

Logistic regression
Log likelihood = -79.34221

分析に利用した変数が全て揃っている人に限定して解析される(Complete case analysis)

Number of obs = 132
LR chi2(5) = 24.03
Prob > chi2 = 0.0002
Pseudo R2 = 0.1315

	attack	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
smokes	4.683533	1.872631	3.86	0.000	2.13912	10.25444
age	1.026456	.0174929	1.53	0.125	.9927368	1.06132
bmi	1.119625	.0559881	2.26	0.024	1.015096	1.234917
hsgrad	1.49904	.6664762	0.91	0.363	.6271443	3.583102
female	1.252987	.567297	0.50	0.618	.5158935	3.043217
_cons	.0044788	.0081094	-2.99	0.003	.0001288	.1557224

31

misstable summarize

Variable	Obs=.	Obs>.	Obs<.	Obs<.		
				Unique values	Min	Max
bmi	22	132		132	17.22643	38.24214

. mi set flog 欠測補完することの宣言とデータの持ち方の指定

. mi register imputed bmi 欠測補完する変数の宣言
(22 m=0 obs. now marked as incomplete)

. mi register regular attack smokes age hsgrad female 欠測補完の予測に使用する変数の宣言

32

Stataでの解析例

. mi impute chained (regress) bmi = attack smokes age hsgrad female, add(20) rseed(2024)

note: Missing-value pattern is monotonic; no iteration performed

Conditional models (monotone):
bmi: regress bmi attack smokes age hsgrad female

Performing chained iterations ...

Multivariate imputation
Chained equations
Imputed: m=1 through m=20
Initialization: monotone
Iterations =
burn-in =

Imputations = 20

added = 20

updated = 0

Iterations =

burn-in =

bmi: linear regression

Variable	Observations per m			
	Complete	Incomplete	Imputed	Total
bmi	132	22	22	154

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)

BMIの欠測を線形回帰で補完、補完セットは20セット、再現性確保のため乱数シードの固定

	attack	smokes	age	bmi	female	hsgrad	marstatus	alcohol	hightar	_mi_id	_mi_miss	_mi_m
1	0	0	60.292	21.11455	0	0	Divorced	Three or more drinks/day	0	1	0	0
2	0	0	45.01435	24.8684	0	1	Married	Less than 3 drinks/day	0	2	0	0
3	0	0	61.95533	30.50274	0	0	Divorced	Three or more drinks/day	0	3	0	0
4	0	0	39.60295	24.62917	1	1	Single	Three or more drinks/day	0	4	0	0
5	0	0	45.47691	22.52744	1	1	Married	Less than 3 drinks/day	0	5	0	0
6	0	0	49.75778	21.87975	0	1	Married	Three or more drinks/day	0	6	0	0
7	0	1	65.39862	17.72857	0	0	Single	Less than 3 drinks/day	0	7	0	0
8	0	0	60.35888	.	0	0	Single	Do not drink	0	8	1	0
9	0	0	51.81486	23.47249	1	0	Divorced	Less than 3 drinks/day	0	9	0	0
10	1	0	61.33911	24.48016	0	1	Divorced	Less than 3 drinks/day	0	10	0	0
11	0	1	57.75695	.	1	1	Married	Less than 3 drinks/day	0	11	1	0

元々のデータセット

	attack	smokes	age	bmi	female	hsgrad	marstatus	alcohol	hightar	_mi_id	_mi_miss	_mi_m
2927	0	0	60.292	21.11455	0	0	Divorced	Three or more drinks/day	0	1	0	19
2928	0	0	45.01435	24.8684	0	1	Married	Less than 3 drinks/day	0	2	0	19
2929	0	0	61.95533	30.50274	0	0	Divorced	Three or more drinks/day	0	3	0	19
2930	0	0	39.60295	24.62917	1	1	Single	Three or more drinks/day	0	4	0	19
2931	0	0	45.47691	22.52744	1	1	Married	Less than 3 drinks/day	0	5	0	19
2932	0	0	49.75778	21.87975	0	1	Married	Three or more drinks/day	0	6	0	19
2933	0	1	65.39862	17.72857	0	0	Single	Less than 3 drinks/day	0	7	0	19
2934	0	0	60.35888	24.02175	0	0	Single	Do not drink	0	8	0	19
2935	0	0	51.81486	23.47249	1	0	Divorced	Less than 3 drinks/day	0	9	0	19
2936	1	0	61.33911	24.48016	0	1	Divorced	Less than 3 drinks/day	0	10	0	19
2937	0	1	57.75695	38.51223	1	1	Married	Less than 3 drinks/day	0	11	0	19

19 個目の補完データセット

	attack	smokes	age	bmi	female	hsgrad	marstatus	alcohol	hightar	_mi_id	_mi_miss	_mi_m
155	0	0	60.292	21.11455	0	0	Divorced	Three or more drinks/day	0	1	0	1
156	0	0	45.01435	24.8684	0	1	Married	Less than 3 drinks/day	0	2	0	1
157	0	0	61.95533	30.50274	0	0	Divorced	Three or more drinks/day	0	3	0	1
158	0	0	39.60295	24.62917	1	1	Single	Three or more drinks/day	0	4	0	1
159	0	0	45.47691	22.52744	1	1	Married	Less than 3 drinks/day	0	5	0	1
160	0	0	49.75778	21.87975	0	1	Married	Three or more drinks/day	0	6	0	1
161	0	1	65.39862	17.72857	0	0	Single	Less than 3 drinks/day	0	7	0	1
162	0	0	60.35888	19.09764	0	0	Single	Do not drink	0	8	0	1
163	0	0	51.81486	23.47249	1	0	Divorced	Less than 3 drinks/day	0	9	0	1
164	1	0	61.33911	24.48016	0	1	Divorced	Less than 3 drinks/day	0	10	0	1
165	0	1	57.75695	39.9687	1	1	Married	Less than 3 drinks/day	0	11	0	1

1 つ目の補完データセット

	attack	smokes	age	bmi	female	hsgrad	marstatus	alcohol	hightar	_mi_id	_mi_miss	_mi_m
3881	0	0	60.292	21.11455	0	0	Divorced	Three or more drinks/day	0	1	0	20
3882	0	0	45.01435	24.8684	0	1	Married	Less than 3 drinks/day	0	2	0	20
3883	0	0	61.95533	30.50274	0	0	Divorced	Three or more drinks/day	0	3	0	20
3884	0	0	39.60295	24.62917	1	1	Single	Three or more drinks/day	0	4	0	20
3885	0	0	45.47691	22.52744	1	1	Married	Less than 3 drinks/day	0	5	0	20
3886	0	0	49.75778	21.87975	0	1	Married	Three or more drinks/day	0	6	0	20
3887	0	1	65.39862	17.72857	0	0	Single	Less than 3 drinks/day	0	7	0	20
3888	0	0	60.35888	27.39241	0	0	Single	Do not drink	0	8	0	20
3889	0	0	51.81486	23.47249	1	0	Divorced	Less than 3 drinks/day	0	9	0	20
3890	1	0	61.33911	24.48016	0	1	Divorced	Less than 3 drinks/day	0	10	0	20
3891	0	1	57.75695	23.31224	1	1	Married	Less than 3 drinks/day	0	11	0	20

20 個目の補完データセット

34

35

30

Stataでの解析例

Stataでの解析例

32

33

Stataでの解析例

Stataでの解析例

36

Stataでの解析例

多重代入法の利点

```
. mi estimate: logistic attack smokes age bmi hsggrad female
. mi estimate, or
```

Multiple-imputation estimates

Logistic regression

DF adjustment: Large sample

Model F test: Equal FMI

Within VCE type: OIM

Imputations = 20

Number of obs = 154

Average RVI = 0.0099

Largest FMI = 0.1070

DF: min = 700.03

avg = 190.953.41

max = 413.350.62

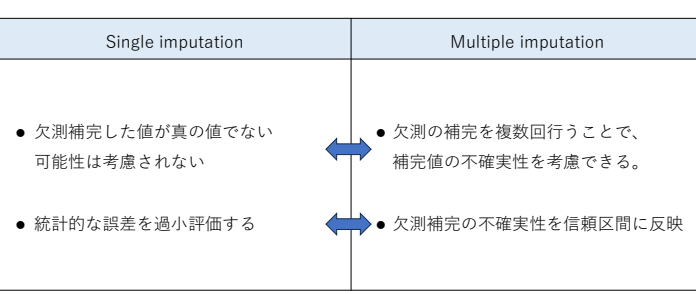
F(5,46129.4) = 3.67

Prob > F = 0.0025

欠測補完した20個のデータセットそれぞれで分析した推定結果を統合して表示

解析ステージと統合ステージが実施され、統合ステージでの最終結果のみ表示される

attack	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]
smokes	3.390795	1.227116	3.37	0.001	1.660204 6.092130
age	1.036013	.0160095	2.29	0.022	1.005106 1.067072
bmi	1.120615	.05552	2.30	0.022	1.016743 1.235008
hsggrad	1.107374	.4009703	0.42	0.672	.5367716 2.626549
female	.9003948	.3761051	-0.25	0.802	.3970777 2.041693
_cons	.0033204	.0057024	-3.33	0.001	.0001156 .095053



Contents

- 1. 分析データの欠測の問題
- 2. 欠測のメカニズム
- 3. よくみられる対処方法
 - ・完全ケース分析
 - ・欠測を示す値の設定
 - ・単一補完法
 - ・多重代入法
- 4. 多重代入法の注意点やポイント
- 5. 報告のポイントと実際の研究例

MARの仮定の確らしさ

アウトカム変数の取り扱い

補完ステージでのデータセット作成数

- 欠測値補完を行う回帰モデル
 - 最終解析に利用する全ての変数(アウトカム含む)、最終解析には使用しなくても欠測値を説明(欠測変数と関連)できる変数(=補助変数)の両方をモデルに投入
 - 補助変数を欠測補完モデルに組み込むことができればMARの仮定を満たすことができる
 - MAR: その他の変数では説明できる状況(欠測がおきる確率は他の変数と関連)→補助変数が必要!
- MARの仮定を満たせないと、最終結果にバイアス

- アウトカム自体の欠測補完、共変量の欠測補完を行う回帰モデルへのアウトカム投入の両方で利用可能
 - 元々、RCTにおける追跡不能や副作用による治療中止などアウトカム測定ができない状況を想定して発展
 - 研究目的とする解析の従属変数と説明変数は、欠測補完において数学的に区別されない
 - アウトカムの利用は精度の高い結果をもたらす

- MIが提案された当初、理論的には補完セットは5組程度で十分とされていた

Rubin D. Biometrika(1976)
- 近年は100-1000組とも
 - データセット数が少ないと、信頼区間の精度が不十分となる可能性
 - 多数のデータセット作成を可能とするコンピューター性能の向上

WhiteRI, et al. Stat Med(2011)
- 少なくとも20組の作成が推奨

Sterne JA, et al. BMJ. 2009

どれくらいの欠測割合が扱える？

- 欠測補完が正しく行われ、MARの仮定が満たされれば、欠測の割合は関係ない
 - 欠測割合は、MIを実施するか否かの指針にはならない
 - MARの仮定が満たされず多重代入の手順も不適切だと、欠測割合が高いほどバイアスが大きくなる可能性
 - 30-50%以上の欠測を補完する場合は、特に注意を払うのが賢明との意見も

Contents

- 1. 分析データの欠測の問題
- 2. 欠測のメカニズム
- 3. よくみられる対処方法
 - ・完全ケース分析
 - ・欠測を示す値の設定
 - ・単一補完法
 - ・多重代入法
- 4. 多重代入法の注意点やポイント
- 5. 報告のポイントと実際の研究例

欠測とMultiple imputationの報告

- 観察的疫学研究報告の質改善のための声明(STROBE声明)の拡張、そのほか文献の報告に関する提案をまとめた推奨内容

Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. BMJ. 2009 Jun 29;338:b2393. doi: 10.1136/bmj.b2393. PMID: 19564179; PMCID: PMC2714692.

- (1) 各変数での欠測割合、可能であれば欠測の起きた理由
- (2) 欠測による対象者除外をした場合の人数と理由
- (3) 欠測のあった集団と欠測の無い集団の特徴比較 (Tableで各変数の比較を示す)
- (4) 欠測に対応した解析方法と欠測の起きたメカニズムの説明

Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. BMJ. 2009 Jun 29;338:b2393. doi: 10.1136/bmj.b2393. PMID: 19564179; PMCID: PMC2714692.

- **MIを利用した場合**
 - (1) 使用ソフトウェア、欠測補完モデルで用いた変数と変数の扱い方などモデルの詳細、補完したデータセット数
 - (2) データの大部分を代入した場合には、観測値と補完値の比較
 - (3) 多重代入法から得た結果とComplete case analysisの結果比較
 - (4) MARの仮定が妥当に関する考察(補助変数を含められたか)
 - (5) 感度分析(MARの仮定からの逸脱の検討)

46

Social, Behavioral, and Metabolic Risk Factors and Racial Disparities in Cardiovascular Disease Mortality in U.S. Adults : An Observational Study.

He J, Bundy JD, Geng S, Tian L, He H, Li X, Ferdinand KC, Anderson AH, Dorans KS, Vasan RS, Mills KT, Chen J. Ann Intern Med. 2023 Sep;176(9):1200-1208.

- 社会的・行動的・代謝的なリスク因子がCVD死亡率とどの程度関連するか、またこれら因子を考慮した上で人種差がCVD死亡にどの程度寄与するかを検討。
- NHANES（National Health and Nutrition Examination Survey）：2年ごとに実施される国民健康栄養調査
- 論文記載内容
 - データの一部と1999～2004年の睡眠データの欠測があること
 - Multiple imputation by chained equation
 - 欠測はMARと仮定したこと(なぜそのように仮定したか記載無し)
 - アウトカム変数およびすべての共変量を欠測補完モデルに投入したこと
 - 10セットの作成(統合方法に関しては記載なし)
 - 感度解析としてComplete case analysis

47

Adverse Pregnancy Outcomes and Long-Term Mortality in Women.

Crump C, Sundquist J, Sundquist K. JAMA Intern Med. 2024 Jun 1;184(6):631-640.

- 1973～2015年までの間に単胎出産を経験した約220万人の女性を対象に、5つの主要妊娠合併症（早産、子宮内発育不全児、妊娠高血圧腎症、その他の高血圧疾患、妊娠糖尿病）と長期的な死亡リスクの関連を評価
- Swedish Medical Birth Register：1973年からの国内全出生前・出生情報レジストリ
- 論文記載内容
 - 妊娠期間または出生児体重が欠測している母親(0.3%)は研究から除外
 - 欠測していない割合の記載：母親のBMI(56%)と喫煙(67%)、それ以外の変数(99%)
 - 研究の共変量と死亡を予測因子として多重代入、20セットの作成
 - 感度解析としてcomplete case analysisとMissing indicator method

48

Breast cancer mortality in 500 000 women with early invasive breast cancer diagnosed in England, 1993-2015: population based observational cohort study.

Taylor C, McGale P, Probert J, Broggio J, Charman J, Darby SC, Kerr AJ, Whelan T, Cutter DJ, Mannu G, Dodwell D. BMJ. 2023 Jun 13;381:e074684.

- イングランドで1993～2015年に早期浸潤性乳癌と診断された女性約50万人を対象に、特に2010～15年に診断された女性に焦点を当て、スクリーニングで発見されたかなど、特徴ごとの絶対死亡リスク推定を目的。
- National Cancer Registration and Analysis Service（NCRAS）：イングランド全域のがん患者が登録、がん疫学研究、治療効果の長期追跡など、臨床研究と公衆衛生政策に利用されるレジストリデータ
- 欠測状況の記載内容「腫瘍サイズ、陽性リンパ節の数、腫瘍のグレード、エストロゲン受容体の状態、HER2の状態などの変数に欠測があった」
- 対応方法の記載内容「多重代入法を使用した。欠測データの詳細と多重代入については、補足テキストS2、表S1、および表S2に記載」

49

Breast cancer mortality in 500 000 women with early invasive breast cancer diagnosed in England, 1993-2015: population based observational cohort study.

Taylor C, McGale P, Probert J, Broggio J, Charman J, Darby SC, Kerr AJ, Whelan T, Cutter DJ, Mannu G, Dodwell D. BMJ. 2023 Jun 13;381:e074684.

- Supplementary Appendixの記載
 - 欠測割合
 - 研究利用した6つの全変数情報が揃っている割合(25.5%)と欠測割合（一つの変数26.7%～6つの全変数0.04%）の提示
 - 欠測のパターンの提示(Table S1)と欠測変数と罹患年度の関係性(Table S2)
 - 欠測割合はがん診断年と強い相関一年々、データ収集入力の努力が増したと考察

50

Breast cancer mortality in 500 000 women with early invasive breast cancer diagnosed in England, 1993-2015: population based observational cohort study.

Taylor C, McGale P, Probert J, Broggio J, Charman J, Darby SC, Kerr AJ, Whelan T, Cutter DJ, Mannu G, Dodwell D. BMJ. 2023 Jun 13;381:e074684.

- Supplementary Appendixの記載 (Multiple imputationに関して)
 - 欠測のある変数は全てカテゴリ変数で予測には順序ロジスティック回帰を利用
 - 予測に用いた独立変数、補助変数の説明
 - 補完データセット数(60セット)
 - 最終推定値と分散はRubin's rulesを用いて得たことの記載

51

- RWDとの親和性
 - 複数変数で欠測発生、MARの仮定を満たせると思われれば、MIを使用することを検討
- 批判
 - MIは主要な統計解析ソフトで利用可能で比較的簡単に適用できるため、解析の適切性が十分検討されないなど、基本的な間違いが生じる恐れ
- 欠測メカニズムをよく考える
 - 利用するRWDのデータ収集プロセスに詳しい専門家、臨床の専門家、疫学家、生物統計家との議論
 - 検証不可能な仮定の元に成り立つため、MARの仮定を逸脱していないか
- 現在でも様々な状況下でのMIの利用について開発・検討が続いている

52

疫学・統計の応用(4)薬剤疫学研究のデザイン

東京大学大学院医学系研究科臨床疫学・経済学
渡邊 秀明

1

より詳しく知りたい場合；参考論文

- 観察データを使用して治療エピソードを構築する際の薬理学的および疫学的考察: シミュレーション研究

L. Pazzagli et al. Pharmacoepidemiol Drug Saf. 2022 Jan;31(1):55-60. doi: 10.1002/pds.5366. Epub 2021 Oct 15..

4

Induction period (誘導期間) Latent period (潜在期間)

- Induction period (誘導期間)：**「薬Xに曝露してから腎保護作用が発現するまでには、数週間？くらいかかるかな」
- Latent period (潜在期間)：**「薬Xの腎保護作用が発現しても、最初は微細な効果で、観測可能（血清Crの違いなど）になるまでは数週間？数ヶ月？かかるかな」

7

What is the appropriate research design in pharmacoepidemiology?

- 薬剤の曝露の定義をどのようにすればバイアスの少ない研究デザインになるのか？
- 同じ薬剤への曝露が継続しているか、中断したかをどう定義するか（Persistence/discontinuation）？ Grace period(猶予期間）

2

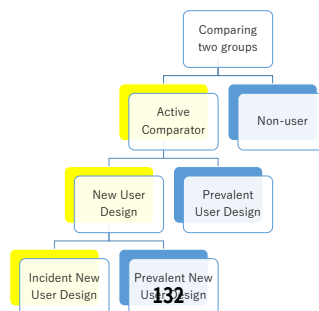
Persistence vs Adherence in general.

- 「患者がどの程度処方スケジュールに従い薬剤を内服したか」
- Adherence: The extent to which a patient acts in accordance with the prescribed interval and dose of a dosing regimen
- 「患者がどの程度薬剤を内服し続けたか、もしくは薬剤の内服開始から中断まで」
- Persistence: The duration of time from initiation to discontinuation of therapy

J Cramer et al. Medication Compliance and Persistence: Terminology and Definitions
<https://doi.org/10.1111/j.1524-4733.2007.00213.x>

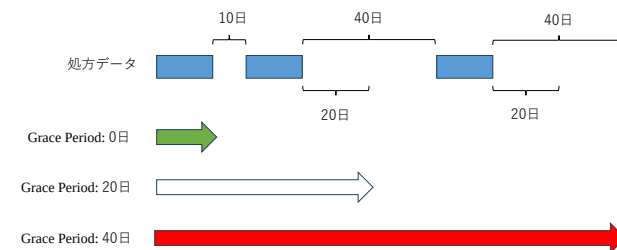
5

これからのお話の結論：



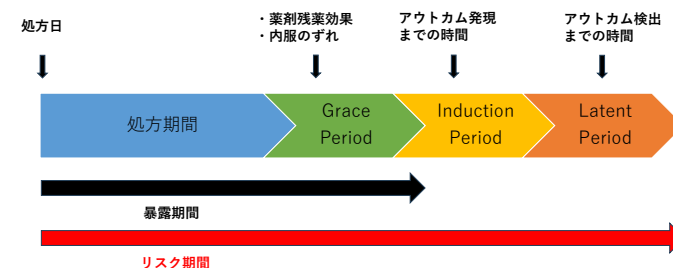
8

Grace period (猶予期間)； 薬剤の残存効果を加味する>半減期より決定



3

Induction period (誘導期間) Latent period (潜在期間)



6

Prevalent user bias(既存処方者バイアス)

- 追跡開始前から、すでに研究対象とする医薬品を使用していた人（既存処方者患者；prevalent user）を曝露群に含めてしまう事によって生じるバイアス
- 既存処方者患者はそれまで副作用が生じなかったから薬を飲み続けられている人。曝露群に副作用が起こりにくい人々が選択的に含まれていることから、選択バイアスが生じていることになる。

Prevalent user biasを評価した論文

観察研究における既存処方者バイアス：
スタチンのメタ分析から得た有効性比較研究への教訓

G Danaei et al. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. PMID: 22223710

10

Prevalent user biasを評価した論文

- スタチンが冠動脈疾患・死亡リスクを低下させることは、数多くのランダム化比較試験で証明。
- しかし、観察研究においてはスタチンの効果に関して系統的ビュー、メタ解析が行われていなかった。
- そのためDanaeiらは、スタチンの冠動脈疾患・死亡のリスク低減効果に関して観察研究のレビューを行うとともに、Prevalent user biasの影響を可視化するため、観察研究側のコホートをPrevalent user（既存処方者）、Incident user（新規処方者）、その両方と層別化し、結果がどう変わるかを比較した。

G Danaei et al. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. PMID: 22223710

11

メタ解析における結果はスタチンの非薬剤使用者に対する死亡のハザード比の比較

- RCT: 0.84 (95%CI 0.77-0.91)
- Incident user vs non user: 0.77 (0.65-0.91)
- Prevalent user and Incident user vs non user: 0.70 (0.64-0.78)
- Prevalent user vs non user: 0.54 (0.45-0.66)

G Danaei et al. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. PMID: 22223710

12

Prevalent user bias：選択バイアス

- # 仮定（薬剤安全性の観点）
 - 薬Xの内服は副作用をもたらす
 - ある患者の元々の体調によっては薬Xの内服を中止せねばならず、さらに元々の体調は患者の死亡リスクに影響する
- 曝露群において、“薬剤を使い続けられた人”だけが選択的に選ばれている。

13

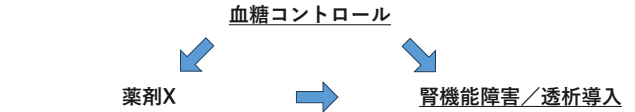
Prevalent user bias：選択バイアス

- # 仮定（薬剤有効性の観点）
 - 薬Xの内服は特に良い効果をもたらさない
 - ある患者の元々の体調によっては薬Xの内服を中止する必要がある、さらに元々の体調は患者の死亡リスクに影響する
- 曝露群において、“薬剤を使い続けられた人”だけが選択的に選ばれている。
- >ベースラインで体調の良い患者（Healthy user）のみ選ばれてしまっている

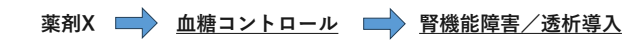
14

Prevalent user bias：中間因子となる可能性

研究：「薬剤Xは非使用者と比べて腎機能障害／透析導入のリスクを抑制するかどうか」



ではなく、



15

Prevalent user bias：
曝露の定義を明確にできない可能性

薬剤の曝露期間や累積曝露量によりアウトカムの発生リスクが変化する際に生じる。
>薬剤XとアウトカムY（腎代替療法導入など）の関連を調べる際に、薬剤Xを10年使用した患者と半年使用した患者、ではおそらくアウトカムYのリスクなどは変わってくるかもしれない。

16

Prevalent user biasへの対処法

- Incident new user design
- Prevalent new user design

17

Incident new user design

- 2つの薬剤を比較する際に、双方の患者を「新規に薬剤を使用し始めた患者」に限定するデザイン
- “新規に薬剤を使用し始めた”の定義は？
 - >Washout period：薬剤処方時点からどの程度振り返れば妥当？
 - >多くの薬剤疫学研究では6-12ヶ月のWashout periodが使われている。しかし、それで本当に妥当かは微妙。

18

Incident User Study Designにおけるウォッシュアウト期間の再考:
6 ～ 12 か月では十分ではない可能性

	Incident user コホート内での誤分類割合		
Washout期間 (休薬期間)	脂質異常症	糖尿病	うつ病
3ヶ月	76.6%	69.9%	66.8%
6ヶ月	53.3%	45.6%	47.7%
12ヶ月	32.8%	25.5%	32.3%
24ヶ月	11.6%	8.1%	13.1%

A. Roberts et al. J Comp Eff res. 2015 Jan;4(1):27-35. doi:10.2217/ceer.14.53. 19

Incident new user design: pitfalls

- サンプルサイズの減少
> 前述の論文では6ヶ月、 1 2 ヶ月のwashout periodで75%、 8 5 %の患者の除外
- 癌など長期の観察期間が必要なものでは薬剤有効性・安全性を追跡できなくなる可能性
- 研究結果の一般化可能性が下がる

E. Johnson et al. The incident user design in comparative effectiveness research PMID: 23023988

Incident new user design: pitfalls

- 研究結果の一般化可能性が下がる
> 薬Aと薬Bが同じクラスの薬剤の場合は「これまで使っていた薬Aの効果がいまいちだから、新しく薬Bに切り替えよう」は評価できない。

E. Johnson et al. The incident user design in comparative effectiveness research PMID: 23023988 21

Prevalent user biasへの対処法

- Incident new user design
- **Prevalent new user design**

Prevalent new user design

- Prevalent user design
- New user design

Then...

- **Prevalent new user design**

22

S. Suissa et al. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores PMID: 27610604 23

Prevalent new user design

- 新薬A(●)と以前から使用されている薬剤B (X)との間で薬剤有効性を比較する場合、Prevalent new user designでは、薬剤Bから新薬Aにスイッチした患者も”新薬AのNew user”としてコホートに含めることを許容

S. Suissa et al. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores PMID: 27610604 24

Prevalent new user design
(Time-based exposure set)

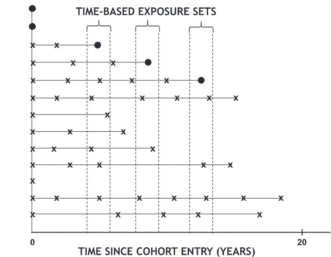


Figure 2. Depiction of a base cohort formed of 13 subjects with time-based exposure sets defined by a small time interval surrounding the timing of the new study drug prescription (●)

25

Prevalent new user design
(Time-based exposure set)

- 薬剤B (X)を使用し続けられた患者 (Healthy user)
- 薬剤B (X)への曝露累積期間が同レベルの患者

Prevalent new user design
(Prescription based exposure set)

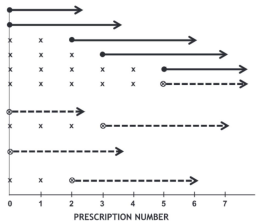


Figure 4. Depiction of the final prevalent new-user comparative cohort with comparator drug subjects matched to study drug subjects on the number of comparator drug prescriptions (X) before the first study drug prescription (●) was given, with the arrows indicating the follow-up period for outcomes

26

27

Prevalent new user design
(Prescription based exposure set)

- 薬剤B (X)を使用し続けられた患者 (Healthy user)
- 薬剤B (X)への曝露累積量が同レベルの患者

Prevalent new user design: pitfalls #1

- ・アウトカム発症者を除外することについて
>既存薬非使用群における見かけ上のX発症リスクが上昇してしまう可能性。
- 対処法：プロペンシティスコアを計算し、マッチングを行う段階では、既存薬使用者のReplacementを行わない

Prevalent new user design: pitfalls #2

- ・ Cohort組み入れ前の既存薬への曝露期間・量が不明

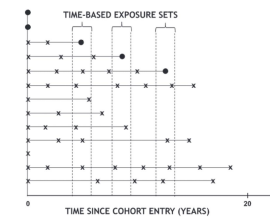
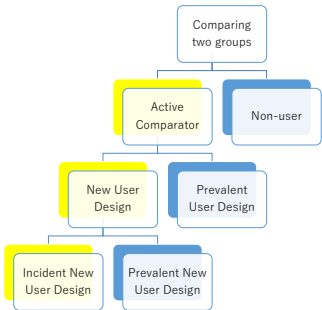


Figure 2. Depiction of a base cohort formed of 13 subjects with time-based exposure sets defined by a small time interval surrounding the timing of the new study drug prescription (x)

Prevalent new user design: pitfalls

- ・ Prevalent new user designでもSelection biasを完全には削除できない
- 結局、新薬A群はIndex date!において既存薬から新薬へのスイッチをしており、この状況は実臨床では「既存薬を使用していて有害事象を経験した」、「既存薬では十分対処できないほど疾患が悪化した」からかもしれない。
- 新薬・既存薬の群間で疾患重症度や、有害事象への感受性に差が出る可能性。
- 薬剤選択に伴うSelection biasを完全に避けるためには、やはりIncident new user designを使用するしかない。

結局どうしたら良いのか？



Active comparatorとは？

- ・ PECOのControl(ターゲット薬の比較群)として、薬剤を使用していない群 (Non-user)ではなく、 Exposureと適応が同じであり、かつExposureの使用が禁忌ではない患者群を選択すること
- >なるべくRCTに模倣する(target trail emulation) という考え方

The gold standard to manage confounding by indication
: Active Comparator

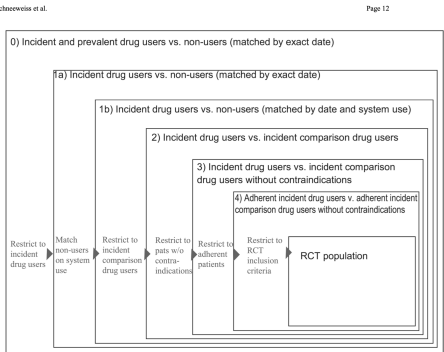


FIGURE 1. Five consecutive population restrictions to approximate RCTs in pharmacoepidemiologic database studies.

S. Schneeweiss et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results PMID: 17909372

研究計画！

- ・ Keywords:
Grace period,
Induction period,
Latent period,
Active comparator,
Incident New user design,
Washout period

XXX database

[PICO]	
P	People with type 2 diabetes stage 4 or worse chronic kidney disease
I	Drug X
C	Drug Y
O	Primary: KRT initiation or LDKT/DDKT
	Secondary: Rate of eGFR decline

Method : XXX database

- Design: Incident New User design
- Active comparator:
 - >WHO-ATC code: drugs (ARB,ACEi, DM medications, insulin)
 - >WHO ICD10 code; disease
- Covariates: heart failure, obesity(BMI), hypertension, dyslipidemia, smoking, albuminuria, MI, unstable/stable angina, stroke, TIA, amputation
- >Labs; HbA1c

37

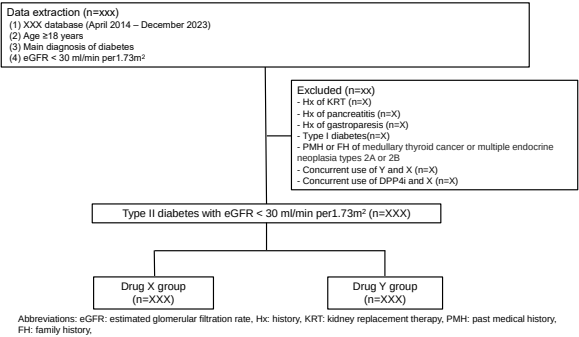
Outcome Definition Repository (ODR)

- Outcome Definition Repository (ODR) は、アウトカム定義を集積・登録・公開するためのデータベース。
- アウトカム定義には、日本の医療情報データベースを用いて実施された研究にて、ある疾患や曝露の状態・有無を決定するために使用されたデータベース上のコードとその組み合わせ（アルゴリズム）を含む。
- 日本薬剤疫学会、日本疫学会、日本臨床疫学会が共同で提供。

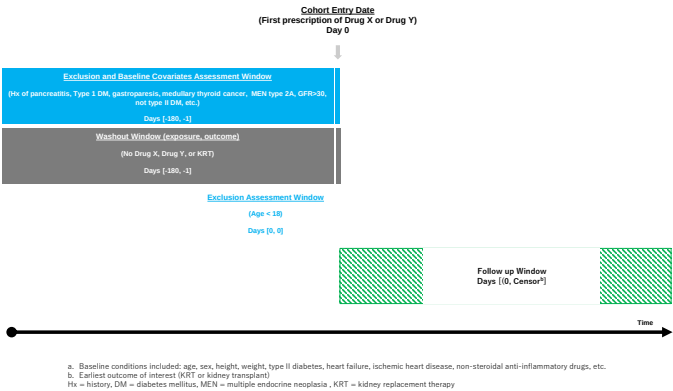
引用 : <https://odr-jp.net>

38

Fig.1



39



a. Baseline conditions included: age, sex, height, weight, type II diabetes, heart failure, ischemic heart disease, non-steroidal anti-inflammatory drugs, etc.
b. Earliest outcome of interest (KRT or kidney transplant)
Hx = history, DM = diabetes mellitus, MEN = multiple endocrine neoplasia, KRT = kidney replacement therapy

41

研究事例紹介 ;
Active comparator, new user design

- 当研究室からの論文-
- Y Kimura et al. J Allergy Clin Immunol Pract. 2024 May 9;S2213-2198(24)00467-7. doi: 10.1016/j.jaip.2024.05.002. PMID: 38734374
- British Medical Journal-
- E Fu et al. BMJ. 2024 Jun 26;385:e078483. doi: 10.1136/bmj-2023-078483. PMID: 38925801

Active comparator, new user design

新規血糖降下薬とメトホルミンと喘息増悪の減少との関連性

Y Kimura et al. J Allergy Clin Immunol Pract. 2024 May 9;S2213-2198(24)00467-7. doi: 10.1016/j.jaip.2024.05.002.

42

Active comparator, new user design

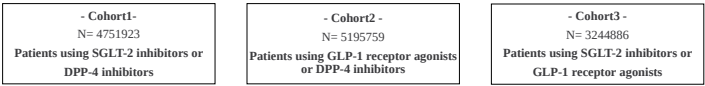
P: 喘息患者
E: DPP4i, GLP1RA, SGLT2i
C: Metformin (Active comparator)
O:喘息の増悪

Active comparator, new user design

SGLT-2 阻害剤、GLP-1 受容体アゴニスト、DPP-4 阻害剤と2型糖尿病患者における高カリウム血症のリスク:
住民を対象としたコホート研究

Active comparator, new user design

P: 糖尿病患者
E/C: SGLT2i, GLP1RA, or DPP4i
O: 高K血症の発症



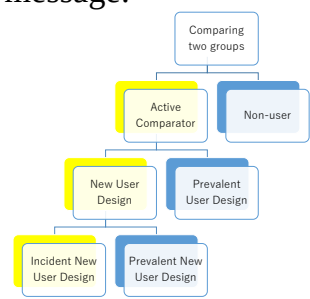
Y Kimura et al. J Allergy Clin Immunol Pract. 2024 May 9;S2213-2198(24)00467-7. doi: 10.1016/j.jaip.2024.05.002.

43

E Fu et al. BMJ. 2024 Jun 26;385:e078483. doi: 10.1136/bmj-2023-078483.

45

Take home message:



NDB・DPC等データベース研究人材育成Webinar

民間企業によるリアルワールドデータ利活用

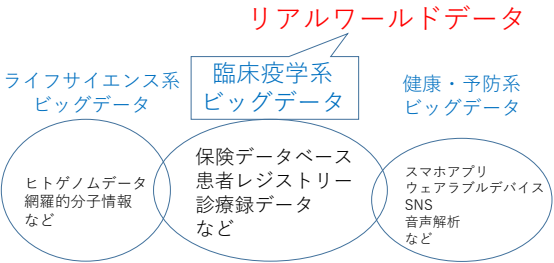
東京大学大学院医学系研究科臨床疫学・経済学
康永秀生

Topics

- 1. リアルワールドデータの概要（復習）
- 2. 民間企業によるRWDの活用
- 3. 【参考】RWDを用いた薬剤研究

1. リアルワールドデータの概要（復習）

様々な医療ビッグデータ



さまざまなRWD

1. 保険データベース (Administrative claims data)
保険者などから収集した診療報酬請求データを研究に二次利用。

- (1) 公的
匿名医療保険等関連情報データベース (NDB)
匿名診療等関連情報データベース (DPCDB)
要介護認定情報・介護レセプト等情報 (介護DB)、等
- (2) 民間
JMDC、DeSC、MDV、等

さまざまなRWD

2. 患者レジストリー(patient registry)

政府や学術団体などが特定の疾患・治療に関する患者データを各医療機関等から収集。

- (1) 公的レジストリ
全国がん登録データベース、匿名感染症関連情報データベース、指定難病患者データベース、小児慢性特定疾病児童等データベース、など
- (2) アカデミックレジストリー
National Clinical Database (NCD)、日本整形外科学会症例レジストリー (JOANR)、日本ICU患者データベース (JIPAD)、日本外傷データベース、等

さまざまなRWD

3 診療録データ (electronic medical records)

電子カルテ、検査データ、など

Medical Information Database Network (MID-NET)
国立病院機構データベース
徳洲会メディカルデータベース
など

データベースの種類	アカデミアによるデータ入手可能性	民間企業によるデータ入手可能性
患者レジストリー (アカデミア)	○	×
患者レジストリー (公的)	○	○
保険データベース (公的)	○	○
保険データベース (商用)	○	○
MID-NET	○	○

RWDの時代

RWDを多施設から恒常的に収集し、蓄積し、大規模データベースを構築。

データベースから個々の研究目的に沿ったデータセットを抽出し、エビデンスを量産。

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。

2. 民間企業によるRWDの活用

民間企業によるRWD活用の可能性

- (1) Unmet medical needsの把握
- (2) 臨床試験デザインにおけるプロセスの効率化
- (3) 製造販売後の安全性評価
- (4) 医療経済評価
- (5) 臨床試験の結果を補完する資料の作成
- (6) 臨床試験の外部対照群にRWDを活用

(1) Unmet medical needsの把握

特定の医薬品・医療機器の使用対象となる疾患の記述疫学分析や診療実態分析(practice pattern analysis)→unmet medical needsの把握

記述疫学分析

患者数や患者背景の把握、有病率・罹患率の推計

診療実態分析

当該疾患を有する患者に実施された手術・処置・処方・リハビリなどの実態を把握。医薬品の場合、用量、処方の期間や継続状況、併用薬の状況、などを把握。特に希少疾患では、患者数や性・年齢構成、重症度の分布といった疫学情報を把握。医薬品・医療機器の製造販売後の診療実態分析にもRWDを活用できる。

結果

ATTRwtの診断：3255例（100万人あたり155.8例）～3992例（100万人あたり191.1例）
ATTRmの診断：67例（100万人あたり3.2例）～106例（100万人あたり5.1例）

アミロイド軽鎖アミロイドーシスの診断：444人（100万人あたり21.2人）

心不全と診断された患者のうち、ATTRwtの診断：1468例（100万人あたり70.3例）から1798例（100万人あたり86.1例）
心不全と診断された患者のうち、ATTRmの診断：50例（100万人あたり2.4例）から61例（100万人あたり2.9例）

【背景】

2018年4月に「医薬品の製造販売後の調査および試験の実施の基準に関する省令」（GPSP省令）が改正され、製造販売後調査に「医療情報データベース」を加えることとした。

厚生労働省は2021年3月に、「承認申請等におけるレジストリの活用に関する基本的考え方」および「レジストリデータを承認申請等に利用する場合の信頼性担保のための留意点」を公表した。

<例1>疾患の疫学データ

トランスサイレチン型アミロイド心筋症の有病率の推計

Estimating the Prevalence of Transthyretin Amyloid Cardiomyopathy in a Large In-Hospital Database in Japan. Cardiol Ther 2019;8:297-316

背景

トランスサイレチン型アミロイド心筋症（ATTR-CM）は、TTR遺伝子の変異によるもの（ATTRm）と野生型（ATTRwt）がある。

ATTR-CMの世界的な有病率は、過小評価されている可能性がある。日本では、疫学的な有病率調査が行われていない。

本研究は、MDVデータを用いて、ATTR-CMの有病率を推計することを目的とした。

結論

この後方視的研究により、日本の大規模な院内データベースにおける9年間のATTR-CMと診断された患者数の推計値が得られた。疾患への認知度を向上させることにより、診断と治療が改善される可能性がある。

<例2> 診療実態分析

サトラリズマブを用いた視神経髄炎スペクトラム障害患者の診療実態

Real-world management of patients with neuromyelitis optica spectrum disorder using satralizumab: Results from a Japanese claims database. Multiple Sclerosis and Related Disorders 2024;84:105502.

背景

サトラリズマブ: 抗インターロイキン-6受容体モノクローナル抗体
視神経髄炎スペクトラム障害 (NMOSD) の治療薬
2つのRCT (SAkuraSkyおよびSAkuraStar) の結果に基づき承認。

NMOSDのリアルワールドでの診療実態、特に併用療法を漸減中の患者管理については不明。
サトラリズマブ治療を受けたNMOSD患者におけるグルココルチコイドや免疫抑制剤の併用、再発など、リアルワールドでの治療パターンを分析した。

19

結論

これらのリアルワールドデータは臨床試験を支持するものである。
サトラリズマブ投与期間 (中央値197.0日) において、大多数 (95.4%) の患者が無再発であった。
無再発でグルココルチコイド・フリーとなった患者数は、サトラリズマブ継続処方のもとで時間の経過とともに増加した。

22

(4) 医療経済評価

医療経済評価
効果の指標：生存年または質調整生存年(quality-adjusted life years, QALYs)
費用：医療費および生産性損失
医療費の算出には、レセプト情報を活用できる。

中央社会保険医療協議会：費用対効果評価の分析ガイドライン
「各健康状態の費用の推計において、(中略) 実臨床を反映した国内におけるレセプトのデータベースを用いることを推奨する」

25

方法

データソース：MDVデータ
Index date: サトラリズマブの初回処方日
研究期間：2018年8月～2022年3月
対象：2020年8月～2022年3月の間にサトラリズマブの初回処方を受け、ICD10コードG36.0 (視神経髄炎) の診断名があり、Index dateより90日以上前から観察可能であった患者。

主要評価項目：
サトラリズマブ投与継続360日後に無再発で経口グルココルチコイドが0mg/日まで減少
副次的評価項目：
再発までの期間、サトラリズマブ継続投与中のIndex date以降の再発数、Index date前後の再発率、併用薬の使用

20

(2) 臨床試験デザインにおけるプロセスの効率化

臨床試験を計画している医薬品・医療機器においては、臨床試験デザインにおけるプロセスを効率化するために、以下の目的でRWDを活用できる。

- ①組み入れ基準と除外基準に基づいた患者数の推計
- ②対象とする患者集団が多い地域などの把握
- ③類似薬に関する情報の把握

特に、既に市販されている類似薬がある場合、処方パターン、服薬コンプライアンス、アウトカムなどを調査することにより、active comparatorの情報を得られる。
これらにより、臨床試験の実施可能性(feasibility)や試験期間設定の根拠資料を得られる。

23

(5) 臨床試験の結果を補完する資料の作成

臨床試験による有効性・安全性の評価結果を補完する資料として、RWDの分析結果を活用することも考えられる。臨床試験のデータに追加して、RWDから得られたデータを利用して、一般化可能性を高められるかもしれない。

<例>
承認済みの医薬品・医療機器について、承認時の臨床試験の組入れ基準が一部の患者層に限られていた場合、適応拡大や添付文書の改訂に当たって、RWDを用いた有効性・安全性評価の結果を補足資料として活用する。

臨床試験に基づき用法・用量の記載が設定されたものの、実臨床ではそれと異なる用法・用量で使用されており、有効性・安全性に大きな影響がないと考えられる場合にも、RWDを用いてそれを検証し、適応拡大や添付文書の改訂に繋げる。

結果

N=131
女性：90.8%、18～65歳：75.6%、グルココルチコイド経口投与：93.1%
アザチオプリン投与：19.1%、タクロリムス投与：18.3%
生物学的製剤の使用：6例 (トシリズマブ1例、エクリズマブ5例)。

Index date前360日間観察可能な111例のうち、再発の回数は平均0.6回
サトラリズマブ投与期間の中央値 (四分位範囲) は197.0 (57.0-351.0) 日

ほとんどの患者 (125/131例、95.4%) はindex date後無再発であった
6例はindex date後90日以内に再発し、うち2例はindex date 後7日以内に初回再発であった。
360日間追跡された21人の患者のうち、6人 (28.6%) がindex date後360日以内に再発することなく0mg/日のグルココルチコイドを処方されていた。

21

(3) 製造販売後の安全性評価

RWDを用いて、対象となる医薬品・医療機器を使用した患者群だけでなく、使用していない対象群も加えて、製造販売後の安全性等の評価ができる。

しかし、保険データベースのみを用いて、薬剤の有害事象の頻度を調べる製造販売後安全性調査を実施することは推奨されないことがある。
入力病名の感度が低く、特に軽症例は記載されないことがあり、有害事象の頻度が過小評価されるかもしれない。

<例>入院中に薬剤性の肝機能障害が発生したとしても、軽度であれば、DPCデータ上に「肝機能障害」とわざわざ書かれないことがある。
休薬と安静以外に特段の治療をすることなく短期間で寛解してしまうような症例では、担当医は律儀に「肝機能障害」と入力しないかもしれない。

DPCデータの他に血液検査結果の情報も含むデータを導入したデータベース (MID-NET、など) の使用が適当である。

24

(6) 臨床試験の外部対照群にRWDを活用

臨床試験において、RWDを外部対照として承認申請等における有効性・安全性の評価に活用することも検討可能である。
特に希少疾患、遺伝子治療、再生医療などの領域において、**症例数不足により対照群を置いたRCTが困難である場合**、患者レジストリーのデータを外部対照として活用することが考えられる。
欧米では先行事例があり、今後日本でも検討の余地はある。

26

27

【アメリカの事例】 CLN2の治療薬Cerliponase alfaの臨床試験
(Schulz A, et al. Study of Intraventricular Cerliponase Alfa for CLN2 Disease.
N Engl J Med 2018;378:1898-1907)

セロイドリボフスチン症2型（CLN2）：トリペプチジルペプチダーゼ欠損を伴う遺伝性疾患であり、小児期から進行性認知症を引き起こすまれなライソゾーム病。
CLN2の治療薬Cerliponase alfa：遺伝子組換えヒトトリペプチジルペプチダーゼ1（hTTP1）酸素前駆体であり、脳室内投与される。

デザイン：非盲検単群用量漸増試験とその継続試験、および患者レジストリーを用いたヒストリカルコントロールによる試験
対象と介入：3～16歳の患児23例に対してCerliponase alfaを30mg、100mg、300mgのいずれかで開始し、その後全例に300mgを少なくとも96週間投与。
外部対照：患者レジストリーから未治療のCLN2患者42例を抽出
主要転帰：CLN2臨床評価尺度の運動領域と言語領域のスコアが2点低下するまでの期間

結果、Cerliponase alfa投与群は外部対照群と比較し、運動機能や言語機能の低下が抑制されると結論づけられた。これらをもとに、Cerliponase alfaは2017年にFDAに承認された。

28

【研究例1】

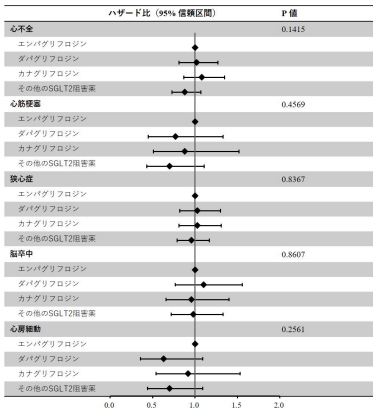
SGLT2阻害薬による循環器疾患の発症リスク抑制効果

Suzuki Y, et al. Comparison of cardiovascular outcomes between SGLT2 inhibitors in diabetes mellitus. *Cardiovascular Diabetology* 2022;21: 67

背景

- ✓ SGLT2阻害薬は近位尿管での糖の再吸収を阻害することで血糖値を下げる。
- ✓ 近年の大規模臨床試験からSGLT2阻害薬は糖尿病の有無に関わらず、循環器疾患（とりわけ心不全）の発症率を低下させることが示された。
- ✓ そのためSGLT2阻害薬は、糖尿病のみならず心不全などの循環器疾患などに用いられるようになっている。
- ✓ しかし、SGLT2阻害薬の薬剤間で循環器疾患の発症リスクを比較した研究は少ない。
- ✓ 日本で使用されている6種類のSGLT2阻害薬について、薬剤間で効果に差があるかどうかを検討した。

31



34

治療群と外部対照群を比較する場合の問題点

- ・治療群と外部対照群の間で組み入れ基準・除外基準が異なっていると、両群の比較が難しくなる。
- ・治療群と外部対照群の背景因子の違いが交絡となり、結果を歪める可能性がある（その場合、傾向スコアなどを用いて交絡調整が必要となる）。
- ・エンドポイントの測定方法が治療群と外部対照群の間で異なっていると、やはり比較が困難になる。
- ・外部対照群を用いてもなおサンプルサイズが不足することもある。

29

方法

データソース：JMDC Claims Database

対象：2005年1月から2020年4月までに糖尿病に対してSGLT2阻害薬が処方され、循環器疾患や透析治療歴のない25,315症例
(年齢中央値52歳、83%が男性、HbA1c中央値7.5%)

エンパグリフロジン (empagliflozin) n=5,302
ダパグリフロジン (dapagliflozin) n=4,681
カナグリフロジン (canagliflozin) n=4,411
イブラグリフロジン (ipragliflozin) n=5,275
トホグリフロジン (tofogliflozin) n=3,074
ルセオグリフロジン (luseogliflozin) n=2,572

アウトカム：心不全、心筋梗塞、狭心症、脳卒中、心房細動の発生

32

【研究例2】

高齢者脳梗塞の再発予防を目的としたワルファリンと直接作用型経口抗凝固薬 (DOAC) の比較

Direct oral anticoagulants versus warfarin for secondary prevention of cerebral infarction and bleeding in older adults with atrial fibrillation. *Journal of the American Geriatrics Society*. 2022;70(7):2029-2039.

近年、DOACは80代や90代の心房細動患者にも使用されている。
脳梗塞の既往のある高齢患者において、DOACがワルファリンと比較して、脳梗塞の再発予防に効果があるかどうかを示すデータは不足している。

35

3. 【参考】RWDを用いた薬剤研究

結果

平均観察期間814 ± 591日
855例の心不全、143例の心筋梗塞、815例の狭心症、340例の脳卒中、139例の心房細動が記録された。

年齢や性別、併存症やその他の糖尿病治療薬で補正後、エンパグリフロジン、ダパグリフロジン、カナグリフロジン、その他のSGLT2阻害薬の間で、心不全、心筋梗塞、狭心症、脳卒中、心房細動の発症リスクはいずれも同等であった。

30

方法

データソース：DPCデータ
期間：2014年1月1日から2020年3月31日まで
デザイン：後ろ向きコホート研究
対象患者：75歳以上の脳梗塞入院患者
弁置換術の手術歴がある患者等は除外。

曝露と対象：

- ✓ DOAC（ダビガトラン、リバーロキサパン、アピキサパン、エドキサパン）
- ✓ ワルファリン

アウトカム：

- ✓ 脳梗塞もしくは出血イベントによる再入院
- ✓ 再入院中の全死亡

36

統計解析

傾向スコアによる安定化逆確率重み付け法
(stabilized inverse probability of treatment weighting)

死亡を競合リスク(competing risk)とする
Fine and Gray部分分布ハザードモデル
(subdistribution hazard model)

結果

N = 106,187
DOAC群 (n = 84,720) とワルファリン群 (n = 21,467)

	DOAC群	ワルファリン群	リスク差	95%信頼区間	
脳梗塞	6.83%	8.92%	-2.09	-2.48	-1.70
出血イベント	1.25%	1.57%	-0.32	-0.49	-0.15
死亡	1.77%	2.53%	-0.77	-0.97	-0.56

	Subdistribution HR	95% CI	P
脳梗塞			
ワルファリン群	Reference		
DOAC群	0.76	0.71-0.81	<.001
出血イベント			
ワルファリン群	Reference		
DOAC群	0.80	0.70-0.91	0.001

【研究例3】

入院患者における薬剤性間質性肺疾患のリスク

Risk of drug-induced interstitial lung disease in hospitalised patients: a nested case-control study. Thorax. 2021;76(12):1193-1199.

考 察

75歳以上の脳梗塞入院患者において、DOAC群は、退院後の脳梗塞もしくは脳内出血による再入院、再入院中の総死亡が、ワルファリン群と比較して有意に少なかった。

背景と目的

- ・ 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、一般化が難しい
- ・ 大規模なリアルワールドデータを用いて、薬剤カテゴリー毎に薬剤性肺障害の発症リスクを評価する。

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年から2016年

対象：薬剤性肺障害と診断されステロイド治療を要した症例

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に記載のある42の薬剤カテゴリーを評価

方 法

統計解析：症例と対照を、性別、年齢（10歳以内）、病院、前後1年の会計年度で1対4マッチング

欠損値補完 + 条件付き多変量ロジスティック回帰分析

結 果

- ・ 約4177万人の入院患者のうち、薬剤性肺障害は2,342例。
- ・ 1：4症例-対照マッチングにより、症例1,541例と対照5,677例を分析対象とした。
- ・ 単変量解析では、5症例未満の処方であった薬剤カテゴリーは除外し、17薬剤カテゴリーを評価
- ・ 多変量解析では、単変量解析で有意となった10薬剤カテゴリーを評価

結果

10薬剤カテゴリーで行っ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	1.90	1.56-2.31	<0.001
Sulfamethoxazole/trimethoprim	2.54	1.04-6.24	0.042
Quinolones	3.10	2.41-3.99	<0.001
Tetracyclines	1.60	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

結 語

大規模リアルワールドデータを用いたコホート内症例対照研究により、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すことができた。

ご清聴ありがとうございました

46

47

自治体によるリアルワールドデータ利活用

自治医科大学 データサイエンスセンター
山名 隼人

自治医科大学のRWD研究

栃木県

- 栃木県全体：
25市町（14市11町）、人口約190万人（全国第19位）

宇都宮市	51万人
小山市	17万人
栃木市	15万人
- 18市町がデータ提供に同意

医療・介護連結データで可能な分析

3. 国民健康保険の集団を対象とした分析

- 特定健診をテーマにした研究
- 小児に関する研究
(↔ JMDCデータなど、健康保険組合の集団)

4. 地域医療分析

- 居住地による差
- 医療提供体制、医療計画 等

本日の内容

- これまでの自治医科大学の取り組み
- 研究の例
- 現在の事業・研究
- 今後の展望

自治医科大学のRWD研究

データ

- 国民健康保険レセプト
- 後期高齢者医療制度レセプト
- 介護レセプト
- 特定健診情報
- 被保険者台帳
(保険の資格開始・終了、要介護度)

期間

- 2014～2019年（約5年分）

個人レベルで連結
→
縦断研究

自治医科大学での人材育成

対象（従来）

- 自治医科大学データサイエンスセンター 大学院生
- 自治医科大学各科 臨床医
(総合診療科 外科 産婦人科 泌尿器科 等)
- その他
(東京大学 獨協医科大学 群馬大学 等)

自治医科大学のRWD研究

データベース

- 県単位のレセプトデータ
- 市町、国民健康保険団体連合会、後期高齢者医療広域連合の協力を得て実施
- 国民健康保険・後期高齢者医療制度の医療レセプト、介護レセプト等入手
- データベース化、医療・介護連結データを作成
→ 教育・研究に利活用

医療・介護連結データで可能な分析

1. 医療レセプトを中心として介護を追加

- 高齢者を対象とした縦断研究
- 要介護度をアウトカム or 変数に加える
- 急性期から回復期

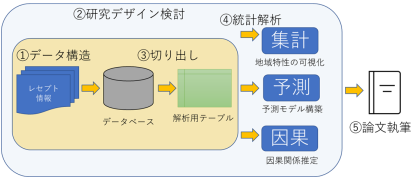
2. 介護サービスの研究

- サービスの効果
- 特定の疾患を持つ患者が受けているサービスの記述

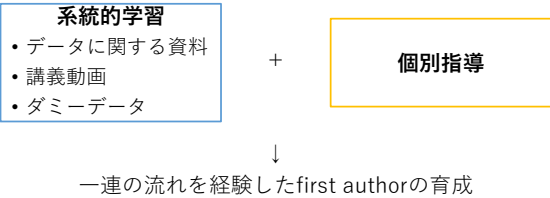
自治医科大学での人材育成

研究の流れ

- ①データ構造の理解
- ②研究デザインの検討
- ③データの切り出し
- ④統計解析
- ⑤論文執筆



研究の流れ



10

要介護状態の高齢者に対する人工呼吸管理の発生率と 1 年後転帰

方法

- Patient 栃木県データ
2014年6月～2018年2月
65歳以上
人工呼吸（挿管下）実施
- Exposure 人工呼吸実施前の要介護状態
- Outcome 1 年以内の死亡、要介護度の変化

13

脊椎圧迫骨折の受傷後の経過

方法

- 栃木県医療・介護レセプトデータ（国保・後期高齢）
- 脊椎圧迫骨折を受傷した65歳以上の患者18,392例
- 受傷後の生存期間、鎮痛薬の処方期間、要介護度の変化を分析

結果

- 1 年死亡率 5.3%
- 鎮痛薬の処方を受けた8,375例では、4か月を過ぎても鎮痛薬を要した患者が22%
- 受傷後に8.2%で要介護度が悪化
受傷前から要介護であることは悪化のリスク因子

16

研究の例

1. 要介護状態の高齢者に対する人工呼吸管理の発生率と 1 年後転帰
2. 脊椎圧迫骨折の受傷後の経過
3. 骨盤臓器脱の要介護状態と治療選択・術後転帰

11

要介護状態の高齢者に対する人工呼吸管理の発生率と 1 年後転帰

結果

• N=4198				
	要介護なし	要支援-要介護1	要介護2-3	要介護4-5
• 死亡	43%	55%	68%	74%
• 要介護度悪化	23%	24%	11%	2%

結論

- 要介護状態にある患者の、人工呼吸管理後の転帰を報告
- 人工呼吸の適応を議論するための資料を提供

14

骨盤臓器脱を有する高齢女性における要介護状態と治療選択および術後転帰

Wada Y et al. *Int J Gynaecol Obstet* 2024 (epub)

背景

- 骨盤臓器脱：骨盤底筋群の機能低下→骨盤内臓器が下垂
- 加齢や身体機能制限に伴い増悪し、保存的加療に抵抗性となる
- 患者の介護状態に応じた治療実態、転帰は不明

Question

- 介護状態別の治療選択や術後転帰の実態は？

17

要介護状態の高齢者に対する人工呼吸管理の発生率と 1 年後転帰

Ohbe H et al. *Critical Care Medicine* 2023;51:584-93.

背景

- 人工呼吸器管理を要する患者の予後は不良で、離脱後にも大きな機能不全を伴う
- 身体機能や認知機能の低下した患者においては、人工呼吸管理の適応の判断は特に困難
- 適応について本人・家族と検討するための資料となるような長期予後のエビデンスは不足

Question

- 要介護状態の高齢者における人工呼吸器管理の発生率と、長期予後・要介護度の変化は？

12

脊椎圧迫骨折の受傷後の経過

Honda A et al. *J Bone Joint Surg* 2024 (epub)

背景

- 脊椎圧迫骨折は骨粗鬆症をベースに生じ、高齢化に伴い増加
- 脊椎の変形や持続する疼痛によりADLを低下させる
- 骨折後の長期経過（予後、鎮痛薬の持続使用、要介護度など）は明らかになっていない

Question

- 脊椎圧迫骨折後の経過と、要介護度増加のリスク因子は？

15

骨盤臓器脱を有する高齢女性における要介護状態と治療選択および術後転帰

方法

- 栃木県医療・介護レセプトデータ（国保・後期高齢）
- 骨盤臓器脱を有する65歳以上の女性3,406例
- 患者の介護状態ごとの治療実態、転帰を分析

結果

- 手術割合： 要介護なし14% vs 要介護あり3.6%
- 術後の尿路感染の割合： 要介護なし8.6% vs 要介護あり37%
- 重篤な合併症は稀

18

自治体によるRWD利活用の現状と課題

現状

- 自治体が保有するリアルワールドデータの研究活用は既に進められている
- 自治体はデータを提供し、研究者が利活用するスキームが中心

課題

- 自治体によるリアルワールドデータ分析と政策への活用
- 研究者の環境（サーバー等）、知識・経験（分析手法等）の共有
- 持続性、人材育成

地域医療構想推進のためのデータ活用

背景

- 高齢化
- 需要↑：医療ニーズの増加（特に入院、介護）
- 供給↓：医療・福祉人材の確保が課題（高齢化、働き方改革）

医療計画

- 都道府県が、医療提供体制の確保を図るために策定
- 期間： 6 年間（第 8 次：2024~2029）
- 医療圏・基準病床数の策定、5疾病6事業の対策、医師確保、等

地域医療構想推進のためのデータ活用

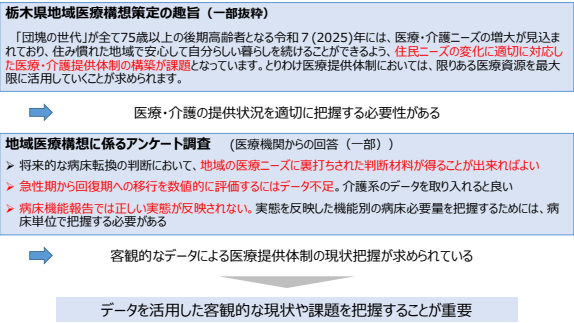
地域医療構想

- 目的：医療機関の機能分化・連携を進め、医療提供体制を確保
- 2025年の「必要病床数」を推計
- 病床機能報告等により、現状と今後の方向性を把握
- 地域医療構想調整会議において協議 → 機能分化・連携
- （都道府県知事による要請・勧告）

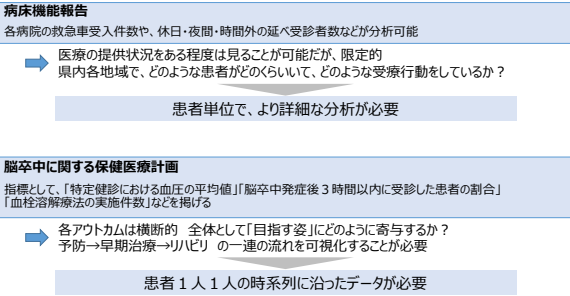
課題

- 調整会議を踏まえた医療機関の自主的な取り組み
- 地域の実情に即した、客観的で有効なデータ分析が必要

栃木県の例



既存のデータ分析の例



地域医療提供体制データ分析チーム構築支援事業

目的

- 都道府県：第8次医療計画、次期地域医療構想の策定のため、医療提供訂正の分析・評価が必要
- データ分析体制の構築が急務
- 地域の感覚とマッチした分析が必要

概要

- 厚生労働省が都道府県を対象にデータ分析チームの構築を支援
- 一部の都道府県からの事例を集積→横展開

地域医療提供体制データ分析チーム：栃木県

概要

- 栃木県と自治医科大学チームを構成、共同で分析を実施
- 各種会議で医療機関に分析結果を報告し意見を聴取（地域医療構想調整会議など）
- 医療の見える化→課題整理→次期地域医療構想の策定
- +
- その他、医療施策の立案
- データ分析に関する県職員の人材育成
- 研究活用

地域医療提供体制データ分析チーム：栃木県

分析の計画

- 5疾病6事業、在宅医療等の提供状況を中心に分析
 - がん、脳卒中、心血管、糖尿病、精神
 - 救急、災害、へき地、周産期、小児、感染症
- 例
 - 市町や医療圏ごとに、県民がどのような病気で、どの医療機関を受診しているか
 - 在宅医療・介護の利用状況
 - 医療圏をまたぐ医療機関への受診
 - 新型コロナウイルスの発生前・感染拡大時・分類位置づけ変更後に、県民の受診行動や医療機関が提供する医療がどう変化したか

地域医療提供体制データ分析チーム：栃木県

活用データ

- 栃木県国民健康保険団体連合会より取得：
 - 国民健康保険、後期高齢者医療制度のレセプトデータ
 - 介護レセプトデータ
 - KDBデータ
- 栃木県内の病院より取得：
 - DPCデータ
- その他：
 - NDBオープンデータなど

分析例：医療圏ごとの入院患者数と流出入

活用データ：DPCデータ

- 医療機関単位での急性期医療を中心とした詳細な診療内容
- 国の調査に参加した医療機関の同意を得て収集
- 県内66病院のうち、64病院のデータを分析
- 様式1・様式3・様式4
- Dファイル・入院EFファイル・外来EFファイル・Hファイル

分析例：医療圏ごとの入院患者数と流出入

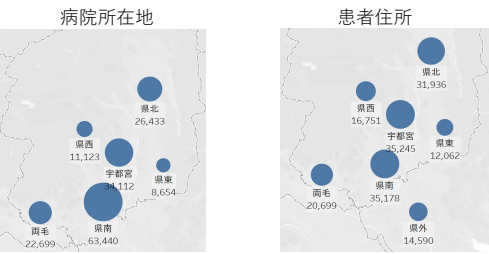
対象

- DPCデータ 様式1
- 2022年4月～2023年3月 県内64病院に入院・退院
- 除外基準
 - 入院時100歳以上
 - 主病名：なし or 疑い病名 or ICD-10でR, V以降
 - 入院経路：院内出生・その他
 - 退院経路：不明・その他
 - 患者郵便番号：欠損・住所特定不能

166,461入院

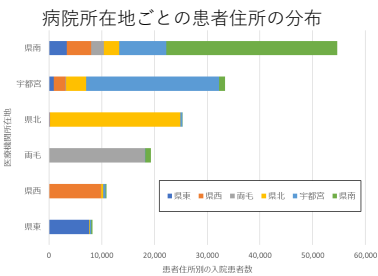
分析例：医療圏ごとの入院患者数と流出入

結果：医療圏別入院患者数



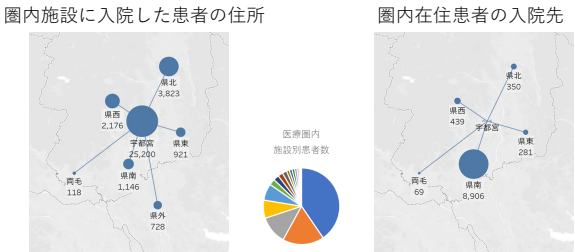
分析例：医療圏ごとの入院患者数と流出入

結果：医療圏別入院患者数



分析例：医療圏ごとの入院患者数と流出入

結果：宇都宮医療圏の例



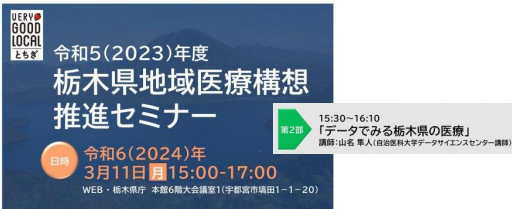
分析チームの取り組み

県各課との連携

- 意見交換会などを実施し、ニーズ・分析テーマを募集
- 例
 - 特定健診： 健診・保健指導と受診・医療費
 - 救急医療： どのような重症度の救急患者がどこに運ばれているか
 - 在宅医療： 医療機関所在地と患者住所
 - がん： 拠点病院の役割、集約化と均てん化
 - 新型コロナウイルス感染症： 受診と入院の実態

分析チームの取り組み

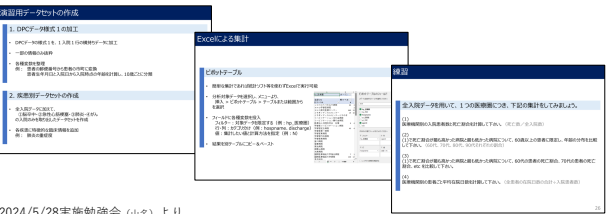
セミナー



- 栃木県ホームページより <https://www.pref.tochigi.lg.jp/e02/2023chiikiyokousouseminar.html>

分析チームの取り組み

県職員向け勉強会



- 2024/5/28実施勉強会 (山名) より

分析中のデータ

DPCデータ

- 患者の重症度など、より詳細な分析
- 様式1 + EFファイル (診療行為) ・Hファイルなど (看護必要度)
- 国民健康保険・後期高齢者医療制度・介護保険レセプトデータ
- 入院 + 外来、医療 + 介護、健診 + 医療
 - 時系列に沿った患者の移動

その他資料との組み合わせ

- 病床機能報告 医療施設調査 人口動態統計 将来人口推計
- 医師・歯科医師・薬剤師調査

研究

戦略的イノベーション創造プログラム（SIP）
「統合型ヘルスケアシステムの構築」
テーマC-2：患者の疾患状態及び施設間動態の可視化を可能とするシステム開発による地域医療構想の実現（研究開発責任者：康永秀生）

概要

- 1) 県レベルの健診・医療レセプト・介護レセプトの連結データベース構築・分析
- 2) 地域レベルのレセプトと他のデータの統合分析
- 3) データサイエンティストおよび臨床疫学研究者の人材育成

↓
データに基づく地域医療分析＋臨床疫学研究

37

自治体によるRWD活用：展望

拡大

- ・佐賀大学と共同研究を開始（SIP）
- ・他の県・大学も募集中

連携

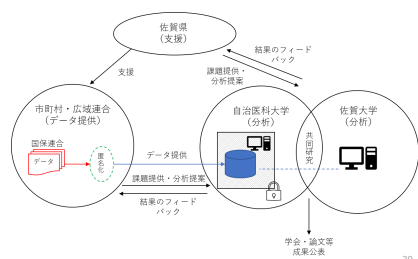
- ・厚労省事業の参加県間： 情報共有・共通化
- ・研究者同士： コンソーシアムを設立
- ・厚労省と

自治体によるRWD活用：構成

必要な体制

- ①県、市町村
- ②地域の大学
- ③データ分析基盤

例：佐賀(SIPで実施中)



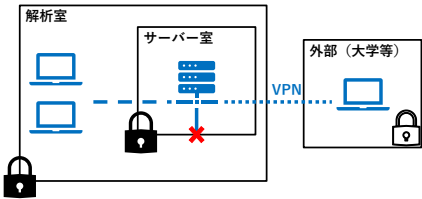
38

39

自治体によるRWD活用：分析基盤

自治医科大学データサイエンスセンターの例

- ・外部から遮断された環境のサーバーでデータを保管
- ・VPN、リモート接続で分析を実施



40

ご清聴ありがとうございました

yamana.hayato@jichi.ac.jp

41

自己対照研究デザイン

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
宮本 佳尚

本日の内容

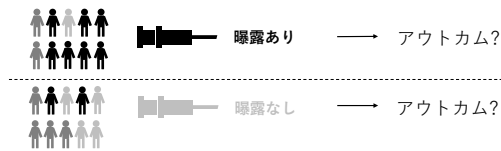
- ・イントロダクション・自己対照研究デザインの概要
- ・ケース・クロスオーバー法
- ・自己対照ケースシリーズ（Self-controlled case series; SCCS）
- ・自己対照研究デザインの前提条件と注意点
- ・さらに学習するために・まとめ

本日の内容

- ・イントロダクション・自己対照研究デザインの概要
- ・ケース・クロスオーバー法
- ・自己対照ケースシリーズ（Self-controlled case series; SCCS）
- ・自己対照研究デザインの前提条件と注意点
- ・さらに学習するために・まとめ

観察研究で因果関係に迫るには？

曝露群・非曝露群でアウトカム発生を比較する（コホート研究）



（弱点）
未測定交絡の問題
適切なコントロール群の設定が難しいことがある

自己対照研究デザインとは

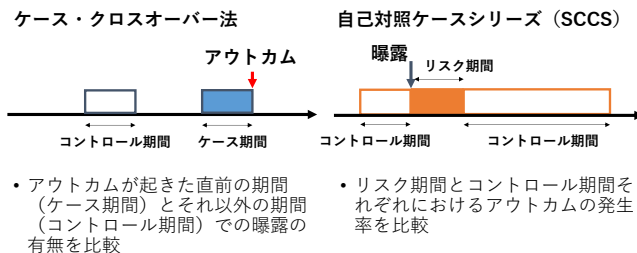
- ・自己対照研究デザインとは、**アウトカムが起こった人のみを対象に、同じ人の異なる時期のデータを比較**することで、曝露とアウトカムの関連に迫る研究デザインである。
- ・自己対照研究デザインでは、**時間で変化しない因子**（性別、遺伝子、生活習慣、社会経済的背景）は自動的にキャンセルアウトされる。

一般的なコホート研究（やケースコントロール研究）で問題となる**未測定交絡の問題**を克服できる
→曝露とアウトカムの関連を**多元的に**評価する一助となる

コホート研究と自己対照研究デザインの比較

コホート研究	自己対照研究デザイン
・ 曝露のあった人となかった人	・ アウトカムが起きた人のみ
・ 上記の 群ごと のアウトカム発生を比較	・ ある特定の興味ある 期間 とそれ以外の 期間 を比較
・ “ 誰が？ ”アウトカムを発生しやすいのか？	・ “ いつ？ ”アウトカムが発生しやすいのか？

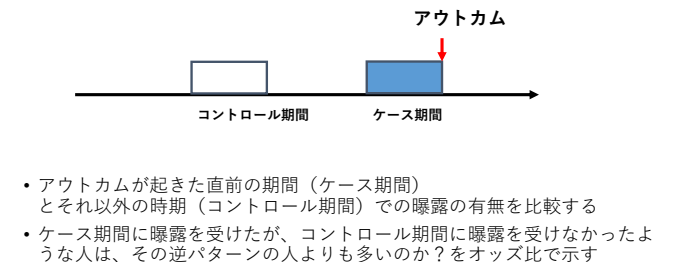
代表的な自己対照研究デザイン



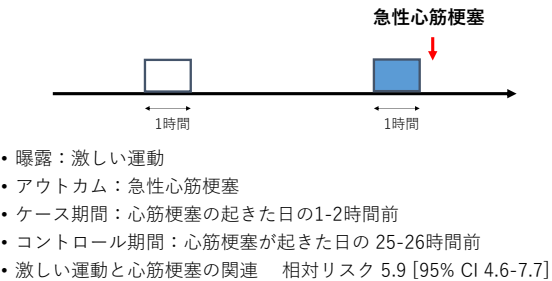
本日の内容

- ・イントロダクション・自己対照研究デザインの概要
- ・ケース・クロスオーバー法
- ・自己対照ケースシリーズ（Self-controlled case series; SCCS）
- ・自己対照研究デザインの前提条件と注意点
- ・さらに学習するために・まとめ

ケース・クロスオーバー法



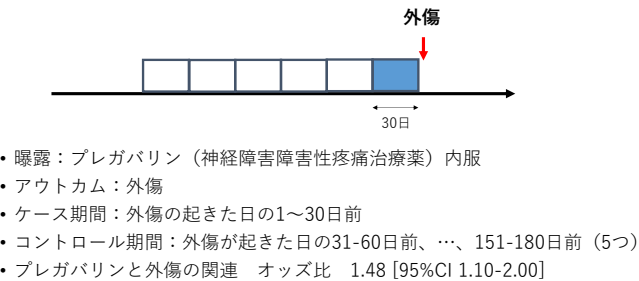
ケース・クロスオーバー法 具体例①
激しい運動と急性心筋梗塞



Mittleman MA, et al. N Engl J Med. 1993;329:1677-83

10

ケース・クロスオーバー法 具体例②
プレガバリンと外傷



Miyamoto Y, et al. Pharmacoepidemiol Drug Saf. 2020;29:558-564.

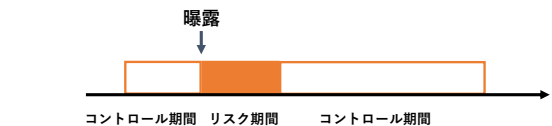
11

本日の内容

- イントロダクション・自己対照研究デザインの概要
- ケース・クロスオーバー法
- 自己対照ケースシリーズ（Self-controlled case series; SCCS）
- 自己対照研究デザインの前提条件と注意点
- さらに学習するために・まとめ

12

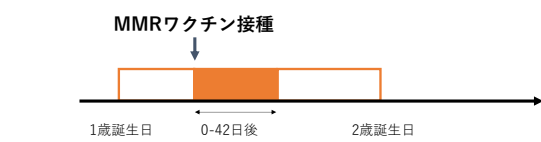
自己対照ケースシリーズ



- リスク期間とコントロール期間それぞれにおけるアウトカムの発生率を比較する
- コントロール期間における発生率を基準とした、リスク期間におけるアウトカムの発生率比=incidence rate ratioが求まる
- リスク期間の長さや曝露からの期間を仮説に基づき柔軟に設定できる

13

自己対照ケースシリーズ 具体例①
MMRワクチンとITP

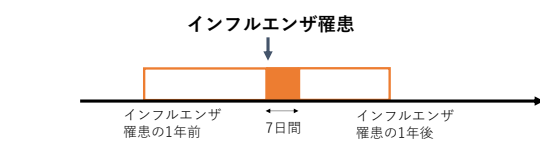


- 曝露：MMR（麻疹・流行性耳下腺炎・風疹）ワクチン接種
- アウトカム：ITP（特発性血小板減少性紫斑病）
- リスク期間：MMRワクチン接種 42日間
- コントロール期間：1-2歳のリスク期間以外の期間
- リスク期間におけるアウトカムの発症率比：3.23 [95% CI 1.53-6.79]

Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.

14

自己対照ケースシリーズ 具体例②
インフルエンザ罹患と急性心筋梗塞



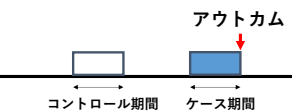
- 曝露：インフルエンザ罹患
- アウトカム：急性心筋梗塞
- リスク期間：インフルエンザ罹患から7日間
- コントロール期間：リスク期間以外の期間
- リスク期間におけるアウトカムの発症率比：6.05 [95% CI 3.86-9.50]

Kwong JC, et al. N Engl J Med 2018;378:345-53.

15

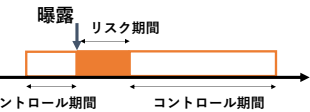
ケースクロスオーバー法とSCCSの比較

ケースクロスオーバー法



- アウトカムが起きた直前の期間（ケース期間）とそれ以外の時期（コントロール期間）での曝露の有無を比較

自己対照ケースシリーズ（SCCS）

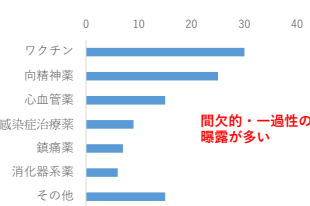


- リスク期間とコントロール期間それぞれにおけるアウトカムの発生率を比較

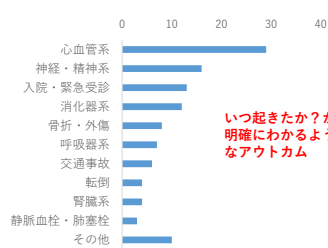
16

電子医療記録データベースを用いた
自己対照研究デザインの研究（2011-2014に出版）

• 曝露のタイプ



• アウトカムのタイプ



Gault N, et al. BMC Med Res Methodol. 2017 8;17:25. 150 作成
(Under a Creative Commons BY 4.0)

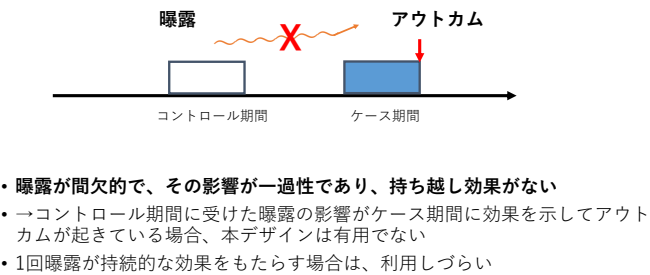
17

本日の内容

- イントロダクション・自己対照研究デザインの概要
- ケース・クロスオーバー法
- 自己対照ケースシリーズ（Self-controlled case series; SCCS）
- 自己対照研究デザインの前提条件と注意点
- さらに学習するために・まとめ

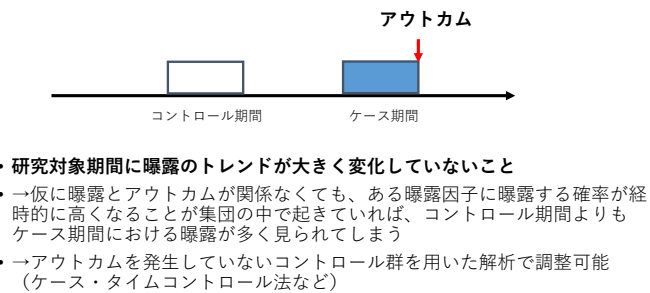
18

ケース・クロスオーバー法的前提条件①



19

ケース・クロスオーバー法的前提条件②



20

SCCS的前提条件①

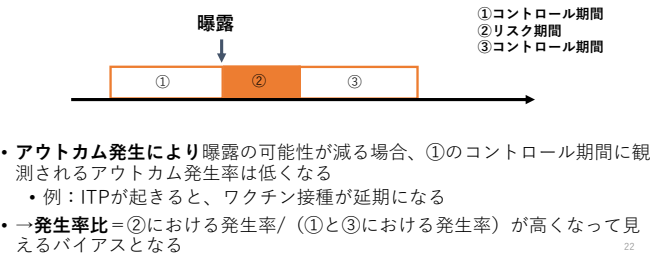
- 反復しないアウトカムの場合、その確率が稀であること
 - 一般にアウトカムが発生した集団全体における、当該アウトカムの累積発症割合が10%未満であればよい
- 反復するアウトカムの場合、それぞれが独立であること
 - 反復するアウトカムがそれぞれ独立でない場合、1回目のアウトカムのみを対象とした解析で対処できる

21

Whitaker HJ, et al. Biom J.2018;60:1110-1120.

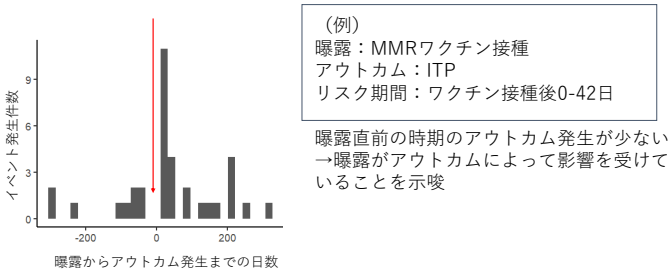
SCCS的前提条件②

- アウトカム発生は曝露に影響を及ぼさない



22

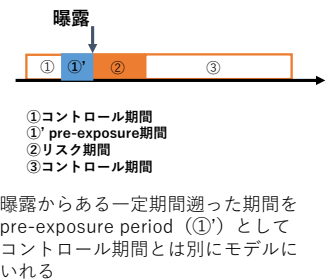
SCCS的前提条件②



Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.
図は<https://sccs-studies.info/index.html>で提供されるデータから独自に作成

23

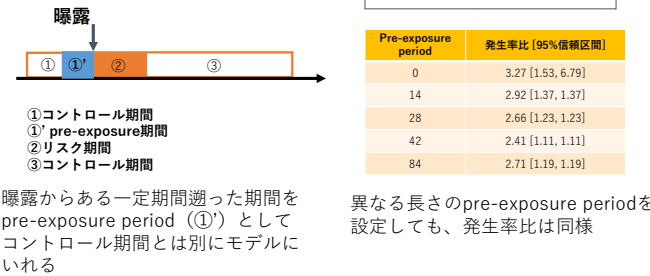
SCCS的前提条件②：対処法



24

Whitaker HJ, et al. Stat Med.2018;37:643-658.

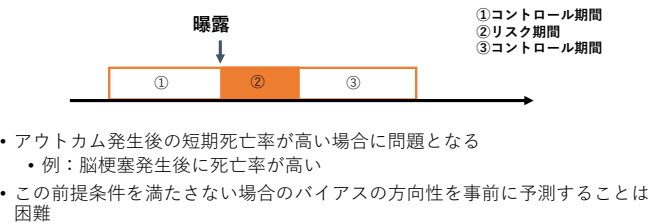
SCCS的前提条件②：対処法



25

SCCS的前提条件③

- アウトカム発生はその後の追跡期間に影響を及ぼさない

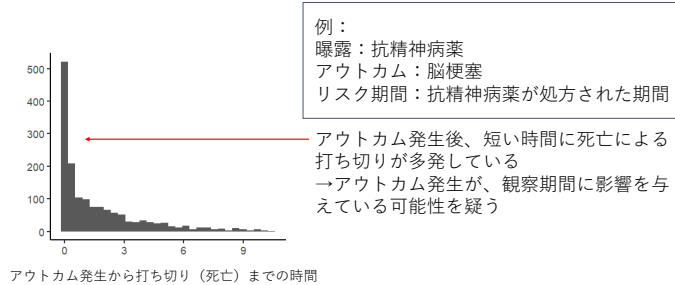


151

Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.

26

SCCS的前提条件③



27

Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.
図は<https://sccs-studies.info/index.html>で提供されるデータから独自に作成

SCCSの前提条件③：対処法

- 死亡した症例を除外する
 - 死亡した症例においても、観察期間をほかの症例と同様に一定にした解析を行う（死亡のタイミングで打ち切りとしない）
- などの方法が提案・利用されている

Whitaker HJ, et al. Stat Med.2018;37:643-658.
Farrington P, et al. Self-Controlled Case Series Studies: A Modelling Guide with R. 2018.

本日の内容

- イントロダクション・自己対照研究デザインの概要
- ケース・クロスオーバー法
- 自己対照ケースシリーズ（Self-controlled case series; SCCS）
- 自己対照研究デザインの前提条件と注意点
- さらに学習するために・まとめ

さらに学習するために

- Zhang, **Case-crossover** design and its implementation in R, Ann Transl Med 2016;4(18):341
 - Rでのケース・クロスオーバー法の実装
- Farrington P, et al. **Self-Controlled Case Series Studies: A Modelling Guide with R.** 2018.
 - SCCSの開発者のFarringtonらによる、発展的な方法まで含んだ本。
- <https://sccs-studies.info/index.html>
 - SCCSの開発者のFarringtonらによる、ウェブサイト。R/Stata/SASコードが提供されている

本日のまとめ

- 自己対照研究デザインとは、**アウトカムが起こった人のみ**を対象に、**同じ人の異なる時期のデータ**を比較することで、曝露とアウトカムの関連に迫る研究デザインである。
- 自己対照研究デザインでは、**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。
- 自己対照研究デザインの代表的な例として、**ケース・クロスオーバー法、自己対照ケースシリーズ**がある。
- 自己対照研究デザインの特徴的な**前提条件**に留意する。

時間依存性交絡

東京大学大学院医学系研究科 臨床疫学・経済学
東北大学病院 高度救命救急センター
大邊寛幸

1

時間依存性交絡・治療の有名な論文 -HIVとZidovudine-

HIV陽性の男性に対するZidovudineは死亡を減らすか？

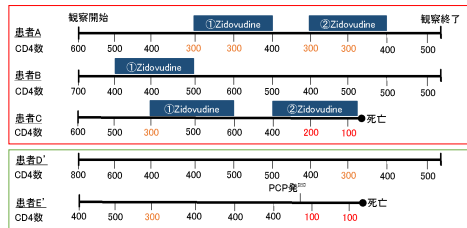
Design 観察研究
Patient AIDS未発症の、HIV陽性男性
Exposure Zidovudineの投与
Control Zidovudineの非投与
Outcome 死亡

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

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は死亡率をリスク比3.6倍上昇させる

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

目次

- 時間依存性交絡と時間依存性治療について
- 時間依存性交絡・治療の対処法 (G-methods)
G-formula, IP-weighting, G-estimation
- G-methodsに必要な仮定
- G-methods (IP-weighting) の具体例の紹介

2

HIV陽性の男性に対するZidovudineは死亡を減らすか？

- 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
→ Zidovudine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
→ Zidovudine治療を開始する
- AIDS発症している場合
→ Zidovudine治療を開始するが、カリニ肺炎などが
重篤な場合はその治療を優先する場合がある

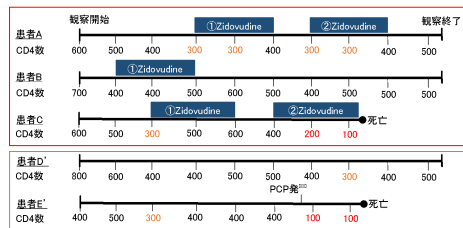
抗HIV治療ガイドラインより

5

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

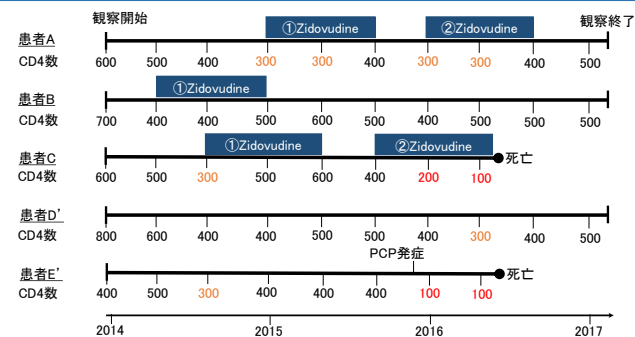
Hernán MA, Brumback B, Robins JM. Epidemiology. 2000;11(5):561-70.

目次

- 時間依存性交絡と時間依存性治療について
- 時間依存性交絡・治療の対処法 (G-methods)
G-formula, IP-weighting, G-estimation
- G-methodsに必要な仮定
- G-methods (IP-weighting) の具体例の紹介

3

HIV陽性の男性に対するZidovudineは死亡を減らすか？



6

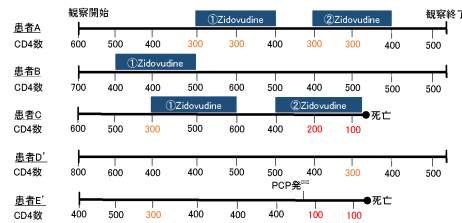
標準的な統計解析ではバイアスを生じる

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)

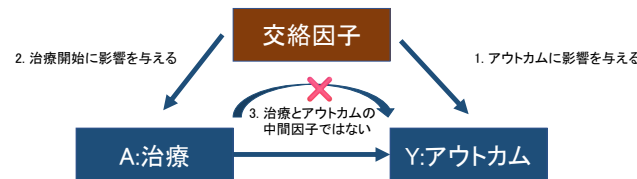
観察期間中に何度も測定し、時間に依存して変化していく変数



10

交絡因子(Confounder)

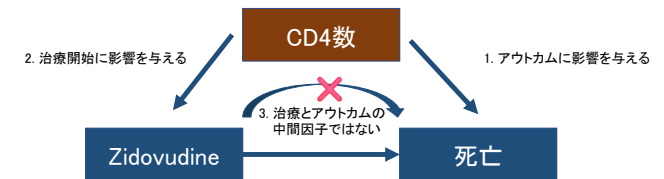
興味あるアウトカム(例えば死亡)のリスク因子(予測因子)かつ、治療開始に影響する(予測する)変数で中間因子でない



11

交絡因子(Confounder)

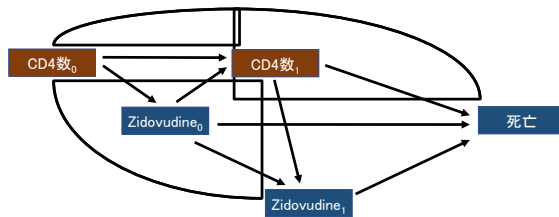
CD4数はZidovudineの開始を予測し、かつ死亡のリスク因子であるので交絡因子である



12

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

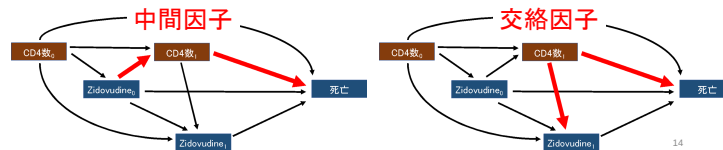


13

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

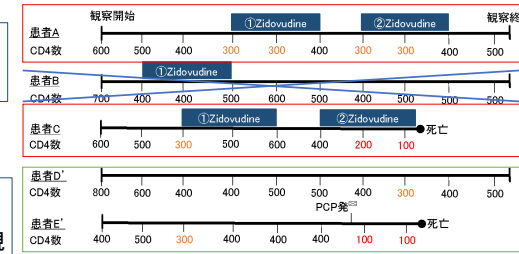


14

時間依存性交絡の問題②選択バイアス

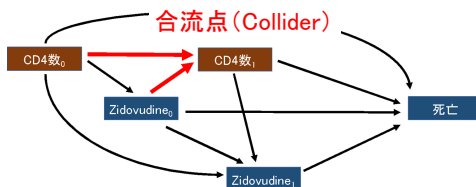
CD4数が300以下の患者のみ含める

患者Bのように初回Zidovudine効果を無視



15

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) と呼ぶ

16

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

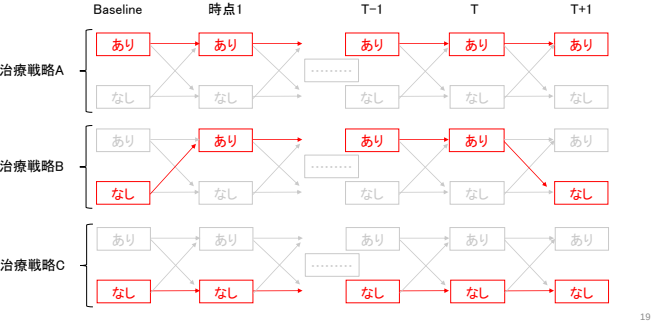
BMJ. 2017 Oct 16;359:j4587.
Epidemiology. 2009;20(4):488-95.
Epidemiology. 2003;14(3):300-6.

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)
あらかじめその時ごとの治療戦略が決定している
“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

18

時間依存性治療(Time varying treatment)



19

時間依存性治療(Time varying treatment)



20

目次

- 1. 時間依存性交絡について
- 2. 時間依存性交絡・治療の対処法(G-methods)
IP-weighting, G-formula, G-estimation
- 3. G-methodsに必要な仮定
- 4. G-methods (IP-weighting) の具体例の紹介

21

時間依存性交絡の対処法(G-methods)

G-methods

- 1. IP Weighting $E[Y^a] = \beta_0 + \beta_1 a \quad W^A = \frac{1}{f(A|L)} \text{ or } SW^A = \frac{f(A)}{f(A|L)}$
- 2. G-formula $E[Y^{a=1}] = \sum_{l=1,0} E[Y|L=l, A=1] \times \Pr[L=l]$
- 3. G-estimation $Y^a - Y^{a=0} = \psi_1 a \quad \text{logit } \Pr[A=1|L, Y^{a=0}] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$

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

25

時間依存性交絡の対処法(G-methods)

G-methods

- 1. IP Weighting ... 治療に対する重み付け擬似集団
- 2. G-formula ... 未来のアウトカムを予測
- 3. G-estimation ... 介入の効果を推定

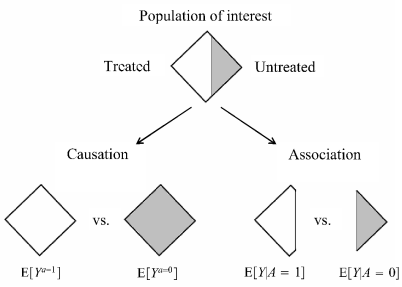
Int J Epidemiol. 2017 01;46(2):756-62.

24

重要用語の説明

- 因果推論
- 疑似集団
- 潜在アウトカム

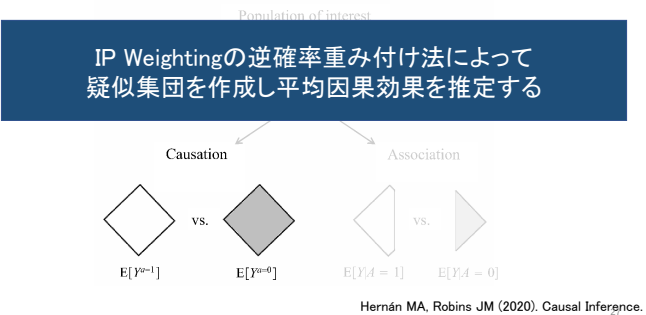
因果推論



時間依存性交絡の対処法の比較

	利点	欠点
IP Weighting	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-estimation	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

因果推論



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”?”を知りたい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。

28

G-methodsに必要な仮定

- 1. Sequential Conditional Exchangeability
 - 未測定の時空間依存性交絡因子が無い
- 2. Consistency -治療方法は単一であり、かつその効果は単一である
- 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
- 4. No misspecification of the model -使用したモデルが正しい

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

31

G-methodsに必要な仮定

- 1. Sequential Conditional Exchangeability
 - 未測定の時空間依存性交絡因子が無い
- 2. Consistency -治療方法は単一であり、かつその効果は単一である
- 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
- 4. No misspecification of the model -使用したモデルが正しい

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

34

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。

29

G-methodsに必要な仮定

- 1. Sequential Conditional Exchangeability
 - 未測定の時空間依存性交絡因子が無い
- 2. Consistency -治療方法は単一であり、かつその効果は単一である
- 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
- 4. No misspecification of the model -使用したモデルが正しい

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

32

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

目次

- 1. 時間依存性交絡と時間依存性治療について
- 2. 時間依存性交絡・治療の対処法(G-methods)
 - G-formula, IP-weighting, G-estimation
- 3. G-methodsに必要な仮定
- 4. G-methods (IP-weighting) の具体例の紹介

30

Sequential Conditional 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

33

G-methodsに必要な仮定

- 1. Sequential Conditional Exchangeability
 - 未測定の時空間依存性交絡因子が無い
- 2. Consistency -治療方法は単一であり、かつその効果は単一である
- 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
- 4. No misspecification of the model -使用したモデルが正しい

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

36

No misspecification of the model

- IP WeightingやG-formulaに使用したモデルが正しいこと
- IP WeightingとG-formulaはalgebraically equivalent(代数的に同値)、つまり見た目は違っても同じ結果を示す
- Doubly RobustではIP WeightingとG-formulaのどちらかのモデルが正しければ良いという拡張性がある

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

38

(再掲)

HIV陽性の男性に対するZidovudineは死亡を減らすか？

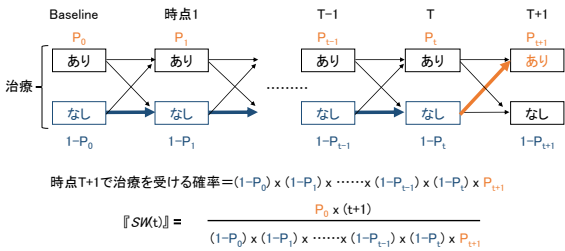
Design	観察研究
Patient	AIDS未発症の、HIV陽性男性
Exposure	Zidovudineの投与
Control	Zidovudineの非投与
Outcome	死亡

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

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



43

目次

1. 時間依存性交絡と時間依存性治療について
2. 時間依存性交絡・治療の対処法 (G-methods)
G-formula, IP-weighting, G-estimation
3. G-methodsに必要な仮定
4. G-methods (IP-weighting)の具体例の紹介

1. 治療に対する逆確率の重み(IPTW)を作成する

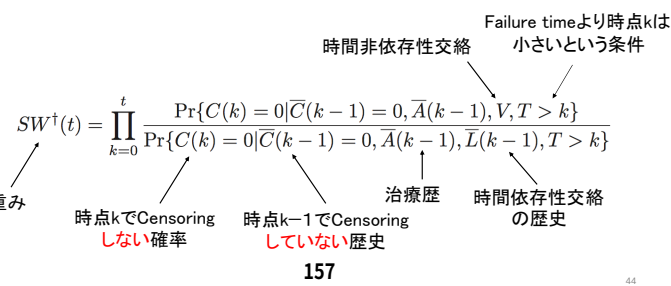
1.1. 治療の重み (treatment weight) の作成

時間非依存性変数	時間依存性変数
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)	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

41

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



44

IP-weightingの実際の手順

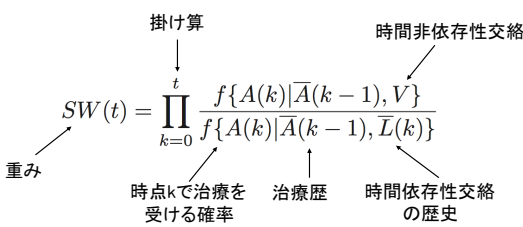
1. 治療に対する逆確率の重み(Inverse Probability Treatment Weight: IPTW)を作成する
 - 1.1. 安定化された重み (Stabilized treatment weight) の作成
 - 1.2. 打ち切りの重み (censoring weight) の作成
 - 1.3. 最終的な重み (overall weight) の作成
2. IPTWで重み付けしたアウトカム-治療回帰モデルを適用する

The Stata Journal 2004;4:402-420.

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

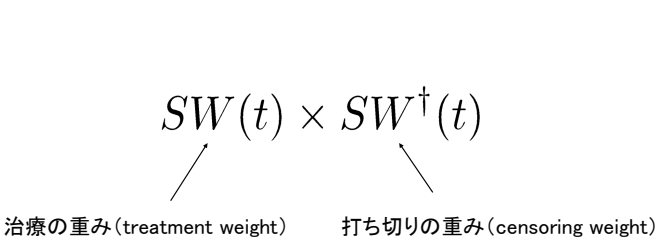
治療を受ける時点ごとに傾向スコアを算出し、最終的に掛け算



42

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



45

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)

46

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G-methods (IP Weighting, G-formula, G-estimation)
- G-methodsは"Sequential Conditional Exchangeability", "Consistency", "Positivity", "No misspecification of the model"を仮定

49

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(overall weight)』により、擬似集団を作り出す
- その擬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

47

まとめ

- IP Weightingにおいて擬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、重み付き回帰モデルを適用し、集団全体の因果効果を推定する

50

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)
IP Weighting 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.

機械学習の疫学応用

東京大学大学院医学系研究科臨床疫学・経済学教室
松居宏樹

この資料は、

- 機械学習の疫学応用に関する資料です。
- 以下を解説します。
 - 機械学習を使った疫学研究の例
 - 具体的実装方法を提示
 - 機械学習を疫学に用いるモチベーションは何か？

機械学習とは

- **機械学習は、コンピューターに明示的なプログラミングなしで学習能力を与えるための学問分野（Arthur Samuel, 1959）**
- 大規模データを、様々なアルゴリズムに基づいてコンピューター（機械）に学習させ、その学習結果をモデル（特徴・規則・関数）として取り出す。
- データをInputすれば、モデル（f(x)）に基づいて、結果(y)が出力される。そのプロセスを機械が自動で気に行えるようにする。

疫学研究における機械学習の利用

- 学習結果として得られたモデルを疫学の枠組みの中でどう使うか？
- 疫学分野においては、（多くの場合）学習手法自体に興味があるわけではない。

- 良く用いられる疫学研究におけるテーマ
 - 予測
 - クラス分け
 - 因果推論

身近な機械学習の例

- 一般線型モデル（線形回帰）
$$Y = X\beta + \epsilon$$
- 一般化線型モデル（ロジスティック回帰モデル）

$$\text{logit}(P(Y = 1)) = \log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = X\beta + \epsilon$$

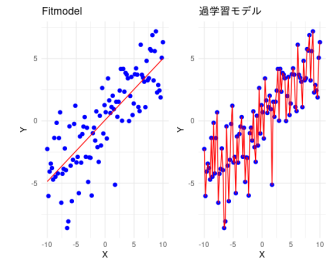
いずれも、結果（ $Y, \text{logit}(P(Y=1))$ ）をモデル（ XB の足し算）で表している。
ここで X はデータ（の中の変数）であり、 β はその係数。

機械学習と過学習（Over fitting）

```
x<- seq(-10, 10, length.out = 100)
y<- 0.5 * x + rnorm(100, sd = 2)
z<- as.character(x)

# 過学習モデルの作成
fit_model <- lm(y~x)
overfit_model <- lm(y~z)
```

与えられたデータに過剰にフィットすれば、誤差は小さくなるが、新しいデータに対する予測能力（汎化性能）がなくなる。



Over Fitting と Under Fitting

$$\mathbb{E}[(\hat{f}(x) - f(x))^2] = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + (\mathbb{E}[\hat{f}(x)] - f(x))^2$$

データから推計したモデル $\hat{f}(x)$ と真のモデル $f(x)$ の誤差（左辺）は
モデル自体のバラつき（右辺第一項：バリエーション）と
真のモデルとデータから推計したモデルの間のずれ（右辺第二項：バイアス）
に分解できる。

Over Fitting: バイアスが小さくなったが、バリエーションが大きくなった状態
Under Fitting: バリエーションが小さくなったが、バイアスが大きい状態

Over Fitting も Under Fitting もしていない、いい塩梅を機械的に探す。

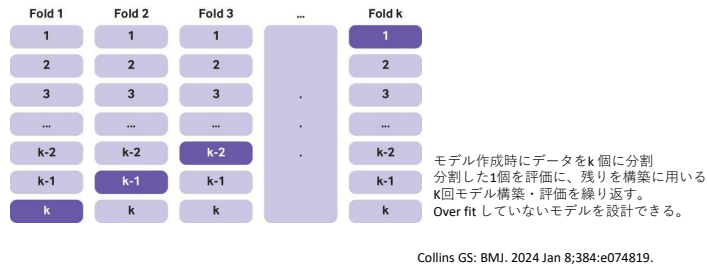
過学習を防ぐ工夫

- モデルの複雑さを制限する（例：パラメータ数を減らす）
- **クロスバリデーションを使用してモデルを作る。**
- 発表時は**バリデーション**を行う。
- 機械学習手法として確立した過学習防止プロセスを活用
 - 正則化（学習時にペナルティをかけることで過学習を防ぐ）
 - Random Forest（データを複数に分割し学習することで過学習を起こしにくい）

Cross Validation

- モデル構築プロセスにおいて、**Over fitting** を避けたい。
- モデルを様々な切り分けて、データセットを作る。
- モデルに必要なパラメータや変数選択などをいろいろ変化させて、モデル構築・評価を繰り返す。
- **Over fit** しない、パラメータや変数群を見つけ出す。

(K-fold) Cross validation



Collins GS: BMJ. 2024 Jan 8;384:e074819.

Validation

- モデル開発データ（Derivation data）とは異なるデータ（Validation data）でモデルの評価を行いOver fitting の影響を除いたモデルの評価を行う。
- モデルの評価に用いる評価指標
 - C-statistics(ROCAUC), PR-AUC, accuracy, Sensitivity, Specificity, PPV, NPV
- Validation プロセス
 - Sample split validation
 - Bootstrapping validation

Sample split validation

- データを予めDerivation と Validation 分割しておく。
 - Random: ランダムにデータを分割する。
 - Non-random: 時間軸・地理空間軸・施設軸でデータを分割する。特定疾患集団などでのモデル精度を報告することもある。
 - 予めモデル構築に使用しないデータを用意することでOver Fitの影響を除いた評価が可能
- 批判
- 小集団の場合、モデル構築時のサンプルサイズが減りかえってOver Fitを増やすのでは？
 - 大集団の場合、そもそもランダムに分割する意義が低いのでは？

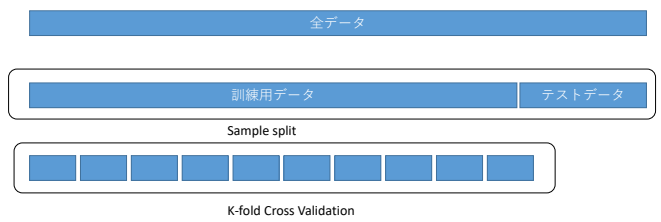
Collins GS: BMJ. 2024 Jan 8;384:e074819.

Bootstrap validation

- 予め全データ(original data)でモデルを構築検証(apparent performance) する。
- Bootstrap data を作成する。
- Bootstrap dataでモデルの構築を行い、bootstrap data での検証（bootstrap performance）とoriginal data での検証（test performance）を繰り返す。
- bootstrap performance – test performance でover fitting 度合い（optimism）を算出し、平均を得る(estimated optimism)。
- apparent performance - estimated optimism でcorrected estimate of performanceを得る。

Collins GS: BMJ. 2024 Jan 8;384:e074819.

多く用いられる実行プロセス



多く用いられる実行プロセス

- 訓練用データにCross validation を適用しモデルを設計→訓練用データでモデル構築→テストデータでのモデル精度報告
- 全データを使ってモデル構築→Bootstrap validation でのモデル精度を報告
- 訓練用データを使ってモデル構築→テストデータでのモデル精度を報告

機械学習を利用する際知っておくべき事

- モデルの評価
 - 識別能の評価
 - ROC, ROCAUC, Accuracy, Sensitivity, Specificity, PPV, NPV, F1-score,
 - 較正曲線(Calibration curve)
 - Net benefit
 - Decision curve
- TRIPOD+AI statement
 - Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378.

機械学習を利用する際知っておくべき事

- Imbalanced data
 - 予測すべきアウトカムが不均衡であるデータ
 - 適切な学習をしないと、意味のないモデルができる。
 - 死亡率が0.1%のデータなら、“全て生存”と回答するモデルの精度は99.9%
 - データ重みづけ
 - Under Sampling, Over sampling, SMOTE
- Sparse data
 - テーブルが非常に多次元だが、その項目のほとんどがNullのデータ
 - 縦持ちデータとして保存しているデータを横展開するとできる。
 - メモリ効率が非常に悪いので、適切なハンドリングが必要
 - Sparse matrix

疫学研究における機械学習

- Okada A, Hashimoto Y, Goto T, Yamaguchi S, Ono S, Ikeda Kurakawa K, et al. 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. Diabetes Care. 2022 Jun 2;45(6):1346–54.
- Konishi T, Goto T, Fujiogi M, Michihata N, Kumazawa R, Matsui H, et al. New machine learning scoring system for predicting postoperative mortality in gastroduodenal ulcer perforation: A study using a Japanese nationwide inpatient database. Surgery. 2022 Apr;171(4):1036–42.
- Matsui H, Yamana H, Fushimi K, Yasunaga H. Development of Deep Learning Models for Predicting In-Hospital Mortality Using an Administrative Claims Database: Retrospective Cohort Study. JMIR Med Inform. 2022 Feb 11;10(2):e27936.

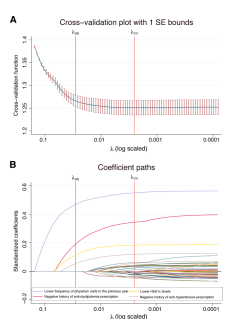
Lasso 回帰を用いた変数選択

1. Okada A, Hashimoto Y, Goto T, Yamaguchi S, Ono S, Ikeda Kurakawa K, et al. 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. Diabetes Care. 2022 Jun 2;45(6):1346–54.

方法

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

Lasso 回帰のクロスバリデーション



Okada A: Diabetes Care. 2022 Jun 2;45(6):1346–54.

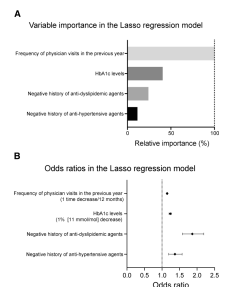
背景

- 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

Lasso 回帰

- 線型回帰を考える。
$$Y = X\beta + \epsilon$$
- 前述の通り、Xをたくさん入れればバイアスは減るがバリエーションは増える (Over Fit する。)
- Xの数を多くしすぎない→Xの数が増えることに罰則を付ける。

結果



Okada A: Diabetes Care. 2022 Jun 2;45(6):1346–54.

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(50個以下)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:
 - trainデータ (80%)とtestデータ (20%)に分割
 - 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を適用)
 - testデータを用いて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

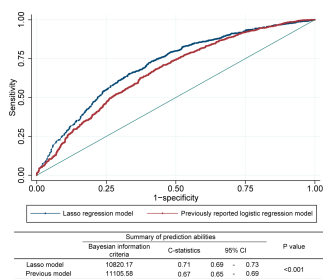
Lasso回帰

- 回帰モデルは以下の誤差を最小にしたい。
$$J = \sum_{i=1}^N \left(y_i - \sum_{j=1}^n x_{ij} \beta_j \right)^2$$
- ここで、Xが増えると、JにペナルティがついてJが増えるように以下の右辺2項を追加してJ'を最小化するようにする。
$$J' = \sum_{i=1}^N \left(y_i - \sum_{j=1}^n x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^n |\beta_j|$$
- λが大きいと、罰則がきつくと、変数が減りやすくなる。
- λをクロスバリデーションを使って決定。

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。

Okada A: Diabetes Care. 2022 Jun 2;45(6):1346–54.



Okada A: Diabetes Care. 2022 Jun 2;45(6):1346–54.

スコア作成 + ほかの機械学習

2. Konishi T, Goto T, Fujiogi M, Michihata N, Kumazawa R, Matsui H, et al. New machine learning scoring system for predicting postoperative mortality in gastroduodenal ulcer perforation: A study using a Japanese nationwide inpatient database. *Surgery*. 2022 Apr;171(4):1036–42.

背景

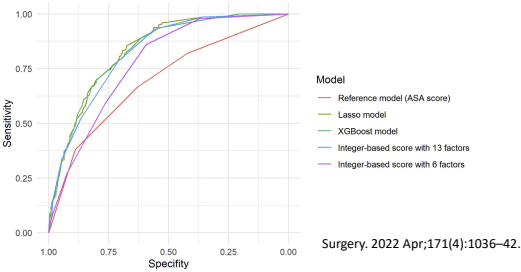
胃十二指腸潰瘍穿孔は一般的な外科的救急疾患である。術後死亡予測としてASA (American Society of Anesthesiology) スコア等が知られる。しかし、それらの予測能は不十分である(Anbalakan K et al. *Int J Surg* 2015; 14: 38–44.)。

- 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

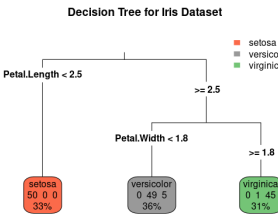
方法

- 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

結果



決定木を用いた機械学習



決定木を用いた機械学習

- Random forest (Bagging)
 - データセットから、Bootstrap サンプルを作成
 - 変数をランダム選択した複数のサブセットを作成。
 - サブセットを基に複数の決定木を作成する。
 - 複数の決定木の平均を基に予測を行う。
- XGBoost (Boosting)
 - データセットを基に、初期モデルを作成
 - 初期モデルの正解との誤差を計算
 - 誤差が小さくなるように（間違えた選択に着目し）、新モデルをUpdate
 - 誤差の計算→Updateを繰り返す

深層学習による高精度リスクスコアの構築と妥当性検証

3. Matsui H, Yamana H, Fushimi K, Yasunaga H. Development of Deep Learning Models for Predicting In-Hospital Mortality Using an Administrative Claims Database: Retrospective Cohort Study. *JMIR Med Inform*. 2022 Feb 11;10(2):e27936.

背景

- 深層学習を応用し、高精度なリスクスコアを構築する。
- 疫学研究におけるリスク調整に利用する。
- DPCデータに含まれる情報のうち、以下の情報のどちらか一方または両方を用いたモデルを構築し、その精度を比較した。
- 入院時の診断名と入院当日に実施された全診療行為（請求情報）
 - 急性心筋梗塞、心不全、脳卒中、肺炎について様式1に含まれる特異的な重症度指標

方法

- 対象
- 2010年7月1日-2017年3月31日退院症例
 - Derivation コホート（95％）とValidation コホート（5％）
- 除外基準
- 入院当日に退院した患者をValidation コホートから除外した。
- 変数
- Outcome: 院内死亡
 - 予測変数
 - 患者の背景情報（年齢、性別、入院前180日間の入院歴）
 - 入院時のICD-10ベースの全診断名
 - 入院当日に行われた全診療行為情報
 - 疾患重症度の情報:
 - 急性心筋梗塞：Killip 分類
 - 心不全：New York Heart Association 分類
 - 脳卒中：Barthel index, Japan Coma Scale (JCS)
 - 市中肺炎：A-DROPスコア

方法

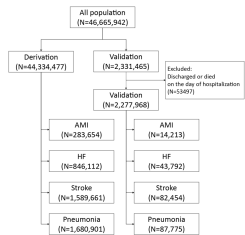
- 全患者の院内死亡を予測するメインモデル
 - 疾患重症度以外の情報を予測因子に利用
- ネットワーク構造
 - 4つの隠れ密層を持つ9層からなる深層学習モデル
 - 各レイヤーは1000ノード
 - 出力層には2ノードのソフトマックス層
 - 20%のドロップアウト層を挟み込む
- 学習
 - 死亡例と生存例の数が大きく異なるため、死亡率の逆数で重み付け
 - Derivation コホートの3%のデータを使った学習ステップを含む早期停止手順
 - Loss Functionは sparse categorical crossentropy を利用
 - 確率的勾配降下法 (stochastic gradient descent:SGD)

方法

- 患者背景情報に加え、以下の疾患重症度を用いて各疾患にモデルを構築
- メインモデルと疾患特異的モデルの予測能力の比較
- 急性心筋梗塞
 - Killip分類
 - 心不全
 - New York Heart Association分類
 - 脳卒中
 - 入院時のBarthel index
 - Japan Coma Scale
 - 肺炎
 - A-DROPスコア

結果

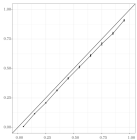
- 46,665,942症例を対象とした
 - Derivation コホート (n=44,334,477)
 - Validation コホート (n=2,331,465)
- 除外基準に従ってValidation コホートの患者を除外し、2,277,968人をValidation コホートとした



結果

Validation of main model

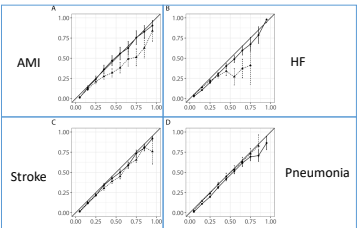
	AUC(95% CI)	Threshold	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV(95% CI)
Main model	0.954 (0.954-0.955)	0.0435	0.920 (0.915-0.924)	0.855 (0.852-0.860)	0.195 (0.192-0.199)	0.996 (0.996-0.997)



メインモデルは高い判別能を示した。観察死亡率と推定死亡率は強い相関があったが、推定死亡率は観察死亡率よりわずかに低かった。

結果

- Validation of disease specific models



AMI, HF, Stroke では、疾患特異モデルより高い相関
Pneumonia では、予測死亡率0.8以上で乖離

考察

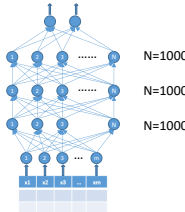
- 深層学習による高精度リスクスコアの構築
 - DPCデータベースから院内死亡率予測モデルを構築
 - 入院初日の患者背景、診断名、診療行為をモデルに投入
 - AUCによる判別能力：全体的に高く、特定疾患サブグループでも優れていた
 - 一般的な疾患重症度指標を用いた疾患別モデルよりも優れた判別能力
 - Calibration curve により、予測死亡率と実死亡に高い相関
- 本研究で開発したモデルは、診療報酬請求データを用いた臨床疫学研究におけるリスク調整に有用と考えられた。

深層学習

- モデル構造が、従来の機械学習モデルより複雑になった。
- 入力（データ）を受け、隠れ層を通り、出力（予測）を行う。
- 隠れ層は巨大な行列計算（データを受け取り、Weightをかけて次に渡す）
- モデル構造は層の接合の仕方や、隠れ層内の処理で様々な構造が作れる。

モデル構造の例

- 4層の全結合中間層(ReLU)を有する多層ニューラルネットワーク
- 入力層は49297ノード
- 各中間層は1000ノードで20%のDrop Outを設定
- 実装はPython 上で行い学習にはTensorFlowを用いた。



機械学習を用いるモチベーション

- 高い精度を求めている？
 - 深層学習はLogistic Regressionより予測能が高いか？
 - 200変数くらいのInput の範囲ではLogistic Regression と各種機械学習モデルに差なし(Christodoulou : Journal of Clinical Epidemiology. 2019 Jun;110:12–22.)
- 変数選択が自動で行われる？
 - Lasso 回帰で決定された係数に一致性（サンプルサイズを無限に大きくなると、母数と推定値が一致する特性）は“ない”。
 - 選択された係数はあくまで予測に寄与した変数にすぎない。

機械学習を用いるモチベーション（私見）

- 高次元データを疫学で利用可能な次元に落とし込みたい。
 - 49297次元あるレセプト情報を、深層学習を使って死亡リスクスコア（1次元）に
 - 77 364次元ある、レセプトと様式1 情報を、分散表現を使って200次元に
- 多様な情報ソースを取り込みたい
 - オミックスデータ
 - 画像データ
 - 自然言語データ
 - 検査値データ
 - 波形データ

まとめ

- この講義では以下を学んだ
 - 機械学習を使った疫学研究の例
 - 具体的実装方法を提示
 - 機械学習を疫学に用いるモチベーションは何か？

2024年9月27日（金）
13:00ー13:50

NDB・DPC等データベース研究人材育成Webinar

SQLの活用:基礎編

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

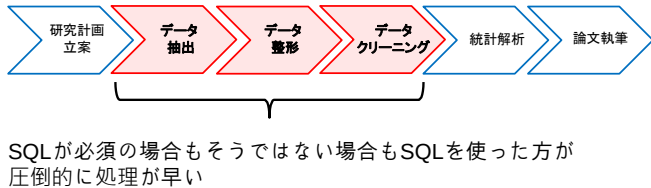
1

SQL概論

2

無断転載・利用を固く禁じます

SQLを使う研究の場面

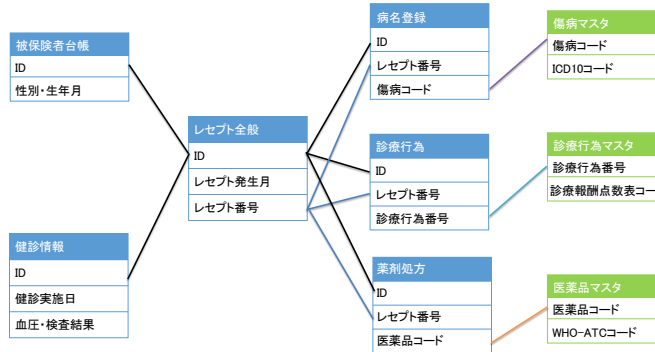


← どちらにしてもSQLを身につけるメリットは大きい

4

無断転載・利用を固く禁じます

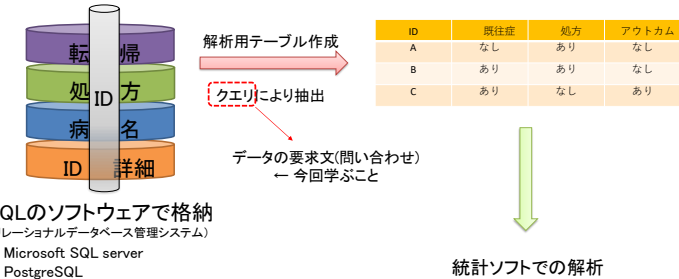
リレーショナルデータベースの構造



5

無断転載・利用を固く禁じます

リレーショナルデータベースの活用法



統計ソフトでの解析

6

無断転載・利用を固く禁じます

ここまでのまとめ:SQL概論

- ① SQL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
- ③ SQLを使うタイミング
研究計画立案を完了した後に、統計解析を行う直前までのテーブル作成

7

SQL入門①

データ外観、SELECT文、条件設定まで

注意:

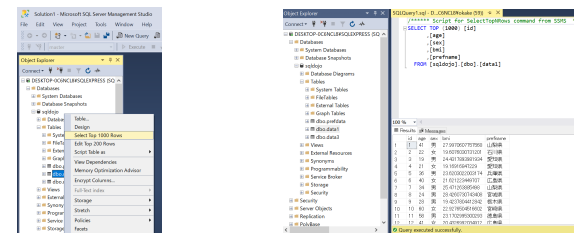
- ・これから挙げるクエリ例は、すべてMicrosoft SQL Serverの記述方法であり、ほかのソフトウェアでは方言が異なることにご注意ください。
- ・また、同じ結果を得るためのクエリの書き方は、必ずしも1通りではなく、わかりやすさを優先して例を載せているので、各自でもっとも適したと思われるクエリを探してみてください。

165

8

無断転載・利用を固く禁じます

それぞれのデータテーブルを眺める



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

9

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

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(下記はMicrosoft SQL Server の方言でもあるのでご注意ください):

- 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

10

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]
```

ID	年齢	性別
A	30	0
B	50	1
C	87	0
D	4	1

ID	性別
A	0
B	1
C	0
D	1

11

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
```

ID	年齢	性別
A	30	0
B	50	1
C	87	0
D	4	1

ID	年齢	性別
B	50	1
C	87	0

ポイント

- * 条件はWHEREでFROMよりも後ろに記述

12

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
```

ID	年齢	性別
A	30	1
B	50	1
C	87	0
D	4	1

ID	年齢	性別
C	87	0
B	50	1
A	30	1
D	4	1

13

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]
```

ID	年齢	性別
A	30	1
B	50	1
C	87	0
C	87	0
D	4	1
D	4	1

ID	年齢	性別
A	30	1
B	50	1
C	87	0
D	4	1

```
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?
```

ID
A
B
C
D

性別
1
0

14

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 '薬剤疫学者' OR [職業] LIKE '臨床疫学者' /*☆☆/
```

ID	年齢	職業
A	30	薬剤疫学者
B	50	臨床疫学者
C	87	社会学者
D	4	経済学者

ID	年齢	職業
A	30	薬剤疫学者
B	50	臨床疫学者

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 ('薬剤疫学者', '臨床疫学者')
- ☆ LIKEの後に部分一致を使いたいときは%を用いて'%疫学者%'とすればよい(後方一致なら2回目の%は不要)

15

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 /*☆☆/
```

ID	年齢	職業
A	30	薬剤疫学者
B	50	臨床疫学者
C	87	社会学者
D	4	経済学者

ID	年齢	職業
A	30	薬剤疫学者
B	50	臨床疫学者

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

16

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]
```

ID	BMI	性別
A	30	男
B	23	男
C	22	男
D	18	女
E	25	女
F	23	女

BMI最大値	BMI最小値	BMI平均値
30	18	23.5

166

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
```

ID	BMI	性別
A	30	男
B	23	男
C	22	男
D	18	女
E	25	女
F	23	女

性別	BMI最大値	BMI最小値	BMI平均値
男	30	22	25
女	25	18	22

18

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]

ID	年齢	性別	alive
A	30	1	1
B	50	1	1
C	87	0	1
D	4	1	1

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]

ID	年齢	性別	adult
A	30	1	1
B	50	1	1
C	87	0	1
D	4	1	0

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]

ID	年齢	性別	generation
A	30	1	middle
B	50	1	middle
C	87	0	older
D	4	1	young

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*/

ID	年齢	性別
A	30	1
B	50	1
C	87	0
D	NULL	1

SELECT *, INTO #1 FROM [データシート名]

WHERE 年齢 is not null /*年齢!=NULLは不可*/

ID	年齢	性別
A	30	1
B	50	1
C	87	0
D	NULL	1

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]

ID	年齢	性別	c年齢
A	30	1	30
B	50	1	50
C	87	0	87
D	NULL	1	0

(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]

ID	年齢	性別	c年齢
A	30	1	30
B	50	1	50
C	87	0	87
D	NULL	1	0

変数作成⑥:CAST/CONVERT関数を用いた日付

☆CONVERT関数を用いて日付と認識させるやり方

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

select convert(date, 誕生年月日, 112)

もしくは、

select cast(誕生年月日 as date)

のようにしても日付として変更出来る

日付認識されたものに、6ヶ月後を指定するなら

DATEADD(month,6,CAST(exam_ymd as date)) とすればできて、6年前でも
DATEADD(year,-6,CAST(exam_ymd as date)) とすればできる。

convertのなかの「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)

変数作成⑦: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	誕生日月日	入院年月日	ID	誕生日月日	入院年月日	年齢
A	19750105	20191001	A	19750105	20191001	
B	19611221	20200502	B	19611221	20200502	
C	19871029	20220819	C	19871029	20220819	

☆SQL自体が日付と変数を認識している場合
SELECT *, DATEDIFF(YEAR, [誕生日月日],[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]が日付認識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[誕生日月日]のところに
SQL Server の方では、CONVERT(date, [誕生日月日], 112)とすると日付に一旦変換される

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
CASTや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

ID	年齢
A	87
B	50
D	4

Table X

ID	性別
A	0
B	1
C	0

Table Y

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
```

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

テーブルの結合②:概略

内部結合INNER JOIN

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

両方のテーブルに存在するIDのみ結合

外部結合OUTER JOIN

左外部結合 LEFT JOIN

基準が左

ID	年齢	性別
A	87	0
B	50	1
D	4	NULL

左にあるIDのみ情報追加

右外部結合 RIGHT JOIN

基準が右

ID	性別	年齢
A	0	87
B	1	50
C	0	NULL

右にあるIDのみ情報追加

全外部結合 FULL JOIN

基準は両方

ID	年齢	性別
A	87	0
B	50	1
C	NULL	0
D	4	NULL

どちらかにあるIDに情報追加

テーブルの結合③:概念

内部結合INNER JOIN

基準は両方
(両方にあるもの)

ID	年齢	性別
A	87	0
B	50	1

両方のテーブルに存在するIDのみ結合

左外部結合 LEFT JOIN

基準が左

ID	年齢	性別
A	87	0
B	50	1
D	4	NULL

左にあるIDのみ情報追加

右外部結合 RIGHT JOIN

基準が右

ID	性別	年齢
A	0	87
B	1	50
C	0	NULL

右にあるIDのみ情報追加

全外部結合 FULL JOIN

基準は両方
(何れにあるもの)

ID	年齢	性別
A	87	0
B	50	1
C	NULL	0
D	4	NULL

どちらかにあるIDに情報追加

テーブル結合④:具体例

ID	年齢
A	87
B	50
D	4

Table X

ID	性別
A	0
B	1
C	0

Table Y

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④*/  
ON X.ID=Y.ID /*XとYのIDをキー変数と宣言⑤*/
```

注)
① XとYの両方にある変数を両方選ぶとエラーが出る
(1) どちらから選んてくるのか不明
(2) 同じ変数名は同じテーブルには存在できない
(3) 名付けるのはX/YではなくてもA/Bなどとしてもよい
(4) INTOの位置に注意
(5) 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

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

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

ID	性別	職業	処方月	薬剤
A	男	ワカナゴ	201401	グラルゲン
A	男	ワカナゴ	201401	リスプロ
A	男	イナダ	201402	カンデサルタン
A	...	...	...	...
A	男	ブリ	202006	エンバグリフロジン
B	女	ヒカリゴ	201401	アスピリン
...	...	...	...	...
C	女	スズキ	202206	クロビドグレル

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

結果の行数は何行でしょうか?
Aさんだけでは

A	男	ワカナゴ	201401	グラルゲン
...	...	...	...	...

に対して、Table Yの2014~2020までの全病名登録月・全病名が結合
されれば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	ワカナゴ	201401	グラルゲン
A	男	ワカナゴ	201401	リスプロ
A	男	イナダ	201402	カンデサルタン
A	...	...	...	...
A	男	ブリ	202006	エンバグリフロジン
B	女	ヒカリゴ	201401	アスピリン
...	...	...	...	...
C	女	スズキ	202206	クロビドグレル

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
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
```

37

## 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\*/

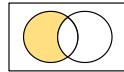
LEFT JOIN

[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

38

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

39

ご清聴ありがとうございました

40

## SQLの活用：中級編

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

1

## 臨床疫学研究における SQLにおけるデータクリーニング

DeSCデータベースを、Microsoft SQL Severのクエリを使い、解析可能なものにする実践的なデータクリーニング

2

## 具体例として使うデータベース・ソフト

### ◎ DeSCデータベース

- ・DeSCヘルスケアが提供、国保・健保・後期高齢者のレセプトと健診情報を含む。
- 他の候補となるデータベース
- JMDCデータベース、MDVデータベースなど

### ◎ Microsoft SQL Server

- ・Microsoftが提供、日本語対応がしっかりしている。
- 他の候補となるソフト
- PostgreSQL, Google BigQuery, Oracle SQL, MySQLなど

3

## 今回のダミー研究案

クリニカルクエスチョン：

「糖尿病を持つ人での貧血の頻度は？」

リサーチクエスチョン：

「健診データで貧血(男性13 g/dL未満、女性12 g/dL未満)を有するかを、透析患者を除外した上で、年齢・性別を調整して、糖尿病を持つ人と持たない人(主解析：病名レセプト発生、感度解析：処方薬使用)で頻度が違うかを確認する」

4

Figure 1のイメージ

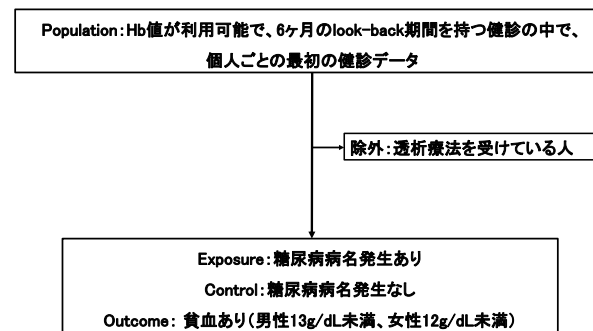
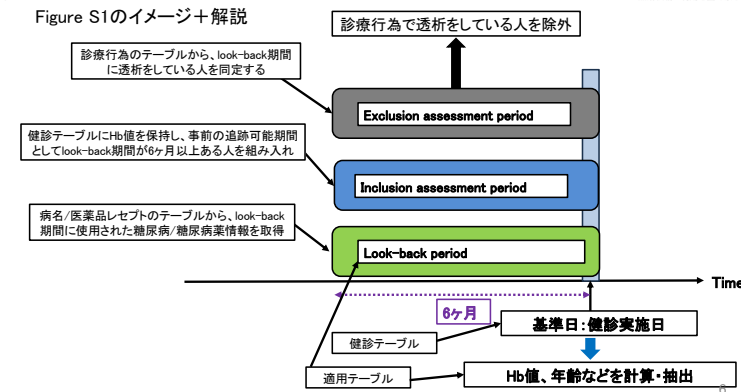
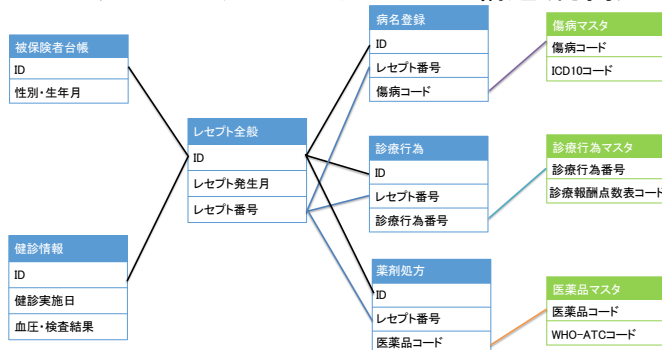


Figure S1のイメージ+解説



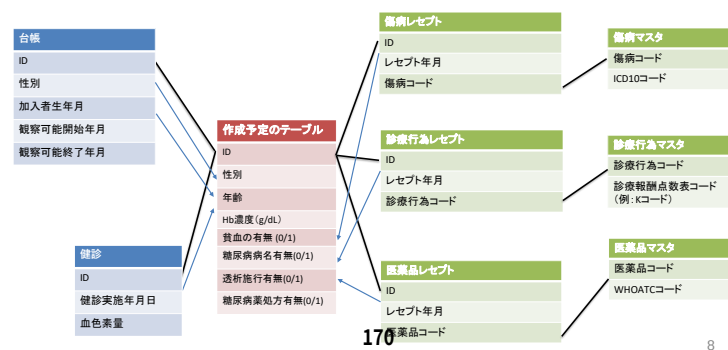
5

## リレーショナルデータベースの構造(再掲)



7

## データベースの構造と作成テーブル目標



8

## テーブル作成の手順

1. 研究のデザイン：横断研究・Look-back期間などの設定（ここまでのプロセス）
2. 対象の抽出：健診でのHb利用可能な人のデータで、初回に絞る
3. 曝露の抽出：Index dateから過去半年以内に糖尿病病名があったか
4. 除外基準の抽出：Index dateから過去半年以内に透析を実施したか
5. 曝露の感度解析の抽出：Index dateから過去半年以内に糖尿病薬処方あったか

9

# 研究対象の抽出 (PECOのpopulation)

Hb値の健診データが利用可能で、  
6ヶ月のlook-back期間がある人

## 対象の抽出①健診データから抽出

健診データでHb値が存在するものを選ぶ

以下のものを一時テーブル(#1)として保存

「健診テーブルでHb値が欠測でないデータ」

WHERE 血色素量 is not null

台帳

| ID | 健診実施年月日   | 血色素量 |
|----|-----------|------|
| A  | XXXX年X月X日 | 15.0 |
| A  | YYYY年Y月Y日 | 15.3 |
| A  | ZZZZ年Z月Z日 | NULL |
| B  | XXXX年X月X日 | 14.3 |
| B  | YYYY年Y月Y日 | NULL |
| Z  | XXXX年X月X日 | 11.2 |
| Y  | XXXX年X月X日 | 12.6 |
| X  | XXXX年X月X日 | NULL |

テーブル#1

| ID | 健診実施年月日   | 血色素量 |
|----|-----------|------|
| A  | XXXX年X月X日 | 15.0 |
| A  | YYYY年Y月Y日 | 15.3 |
| B  | XXXX年X月X日 | 14.3 |
| Z  | XXXX年X月X日 | 11.2 |
| Y  | XXXX年X月X日 | 12.6 |

ポイント

- \* 条件はWHEREで記載
- \* NULLは「IS NULL」や「IS NOT NULL」で条件を指定

10

## 対象の抽出②性別/誕生年月/観察開始年月を取得

以下のクエリで、性別/誕生年月/観察開始年月を追加し#2として保存する

/\*もとのテーブルに下記追加\*/  
性別, 誕生年月, 観察開始年月  
もとのテーブル X  
LEFT JOIN  
台帳 Y  
ON X.ID= Y.ID

テーブル#1

| ID | 健診実施年月日   | 血色素量 |
|----|-----------|------|
| A  | XXXX年X月X日 | 15.0 |
| A  | YYYY年Y月Y日 | 15.3 |
| B  | XXXX年X月X日 | 14.3 |
| Z  | XXXX年X月X日 | 11.2 |
| Y  | XXXX年X月X日 | 12.6 |

テーブル#2

| ID | 健診実施年月日   | 血色素量 | 性別 | 誕生年月    | 観察開始年月  |
|----|-----------|------|----|---------|---------|
| A  | XXXX年X月X日 | 15.0 | 男性 | YYYY年Y月 | ZZZZ年Z月 |
| A  | YYYY年Y月Y日 | 15.3 | 男性 | YYYY年Y月 | ZZZZ年Z月 |
| B  | XXXX年X月X日 | 14.3 | 女性 | YYYY年Y月 | ZZZZ年Z月 |
| Z  | XXXX年X月X日 | 11.2 | 女性 | YYYY年Y月 | ZZZZ年Z月 |
| Y  | XXXX年X月X日 | 12.6 | 男性 | YYYY年Y月 | ZZZZ年Z月 |

ポイント

- \* #1と台帳テーブルを連結する
- \* LEFT JOINの形式は基礎編を復習

11

12

## 対象の抽出③look-backと年齢を計算

Lookback変数と年齢を#2に追加し、#3として保存する

DATEDIFF(MONTH,  
CAST(観察開始可能年月 + '/'01' as date),  
CAST(健診実施年月日 as date))  
as Lookback月数,  
DATEDIFF(YEAR,  
CAST(誕生年月 + '/'01' as date),  
CAST(健診実施年月日 as date))  
as 年齢

テーブル#3

| ID | 健診実施年月日   | 血色素量 | 性別 | 誕生年月    | 観察開始年月  | Lookback月数 | 年齢 |
|----|-----------|------|----|---------|---------|------------|----|
| A  | XXXX年X月X日 | 15.0 | 男性 | YYYY年Y月 | ZZZZ年Z月 | 45         | 45 |
| A  | YYYY年Y月Y日 | 15.3 | 男性 | YYYY年Y月 | ZZZZ年Z月 | 57         | 46 |
| B  | XXXX年X月X日 | 14.3 | 女性 | YYYY年Y月 | ZZZZ年Z月 | 40         | 70 |
| Z  | XXXX年X月X日 | 11.2 | 女性 | YYYY年Y月 | ZZZZ年Z月 | 4          | 16 |
| Y  | XXXX年X月X日 | 12.6 | 男性 | YYYY年Y月 | ZZZZ年Z月 | 2          | 30 |

ポイント

- \* 日付計算の方法(CAST、差のDATEDIFF)は基礎編を復習

13

## 対象の抽出④対象を絞り、順序をつける

#4で対象を絞り、#5で順位付け変数を作成

先ほどのテーブルを下記で絞る(#4)

WHERE Lookback月数 >=6 AND 年齢 >=18

#4の一部に「健診何回目」の変数を加え  
一時テーブルとして保存(#5)

ROW\_NUMBER() over(partition by ID  
order by 健診実施年月日)  
as 健診何回目

ポイント

- \* ROW\_NUMBERで番号を振る

テーブル#3から□を削除して#4を作成

| ID | 健診実施年月日   | 血色素量 | 性別 | 誕生年月    | 観察開始年月  | Lookback月数 | 年齢 |
|----|-----------|------|----|---------|---------|------------|----|
| A  | XXXX年X月X日 | 15.0 | 男性 | YYYY年Y月 | ZZZZ年Z月 | 45         | 45 |
| A  | YYYY年Y月Y日 | 15.3 | 男性 | YYYY年Y月 | ZZZZ年Z月 | 57         | 46 |
| B  | XXXX年X月X日 | 14.3 | 女性 | YYYY年Y月 | ZZZZ年Z月 | 40         | 70 |
| Z  | XXXX年X月X日 | 11.2 | 女性 | YYYY年Y月 | ZZZZ年Z月 | 4          | 16 |
| Y  | XXXX年X月X日 | 12.6 | 男性 | YYYY年Y月 | ZZZZ年Z月 | 2          | 30 |

テーブル#5

| ID | 健診実施年月日   | 血色素量 | 性別 | 誕生年月    | 観察開始年月  | Lookback月数 | 年齢 | 健診何回目 |
|----|-----------|------|----|---------|---------|------------|----|-------|
| A  | XXXX年X月X日 | 15.0 | 男性 | YYYY年Y月 | ZZZZ年Z月 | 45         | 45 | 1     |
| A  | YYYY年Y月Y日 | 15.3 | 男性 | YYYY年Y月 | ZZZZ年Z月 | 57         | 46 | 2     |
| B  | XXXX年X月X日 | 14.3 | 女性 | YYYY年Y月 | ZZZZ年Z月 | 40         | 70 | 1     |

14

## 対象の抽出⑤最初の健診を同定する

最初の健診に限定する(#6)

WHERE 健診何回目 =1

データのID重複を確認する  
SELECT COUNT(ID) FROM #5  
SELECT COUNT(DISTINCT ID) FROM #5  
SELECT COUNT(ID) FROM #6  
SELECT COUNT(DISTINCT ID) FROM #6

ポイント

- \* IDの重複を確認することにより、一人1行になったことを確認する

15

## 対象の抽出⑥アウトカムを定義する

PECOのoutcomeに相当する変数を作成する

CASE WHEN 性別=女性 AND 血色素量 <12 then 1  
WHEN 性別=男性 AND 血色素量 <13 THEN 1 ELSE 0 END AS 貧血

テーブル#7

| ID | 健診実施年月日   | 血色素量 | 性別 | 誕生年月    | 観察開始年月  | Lookback月数 | 年齢 | 健診何回目 | 貧血 |
|----|-----------|------|----|---------|---------|------------|----|-------|----|
| A  | XXXX年X月X日 | 15.0 | 男性 | AAAA年A月 | DDDD年D月 | 45         | 45 | 1     | 0  |
| B  | YYYY年Y月Y日 | 14.3 | 女性 | BBBB年B月 | EEEE年E月 | 40         | 70 | 1     | 0  |
| Z  | ZZZZ年Z月Z日 | 11.3 | 女性 | CCCC年C月 | FFFF年F月 | 12         | 37 | 1     | 1  |

ポイント

- \* CASE WHENの使用方を復習

16

# 曝露の抽出 (PECOのexposure)

傷病テーブルから該当患者の糖尿病病名が健診の過去  
6ヶ月以内に発生しているかの変数を作る

## 曝露の抽出①糖尿病傷病コードを抽出

傷病マスタから、糖尿病の傷病コードを抽出(#糖尿病傷病コード)

WHERE ICD10コード LIKE 'E10%' OR  
ICD10コード LIKE 'E11%' OR  
ICD10コード LIKE 'E12%' OR  
ICD10コード LIKE 'E13%' OR  
ICD10コード LIKE 'E14%'

テーブル#糖尿病傷病コード

| 病名コード   |
|---------|
| XXXXXX1 |
| XXXXXX2 |
| XXXXXX3 |

ポイント

- \* LIKEの時に使える、部分一致の%の用法を確認

17

18

## 曝露の抽出②糖尿病病名発生月を絞る

傷病レセプトのテーブルから、IDが#糖尿病傷病レセに含まれるもの、かつ傷病コードが#糖尿病傷病コードに含まれるものを取り出す(#糖尿病傷病レセ)

```
WHERE 傷病コード IN (SELECT DISTINCT 傷病コード FROM #糖尿病傷病レセ)
AND ID IN (SELECT DISTINCT ID FROM #7)
```

```
#糖尿病傷病レセに、健診年月日を結合(#糖尿病傷病レセ2)
FROM 傷病レセプト X
LEFT JOIN
#7 Y
ON X.ID = Y.ID
```

ポイント

\* サブクエリを使って中間テーブルを節約する

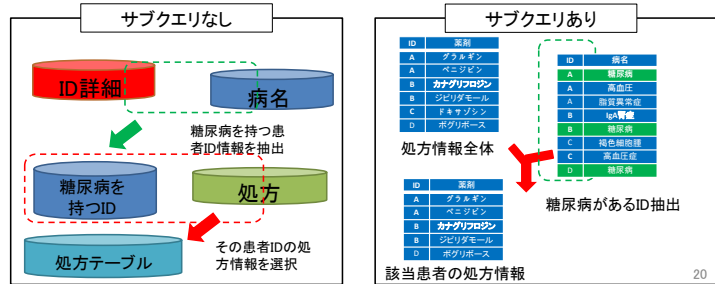
テーブル#糖尿病傷病レセ2

| ID | レセプト年月  | 健診実施年月日   |
|----|---------|-----------|
| A  | XXXX年X月 | AAAA年A月A日 |
| A  | YYYY年Y月 | AAAA年A月A日 |
| D  | XXXX年X月 | XXXX年X月X日 |
| D  | YYYY年Y月 | XXXX年X月X日 |
| D  | ZZZZ年Z月 | XXXX年X月X日 |

19

## 発展編①:サブクエリとは

☆SQL文が多くなるデメリット → 中間生成物が増えてHDDの負担になる  
サブクエリ: 中間生成物を作るクエリをメインのクエリに入れる

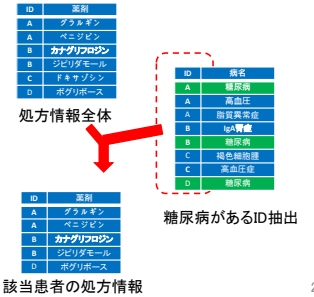


20

## 発展編②:WHERE句に置くサブクエリ

☆WHERE内のサブクエリ  
→ IDの使用条件をINで例示する！

```
クエリ例
SELECT *
FROM [処方テーブル]
WHERE ID in
(SELECT distinct ID
from [病名テーブル]
where 病名 like '糖尿病')
```



21

## 曝露の抽出③糖尿病病名ありを変数作成

★ 糖尿病病名ありの変数を作成

```
WHERE 診療年月日 BETWEEN 健診実施年月日の6ヶ月前 AND 健診実施年月日
```

上記条件で選択されるもので変数作成(#糖尿病傷病レセ3)  
1 AS 糖尿病病名あり

テーブル#糖尿病傷病レセ3

| ID | 糖尿病病名あり |
|----|---------|
| A  | 1       |
| D  | 1       |
| F  | 1       |
| X  | 1       |

22

## 曝露の抽出④糖尿病病名有無を変数追加

#7にさきほどの糖尿病病名ありをLEFT JOINで結合して#8を作る

| ID | 健診実施年月日   | 血圧値  | 性別 | 誕生年月    | 健診開始年月  | Lookback月数 | 年齢 | 健診項目 | 貧血 | 糖尿病病名あり |
|----|-----------|------|----|---------|---------|------------|----|------|----|---------|
| A  | XXXX年X月X日 | 15.0 | 男性 | AAAA年A月 | DDDD年D月 | 45         | 45 | 1    | 0  | 1       |
| B  | YYYY年Y月Y日 | 14.3 | 女性 | BBBB年B月 | EEEE年E月 | 40         | 70 | 1    | 0  | NULL    |
| C  | ZZZZ年Z月Z日 | 11.3 | 女性 | CCCC年C月 | FFFF年F月 | 12         | 37 | 1    | 1  | NULL    |

上記のように#8には、NULLが発生するので、NULLを0に置換しておく(#9)

ISNULL(糖尿病病名有り,0) AS 糖尿病病名有無

| ID | 健診実施年月日   | 血圧値  | 性別 | 誕生年月    | 健診開始年月  | Lookback月数 | 年齢 | 健診項目 | 貧血 | 糖尿病病名有無 |
|----|-----------|------|----|---------|---------|------------|----|------|----|---------|
| A  | XXXX年X月X日 | 15.0 | 男性 | AAAA年A月 | DDDD年D月 | 45         | 45 | 1    | 0  | 1       |
| B  | YYYY年Y月Y日 | 14.3 | 女性 | BBBB年B月 | EEEE年E月 | 40         | 70 | 1    | 0  | 0       |
| C  | ZZZZ年Z月Z日 | 11.3 | 女性 | CCCC年C月 | FFFF年F月 | 12         | 37 | 1    | 1  | 0       |

ポイント

\* ISNULL関数でNULLを消しておく統計ソフトでの処理が楽

23

## 除外基準の抽出

傷病テーブルから該当患者の診療行為レセで健診の過去6ヶ月以内に透析を施行しているかの変数を作る

## 除外基準の抽出①透析コードを抽出

```
SELECT 診療行為コード
INTO #透析コード
FROM 診療行為マスタ
WHERE 診療報酬点数表コードがJ38 and 診療報酬点数表コードがJ42
```

あとは曝露の部分の②～④参照して変数を#9に追加して#10とする

ポイント

\* 診療行為コードなどのコードは、データベースによって方式が違う  
\* 年度により、コードが違うことがあるので、マスタを確認すること

25

## 曝露(感度解析用)の情報抽出

医薬品テーブルから該当患者のレセで健診の過去6ヶ月以内に糖尿病薬を処方しているかの変数を作る

## 曝露薬剤の抽出①糖尿病医薬品コードを抽出

```
SELECT DISTINCT 医薬品コード
INTO #糖尿病薬コード
FROM 医薬品マスタ
WHERE WHOATCコード like 'A10%'
```

あとは曝露の部分の②～④参照して変数を#10に追加して#11とする

ポイント

\* 世界共通のWHO-ATCコードで薬剤を指定する  
\* ATCコードは、WHOでない他のATCコードもあるので確認が必要

26

27

結果（完全なるダミーの結果、岡田入力）

| ID | 研究実施年月日   | 年齢   | 性別 | 調査年月    | 調査開始年月  | Loadback月数 | 年齢 | 研究計画日 | 資金 | 結果発表の有無 | 透明化の有無 | 結果発表の方向性 |
|----|-----------|------|----|---------|---------|------------|----|-------|----|---------|--------|----------|
| A  | XXXX年X月X日 | 15.0 | 男  | YYYY年Y月 | ZZZZ年Z月 | 45         | 45 | 1     | 0  | 1       | 0      | 1        |
| B  | XXXX年X月X日 | 14.3 | 男  | YYYY年Y月 | ZZZZ年Z月 | 9          | 28 | 1     | 0  | 0       | 0      | 0        |
| C  | XXXX年X月X日 | 11.2 | 女  | YYYY年Y月 | ZZZZ年Z月 | 40         | 70 | 1     | 1  | 0       | 0      | 0        |
| D  | XXXX年X月X日 | 12.6 | 男  | YYYY年Y月 | ZZZZ年Z月 | 80         | 45 | 1     | 1  | 1       | 0      | 0        |
| E  | XXXX年X月X日 | 14.3 | 男  | YYYY年Y月 | ZZZZ年Z月 | 19         | 45 | 1     | 0  | 0       | 0      | 0        |
| F  | XXXX年X月X日 | 11.2 | 女  | YYYY年Y月 | ZZZZ年Z月 | 16         | 45 | 1     | 1  | 0       | 0      | 0        |
| G  | XXXX年X月X日 | 15.0 | 男  | YYYY年Y月 | ZZZZ年Z月 | 40         | 45 | 1     | 0  | 0       | 0      | 0        |
| H  | XXXX年X月X日 | 9.8  | 女  | YYYY年Y月 | ZZZZ年Z月 | 6          | 24 | 1     | 1  | 0       | 0      | 0        |
| I  | XXXX年X月X日 | 11.2 | 男  | YYYY年Y月 | ZZZZ年Z月 | 59         | 78 | 1     | 1  | 0       | 0      | 0        |
| H  | XXXX年X月X日 | 11.0 | 女  | YYYY年Y月 | ZZZZ年Z月 | 26         | 80 | 1     | 1  | 1       | 1      | 0        |
| K  | XXXX年X月X日 | 14.3 | 男  | YYYY年Y月 | ZZZZ年Z月 | 30         | 45 | 1     | 0  | 0       | 0      | 0        |

ポイント

- \* 最終的なテーブルをイメージしてクエリ作成を行う。
- \* （今回は省略したが）仮テーブル作成直後に都度中身を確認する

本日のまとめ:SQLの活用:中級編

- ① 研究デザインがしっかりしていないと変数が作れないので、CQ→RQのプロセスをしっかりと行う。  
（例:交絡は既存研究を踏まえて網羅されているか？）
- ② 最終的なテーブルをイメージしてクエリ作成を行う。
- ③ 各テーブル同士を繋ぐ、キー変数(ID、日付)を把握する。
- ④ CASE WHENやGROUP BYを使い変数を作成して、最終的にLEFT JOINで繋げていく。

ご清聴ありがとうございました



2024年9月27日（金）  
15:00 - 15:50

NDB・DPC等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 上級編

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

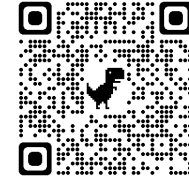
大規模なレセプトデータベースを解析する場合に当たって、

- レセプトデータベースの構造を理解する。
- 解析用データセットの設計・構築の流れを学習する。

## 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/iryohoken/reseputo/index.html](https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryohoken/reseputo/index.html)

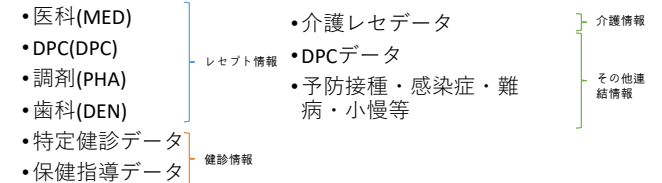
## ここ最近の研究成果

- 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.
- 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;50161-6420(21)00719-3.
- 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):1379-48.
- 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;80(4):714-27.
- 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.
- Kimura Y, Suzukawa M, Jo T, Hashimoto Y, Kumazawa R, Ishimaru M, et al. Epidemiology of severe childhood asthma in Japan: A nationwide descriptive study. Allergy. 2024 Jan 7;all.16008.
- Kimura Y, Jo T, Hashimoto Y, Kumazawa R, Ishimaru M, Matsui H, et al. Epidemiology of patients with lymphangioleiomyomatosis: A descriptive study using the national database of health insurance claims and specific health checkups of Japan. Respir Investig. 2024 May;62(3):494-502.

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、介護情報等との接続も可能になった。

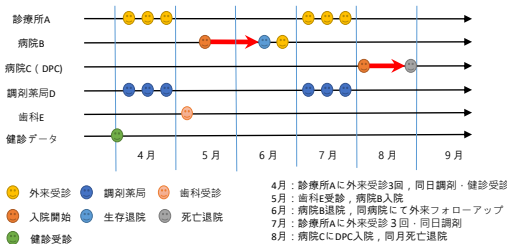


今回の演習では主として医療レセプトのデータハンドリングを学ぶ

NDBデータ構造

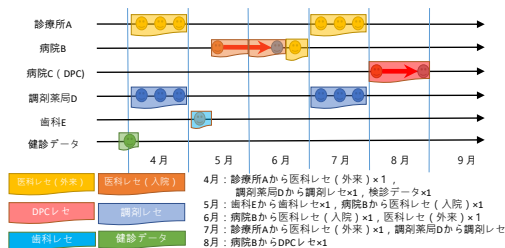
## NDBデータ の大まかな構造

患者のエピソードで考えると以下の様になる。



NDBデータ構造

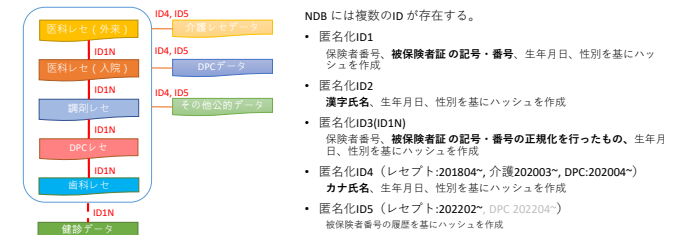
## NDBデータ の大まかな構造



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で横断的の紐付けを行う事が出来る

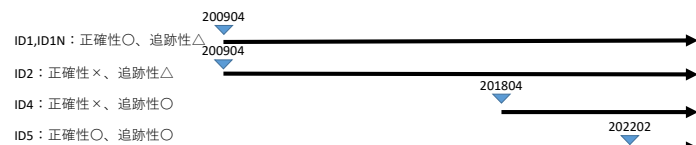


- NDBには複数のIDが存在する。
- 匿名化ID1  
保険者番号、被保険者証の記号・番号、生年月日、性別を基にハッシュを作成
- 匿名化ID2  
漢字氏名、生年月日、性別を基にハッシュを作成
- 匿名化ID3(ID1N)  
保険者番号、被保険者証の記号・番号の正規化を行ったもの、生年月日、性別を基にハッシュを作成
- 匿名化ID4（レセプト:201804~, 介護202003~, DPC:202004~）  
カナ氏名、生年月日、性別を基にハッシュを作成
- 匿名化ID5（レセプト:202202~, DPC 202204~）  
被保険者番号の履歴を基にハッシュを作成





## レセプトにおけるIDの特性と出現時期



近年登場したID5によって、正確な個人追跡が可能となった。  
過去のデータはID0を用いる縦断追跡が必要  
ID5を用いた"ID0の答え合わせ"はできるはず。

## 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。）

## 縦断ID作成の事例

学術論文でOnlineになったIDの作成法  
National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB): Outline and Patient-Matching Technique  
<https://www.biorxiv.org/content/early/2018/03/29/280008.article-metrics>

## NDBデータ構造

### NDBデータの大まかな構造

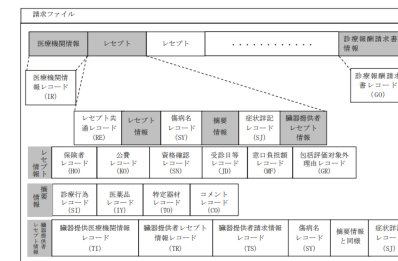
- レセプトの記載内容がそのまま登録されている。

## レセプトデータの構造化

共通の様子が公開されている。

- 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](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/igynaiyo/igynaiyo\\_03/index.html](https://www.ssk.or.jp/igynaiyo/igynaiyo_03/index.html)

電子レセプトのCSV情報及びレセプト電子データのサンプルデータ

|            |                   |            |
|------------|-------------------|------------|
| → 医科: 18件  | 0_COMMON001_MED   | ファイル名: 001 |
| → PC: 8件   | 1_MEDICAL001_0001 | ファイル名: 001 |
| → 歯科: 21件  |                   |            |
| → 産科: 18件  |                   |            |
| → 訪問看護: 5件 |                   |            |

## レセプトデータの構造化

- 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\\_MEDICAL001\\_0001\\_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 —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。

- 抽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
- ☒ 単抽出依頼テンプレート（標準） [XLS形式: 1,210KB]
- ☒ 単抽出依頼テンプレート（標準） [XLS形式: 1,206KB]
- ☒ 単抽出依頼テンプレート（標準） [XLS形式: 1,530KB]

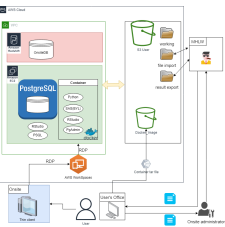
DB仕様は機械判読可能な形で

- エクセル内に提供ファイルのカラム型情報あり

| 項目       | データ型 | 単位   | 備考   | 単位   | 備考   |
|----------|------|------|------|------|------|
| 1. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 2. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 3. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 4. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 5. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 6. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 7. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 8. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 9. 診療科目  | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 10. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 11. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 12. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 13. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 14. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 15. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 16. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 17. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 18. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 19. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |
| 20. 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 | 診療科目 |

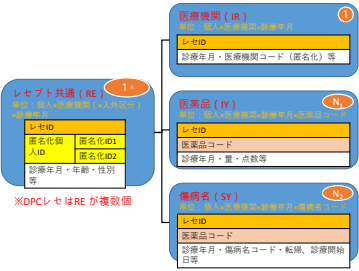
特別抽出仕様NDB データベース

- 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。



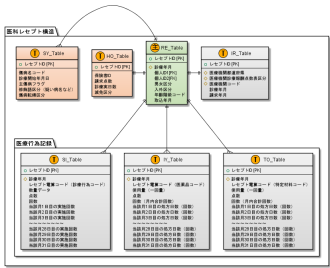
NDBデータ構造

- 複数のテーブル（右図）で1レセの情報がなる。
- 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）とはレセプト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



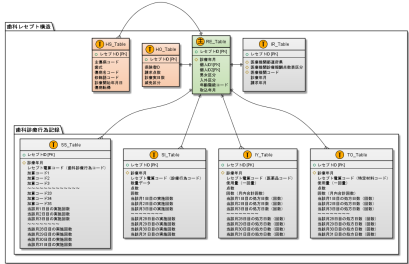
医療レセプトの情報

- 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



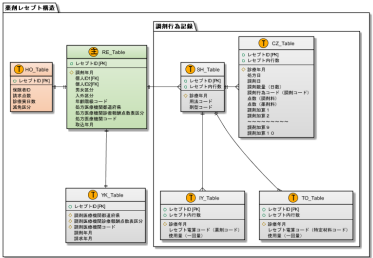
歯科レセプトの情報

- 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



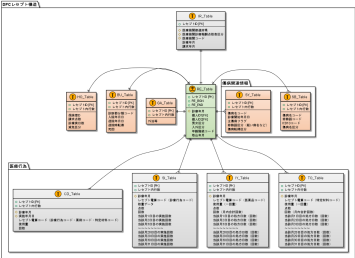
調剤レセプトの情報

- 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



NDBデータ構造に触れる

- 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。
- 演習1にアクセスしてください。



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

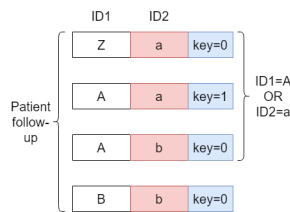
## オンサイトセンターでの研究の実際

- 症例抽出計画の策定
  - Inclusion Period の決定
  - 症例抽出条件の（キーレセプトの抽出）決定
  - 抽出する情報の決定
- 症例の追跡
  - 症例追跡期間の決定
- レセプト情報の整理（コホート化）
- 解析



## オンサイトセンターでの症例の実際

- ## 症例の追跡
- 症例追跡期間の決定
    - ・ 許諾された全期間でキーレセプト前後情報を追跡する。
  - 症例の追跡
    - ・ キーレセプトを基にID1,ID2を再帰的に追跡する。
    - ・ 'ID1 = OR ID2 = 'の条件一回で追跡を打ち切るのはダメ
  - 追跡したID1,ID2を基にID0を作成する。
    - ・ 別人を誤って追跡した場合は後でその症例を除くが諦める。
    - ・ ランダム抽出する場合は、ID0単位で抽出が望ましい。
  - 演習（NDBの構造）の演習10に取り組んでください。



## 解析用データセットを設計する

NDBデータベースに格納されている情報

DBから得られるタイムスタンプ (時間軸の粒度)



DBから得られる集計単位



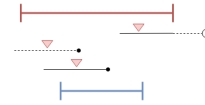
DBから得られる情報



## オンサイトセンターでの研究の実際

## 症例抽出計画の策定

- **Inclusion Period の決定**
  - 許諾された全期間で症例を検索する。
  - 容量削減や抽出プロセス短縮のために、期間を絞ることはある。
- **レセプト条件は単一レセプト内で完結する条件にする。**
  - 他のレセプトを参照して決定する条件は好ましくない。
  - 例えば、“脳梗塞病名のついた症例のレセプト”や“tPA実施症例のレセプト”は検索できる。
  - “脳梗塞発症後”、“リハビリテーション”を行ったレセプトの検索は危険
- **研究に利用する情報を決める。**



## オンサイトセンターでの症例の実際

## レセプト情報の整理（コホート化）

- ・対象となるIDを基にレセプトを検索
- ・レセプトから**必要となる情報**を抽出する
  - ・実施医療行為や傷病名に加え特定健診情報等
  - ・情報抽出過程はレセプト単位で独立させる。
  - ・MEDISの標準マスタなどを活用
- ・時系列に情報を並べる（解析データベースの構築・コホート化）。



解析用データセットを設計する。

- レセプトデータはテーブル構造が複雑で解析しにくい
- 解析しやすいデータとは？

- **収集用（格納用）のデータ構造と解析用のデータ構造は異なる。**
  - R, Stata, SPSS, SAS, Python いずれも、基本的に1つの表（テーブル）に情報を集約する。
    - 行：解析粒度
    - 列：情報
- | 患者ID | 年齢 | 性別 | 治療薬1 | ..... | 転帰 | 最終観察 |
|------|----|----|------|-------|----|------|
| 001  | 75 | 男  | 1    | 0     | 0  | 0    |
| 002  | 78 | 女  | 0    | 1     | 0  | 0    |
| 003  | 72 | 男  | 1    | 0     | 0  | 0    |
| 004  | 76 | 女  | 0    | 1     | 0  | 0    |
| 005  | 74 | 男  | 1    | 0     | 0  | 0    |

解析粒度（テーブル1行）

177 データ次元



## オンサイトセンターでの研究の実際

サンプルクエリ

```

-- 以下为 TBL_SQZ2 和 TBL_SQZ3 中新增数据与删除数据的 SQL 执行操作记录。
-- 以下为新增
SELECT 'dblink exec'
INSERT INTO TBL_SQZ2 Extract
SELECT DISTINCT SQZ2_NO,PRAC_NO FROM insight_db.scr.mds.user_info,T_TPC_HCD ED WHERE PRAC_NO = '2021019' AND random() <=0.05
UNION
SELECT DISTINCT SQZ2_NO,PRAC_NO FROM insight_db.scr.mds.user_info,T_TPC_HCD ED WHERE PRAC_NO = '2021019' AND random() <=0.05
-- 以下为删除
-- test
SELECT SQZ2_NO FROM dblink('SQZSHS755') SELECT SQZ2_NO FROM TBL_SQZ2 Extract LIMIT 10 SQZSHS755 as (SQZ2_NO varchar(200))

```

- 演習（NDBの構造）の演習4に取り組んでください。



## 解析用データセットを設計する

- NDBに含まれるデータとその構造を学習した。
- NDBデータは構造として使いにくい
  - 情報がテーブル間にまたがって存在
  - テーブルの中は月単位情報と日単位情報がバラバラに
  - レセプト単位・処置単位でとれる情報が一つのテーブルに
- 中級編で触れたような、解析しやすい形のデータセットを作成したい。
- 設計の考え方を学ぶ



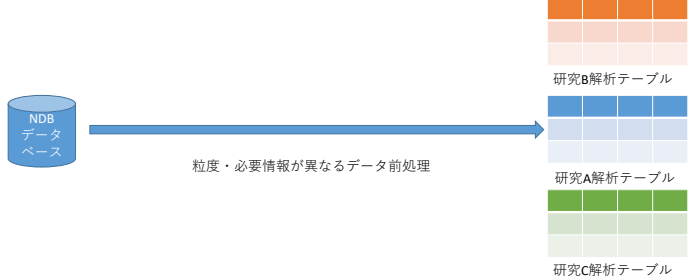
## データハンドリング

- 必要な解析粒度は解析手法によって異なる。
  - 患者単位
  - 入院単位
  - 地域単位
  - 病院単位
  - 期間単位
  - 患者日単位
- 必要なデータ次元も研究によって異なる。
  - 特定治療薬・特定診療行為の有無
  - 特定診断名

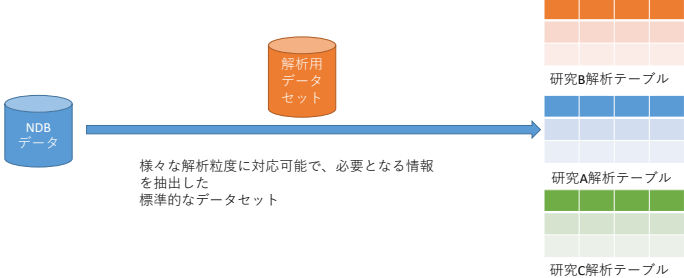
| 解析粒度（テーブル1行） |    |    |       |       |    |      |  |
|--------------|----|----|-------|-------|----|------|--|
| 患者ID         | 年齢 | 性別 | 治療薬 1 | ..... | 転帰 | 最終観察 |  |
|              |    |    |       |       |    |      |  |



NDBデータと解析用データ



NDBデータと解析用データ

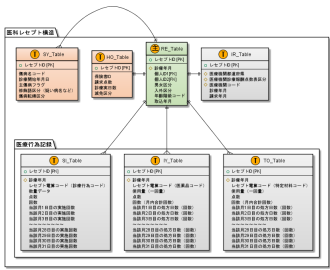


解析用データセットの設計

- NDBデータに含まれる情報（エンティティ）をリストアップする。
- 各エンティティに紐づくの属性を整理する。
- 各エンティティの時間粒度を整理する。
- 各エンティティ間の接続・依存関係（1:N, 1:1, N:M）を整理する。
- 時間粒度と紐づく属性が似ている情報を1つのテーブルにまとめる。
  - この際、データベース正規化をどこまで行うかは個別事例による。

医科レセプトの情報（エンティティ）

- 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



医科レセプトの情報（エンティティ）

- 患者
- 最終転帰
- レセプト発生
- 患者の外来受診
- 患者の入院
- 診断
- 診療行為記録（診療行為・医薬品・特定材料）

医科レセプトエンティティに紐づく属性

- 患者
  - 患者ID・年齢・性別・観察開始年月日・観察終了年月日
- 最終転帰
  - 患者ID・死亡転帰
- レセプト発生
  - 患者ID・レセプトID・診療年月・月額保険費用・保険者区分・レセプト種別
- 患者の外来受診
  - レセプトID・外来受診日・医療機関
- 患者の入院
  - レセプトID・入院日・退院日・医療機関
- 診断
  - レセプトID・診療年月・傷病名区分コード・修飾語コード・診療開始年月日
- 診療行為記録（診療行為・医薬品・特定材料）
  - レセプトID・レセプト電算コード・実施日・使用量・回数・点数

医科レセプトエンティティ時間粒度

- 患者
  - 患者ID・年齢・性別・観察開始年月日・観察終了年月日
- 最終転帰
  - 患者ID・死亡転帰・死亡観察日
- レセプト発生
  - 患者ID・レセプトID・診療年月・月額保険費用・保険者区分・レセプト種別
- 患者の外来受診
  - レセプトID・外来受診日・医療機関
- 患者の入院
  - レセプトID・入院日・退院日・医療機関
- 診断
  - レセプトID・診療年月・傷病名区分コード・修飾語コード・診療開始年月日
- 診療行為記録（診療行為・医薬品・特定材料）
  - レセプトID・レセプト電算コード・実施日・使用量・回数・点数

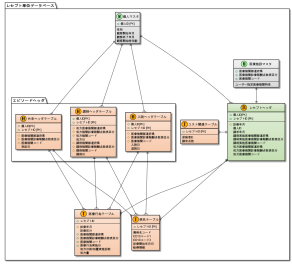
用意するテーブル

過去に実績のある研究用データセットの仕様

- 患者IDテーブル
  - 患者個人に対し1行
- レセプトヘッダテーブル
  - レセプト（診療年月）ごとに1行（DPCレセプトについては複数行）
- 医療行為テーブル
  - 診療行為・医薬品・材料の日ごとの実施状況
- 診断テーブル
  - 月×診療開始年月日毎（DPCレセプトについては入院日毎）の診断情報
- アウトカムテーブル
  - 死亡転帰

解析用データベース構造の一例

- 当教室で解析用DBとして用いているData Model



## 解析用データベースから研究用テーブルを作る

レセプトを跨いだ症例情報を取得する（研究用テーブル）。

- Date Zero の設定
- Washout Period の設定
- Follow-up と打ち切りの設定
- Exposure の設定
- Outcome の設定
- Covariate の設定
- 中級編で学んだこと

## 解析データベースを構築する

- 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。
- 参考のページにアクセスしてください。



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

## まとめ

大規模なレセプトデータベースを解析する場合に当たって、

- レセプトデータベースの構造を学んだ。
- 解析用データセットの設計・構築の流れを学習した。

- 演習用ページを利用して、理解をさらに深めてください。