

データマネジメント

東京大学大学院医学系研究科臨床疫学・経済学
松居宏樹

情報管理の失敗事例

- K大学において、学生が、研究利用目的の患者情報をUSBメモリに保存し、それを紛失
- A病院において、患者個人情報を含むノートPCとUSBメモリを医師が院外に持ち出し、車上荒らしによりそれらを盗難される。
- 研究過程情報が記録された研究ノートを紛失

→研究者にペナルティと研究の中止という**失敗**

この講義の目的

- 臨床データの分析におけるデータマネジメントの実例を学ぶ

(i)臨床疫学研究に関連する情報管理の実例を学ぶ

(ii)大規模データベースから、目的にあった標準データセットを抽出するプロセスを復習する。

(iii)大規模データベースから、目的にあった標準データセットを抽出するプロセスを新たに設計する。

情報管理の失敗事例

- セキュリティ確保のための規制が多く、利用者が増えない国保有データベース

- 個人情報保護が課題となり、研究利用が進まない学会主導データベース

→研究は未実施となる**失敗**。

この講義の目的

- 臨床データの分析におけるデータマネジメントの実例を学ぶ

(i)臨床疫学研究に関連する情報管理の実例を学ぶ

(ii)大規模データベースから、目的にあった標準データセットを抽出するプロセスを復習する。

(iii)大規模データベースから、目的にあった標準データセットを抽出するプロセスを新たに設計する。

情報の管理とは

- 情報管理を失敗すると
研究のアウトプットが棄損される。

- 情報管理で重要なのはCIA
 - Confidentiality : 守秘性（機密性）
 - Integrity : 完全性
 - Availability : 可用性

臨床疫学研究における情報管理は、研究のアウトプットの長期的**最大化**を目指して**可用性を損なわないように守秘性と完全性を高める作業**

データ（情報）の管理

- 臨床データの研究利用における種々の規制
 - 臨床研究法
 - 個人情報保護法
 - 人を対象とする医学系研究に関する倫理指針

→研究者は情報の適切な管理を求められる

データベースを管理するとは

- 臨床疫学研究における情報管理は、**研究のアウトプットの長期的最大化を目指して**
- **可用性を損なわないように**守秘性と完全性を高める作業

研究のアウトプット

- 研究のアウトプットの単純モデル
1年間に処理可能な研究数

$$S = \frac{365 \times T}{D}$$

T: 同時進行研究数 (Thread)
D: 1研究の平均期間 (Duration)

ただし

$$S \leq S_{max}$$

Smax: 供給される研究Seedの数

Thread(同時研究数) を増やす

同時に研究できるActive な研究者を増やし、それに対応する。

- 多診療科ネットワークが必要
- 同時アクセス可能な解析環境が必要

コツ

- 標準化データ構造と教育でボトルネックが生じないようにする必要

研究のアウトプット (現在)

- 年間100本以上の臨床疫学研究論文
- 同時研究本数は25-30程度
- Durationは3か月から1年程度

$$365 * (27/100) = 98.6$$

Duration(研究期間) を短縮する

研究着想・データ取得・解析用データ作成・解析・論文執筆から短縮可能なプロセスを探す

- 倫理手続きの整備 (データ取得)
- データ構造の標準化と教育 (解析用データ作成)

コツ

- 誰が・いつ・何をしたかがしっかりわかる構造で解析用データを作成

研究生産性を上げる方法

- Smax を増やす
 - 臨床現場のClinical Question を吸い上げる。
- Thread を増やす
 - 同時に研究できるActive な研究者を増やし、それに対応する。
- Duration を短縮する
 - 研究着想・データ取得・解析用データ作成・解析・論文執筆から短縮可能なプロセスを探す

研究のアウトプットを最大化するために

- システム運用者ができること
 - 情報の守秘性・完全性を高める
 - セキュリティ・バックアップ
 - 可用性を高める
 - 同時アクセス可能な解析環境が必要
 - データ構造の標準化と教育 (解析用データ作成)

Smax を増やす

臨床現場のClinical Question を吸い上げる。

- CQをResearch Question に変換する手伝いが必要
 - 多診療科ネットワークが必要
- コツ
- ある程度診療科から独立して (医師は診療科の垣根を超えるのにハードルがある)
 - どういった情報があり、何ができるのかを示していく

情報の守秘性

情報へのアクセスを認められた者だけが、その情報にアクセスできる状態を確保すること

- 物理的・技術的にデータにアクセスできる研究者を制限する。
- 研究倫理指針上では個人情報の安全管理に関連する規定に該当する。

情報の完全性

情報が破壊、改ざん又は消去されていない状態を確保すること

- データのバックアップやバージョン管理
- データのアクセス・編集履歴（ログ）を適切に管理
- 研究倫理指針上では研究に係る試料及び情報等の保管に関連する規定に該当する。
- (iii)で紹介する研究の再現性にも関連する。

研究情報管理の事例

情報を管理するための工夫

- 個人の研究なら、技術的な工夫はあまり必要ない。
 - データの利用端末を制限
 - 研究終了後に当該端末を破壊

情報の可用性

情報へのアクセスを認められた者が、必要時に中断することなく、情報及び関連資産にアクセスできる状態を確保すること

- 研究に参画した研究者がシームレスに研究データにアクセスできる。
- 冗長なシステム構成
- 利便性の向上
- (ii)で話す研究目的の標準化データもこの一環

研究情報管理の事例

情報を管理するための工夫

- チームで研究を行う場合、権限の分離が必須
 - データの管理者
 - データの利用者
 - データのアクセス権がない者
- ユーザー認証を行うシステムを利用
 - データのアクセス権がない者の排除

• どのように実施するか？

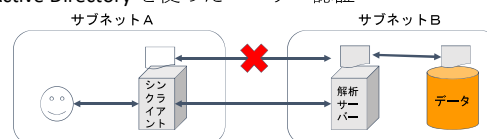
ここまでのまとめ

- 臨床疫学研究における情報管理は、研究のアウトプットの長期的最大化を目指して
- 可用性を損なわないように守秘性と完全性を高める作業

システム構築の事例

解析サーバーへのRDP接続

- シンクライアントの利用
- ネットワーク分離
- リモートデスクトップ（RDP）を用いたデータ利用
- Active Directory を使ったユーザー認証



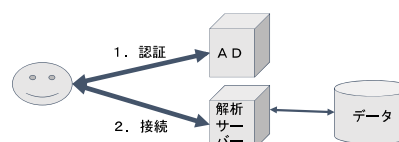
演習

- シナリオ
 - あなたは医療機関の職員です。院内データを活用して、研究論文を発表するプロジェクトが立ち上がり、データ管理を任せられました。
 - データ：患者レベルの個票データ（匿名化前）
 - 利用者：医師10名、大学院生5名
 - 外部持ち出しは禁止
 - 予算：年間100万円以内
- どういった場所にデータを置きますか？
- どういった機材・システムを用いますか？
- アクセス制御はどう行いますか？
- ログ管理はどうしますか？

ユーザー認証システム

Windows Server のActive Directory を利用

- 研究参加時にユーザー登録
- ユーザーの権限をユーザーグループで分離
- システムへのログインをActive Directoryで実施



研究情報管理の事例

機密性を高めるための工夫

- データの保存された媒体には、データ管理者のみがアクセスできる。
- データの利用者はデータの解析を行える。
- データの利用者はコピー・持ち出し出来ない。

アクセスログ取得の事例

- システムへのアクセスログ
 - Active Directory は認証の記録を保存できる。
 - ネットワーク内でのサーバー通信ログも保存できる。
- 情報の閲覧編集ログ
 - 商用のNASはおおよそ編集ログなどが保存できる
- ログの監査計画の策定
 - ログ自体を改ざんが困難な領域へ移動
 - 少なくとも何かあったときに参照できるように

研究情報管理の事例

その他の機密性を高めるための工夫

- デバイスの認証とドメインポリシー
 - 管理外のデバイスがシステムに接続できないようにする。
 - デバイスの動作は中央管理
- データの削除方法
 - データの保存されたことのある媒体は適切なデータ削除を行ってから破棄する。

研究情報管理の事例

可用性を高めるための工夫

- ユーザーの利便性向上
 - 複数サイトからのアクセス
 - 在宅ワークへの対応
-
- 運用レベルで、研究がしやすいデータセットの整備→(ii)で説明

研究情報管理の事例

完全性を高めるための方法

- データのバックアップ
- バックアップの世代管理
- 情報のアクセスログを保存

可用性を高めるための工夫

- 大規模データの場合
 - ユーザーがデータファイルにアクセスできるだけでは利便性が向上しない
 - 例えば、ユーザーがアクセス可能なDBを用意する。
 - 例えば、ユーザーが利用しやすい標準的なデータ様式を用意する。

→ユーザーの利便性の向上

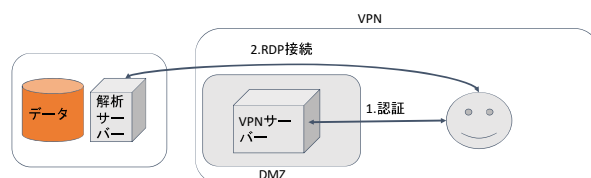
バックアップの事例

当教室では複数のバックアップを併用している

- 複数のデータ保管サーバーを同期
 - 一つのサーバーが破損してもデータは維持できる。
 - データが保存されたボリュームイメージを週次でバックアップ
 - 複数世代保持することで、研究者の意図的なデータ廃棄などに対応
- 特に重要なデータは暗号化したうえでクラウド上に保存
 - 研究室の火災などに対応
- テープバックアップシステムも導入

可用性を高めるための工夫

- 複数施設が研究に参加する場合や在宅ワークの利用者がいる場合
- Virtual Private Network(VPN)による接続



臨床研究データの管理

最低ラインはどんなもの？

- システムへのアクセス管理
 - 複数名のユーザーならDomain Controller
- 複数サブネットを設定できるネットワークスイッチ（L3スイッチ）
- データへのアクセスを管理できる機材
 - 高機能なNASなら20TBで20万くらい
- シンクライアントなら安いもので5万-10万
- 継続的にデータの管理体制を維持・改善できる人材を研究チームに必ず一人入れる

なぜデータベースからデータを抽出するのか？

- プロセスの標準化によるデータ可用性の向上
- 抽出プロセスとデータをセットで保存することでデータの完全性を保証

Take Home Message

臨床疫学研究における情報管理は、研究のアウトプットの長期的最大化を目指して

可用性を損なわないように守秘性と完全性を高める作業

今日話した内容を理解して実行できる人材を用意しましょう。

DPCデータ抽出のプロセス

- 対象となるエピソード(入院・外来)を抽出するためのエピソードキー(抽出条件)となる病名(ICD-10)、診療行為(レセ電コード)に基づき抽出する。
- 抽出したエピソードの前後全てのエピソードを個人ID（病院ID + 患者ID)に基づき抽出する。
- 抽出したエピソードに、研究に必要な診療行為情報（処置・手術・検査・薬剤等）のデータがあれば、抽出したエピソードの各行に情報を付与して、解析用テーブルを作成する

この講義の目的

- 臨床データの分析におけるデータマネジメントの実例を学ぶ

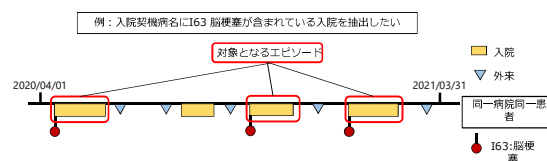
(i)臨床疫学研究に関連する情報管理の実例を学ぶ

(ii)大規模データベースから、目的にあった標準データセットを抽出するプロセスを復習する。

(iii)大規模データベースから、目的にあった標準データセットを抽出するプロセスを新たに設計する。

DPCデータ抽出のプロセス

- 対象となるエピソード(入院・外来)を抽出するためのエピソードキー(抽出条件)となる病名(ICD-10)、処置・手術・検査・薬剤(レセ電コード)で抽出する

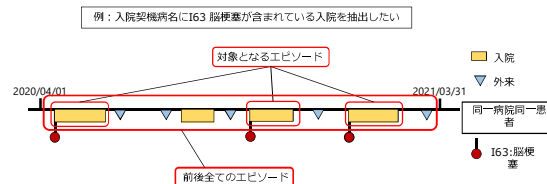


目的

- DPCデータベースからのデータ抽出のプロセスを理解する
- データ抽出依頼者とデータ抽出班とのコミュニケーション・エラーを理解する
- 抽出依頼書を作成するために必要な情報を取得できるようになる

DPCデータ抽出のプロセス

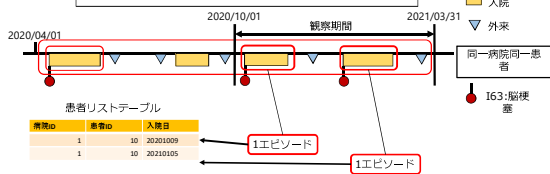
- 対象となるエピソードの病院IDと患者IDで、そのエピソードの前後全てのエピソードを抽出する



DPCデータ抽出のプロセス

3. そのうち、観察期間に該当するエピソードを入院日・退院日・外来受診日を用いて抽出し、各エピソード1行の患者リストテーブルを作成する

例：入院契機病名にI63 脳梗塞が含まれている入院を抽出したい
観察期間は2020年10月～2021年3月とする



データ抽出依頼者とデータ抽出班の コミュニケーション・エラーの防止策

①抽出依頼書のフォーマットを統一する

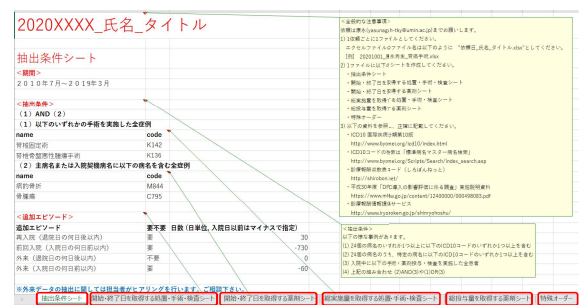
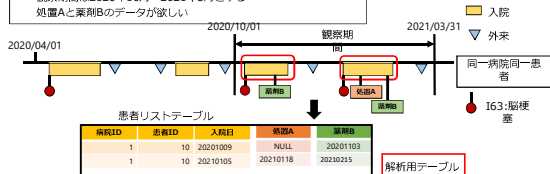
②抽出依頼書の各シートの書き方や注意事項を明確に記載する (処置セット、薬剤セットやDPCデータベース関連資料を活用する)

③抽出依頼書とともにチェックリストも提出してもらう

DPCデータ抽出のプロセス

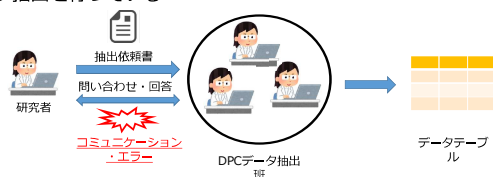
4. 対象となるエピソード内に、研究に必要な処置・手術・検査・薬剤のデータがあれば、患者リストテーブルの各行に1列ずつ追加して、解析用テーブルを作成する

例：入院契機病名にI63 脳梗塞が含まれている入院を抽出したい
観察期間は2020年10月～2021年3月とする
処置Aと薬剤Bのデータが欲しい



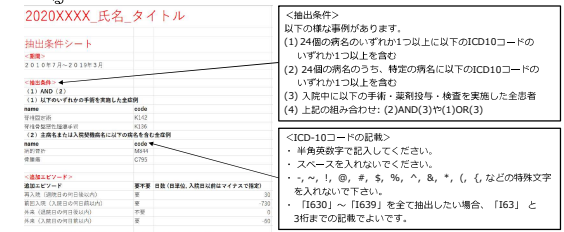
当研究室でのデータ抽出体制

- 研究者が各自でDPCデータの抽出をしていない
- 研究者が作成した抽出依頼書をもとにDPCデータ抽出班がデータ抽出を行っている



データ抽出依頼者とデータ抽出班のコミュニケーション・エラーの防止策

②抽出依頼書の各シートの書き方や注意事項を明確に記載する



データ抽出依頼者とデータ抽出班の コミュニケーション・エラーの防止策

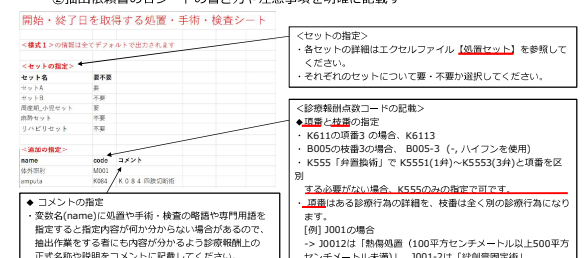
①抽出依頼書のフォーマットを統一する

- 抽出条件シート
- 開始・終了日を取得する処置・手術・検査シート
- 開始・終了日を取得する薬剤シート
- 総実施量を取得する処置・手術・検査シート
- 総投与量を取得する薬剤シート
- 特殊オーダーシート

以上6シートに必要な項目を記載してもらう

データ抽出依頼者とデータ抽出班のコミュニケーション・エラーの防止策

②抽出依頼書の各シートの書き方や注意事項を明確に記載する



演習：抽出依頼書を書いてみよう！！

【演習用_DPC研究計画書.doc】を読んで、この研究の抽出依頼書の空欄に必要項目を入力し、抽出依頼書を完成させてください。

空欄は【演習用_抽出依頼書_問題.xlsx】の  で囲まれたセルです。

しろばんねっとは最新年度診療報酬点数表を参照してください。

解答時間の目安は10分です。

それでは始めてください！！

大規模な医療データの具体例

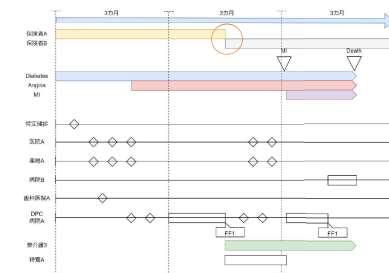
- DPCデータベース
 - 全国の病院から収集した入院退院サマリ情報
 - 年間6TBくらい
- JMDCデータベース（レセプトデータベース）
 - 一部保険者（社保）から取得した医療レセプト情報
 - 17年分くらい
 - 特定健診情報
 - 1.5TBくらい
- DeSCデータベース（レセプトデータベース）
 - 一部保険者（社保・国保・後期高齢）から取得した医療レセプト情報
 - 8年分くらい
 - 特定健診情報
 - 2.5TBくらい
- NDBデータベース（レセプトデータベース）
 - 全保険者から取得した全国の医療レセプト請求情報
 - 10年分くらい
 - 特定健診情報
 - 10TBくらい

演習：抽出依頼書を書いてみよう！！

解答例は【演習用_抽出依頼書_解答例.xlsx】を参照してください。

【演習用_論文.pdf】は実際にこの研究テーマでpublishされた論文です。

Patients time flow とデータベース



架空個人の9カ月の医療・介護関連記録
糖尿病が発覚し、
その後狭心症のオペを行った。
退院後は要介護度3となり、
特養で短期入所していたものの
心筋梗塞を発症し入院、
転院したものの
死に至った。

観察期間の途中で、加入保険者が変わっている。

この講義の目的

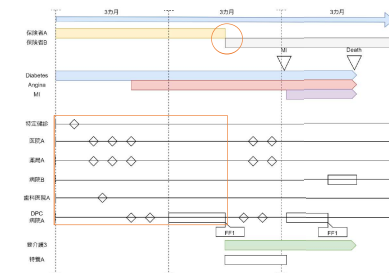
- 臨床データの分析におけるデータマネジメントの実例を学ぶ

(i)臨床疫学研究に関連する情報管理の実例を学ぶ

(ii)大規模データベースから、目的にあった標準データセットを抽出するプロセスを復習する。

(iii)大規模データベースから、目的にあった標準データセットを抽出するプロセスを新たに設計する。

Patients time flow とデータベース

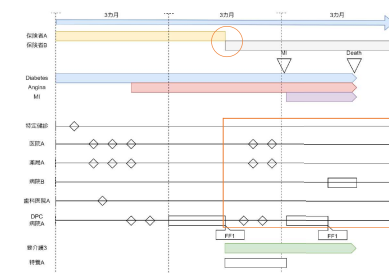


保険者Aのデータベースでは、以下が含まれる。
- 特定健診受診時の検査結果（HbA1C等）
- 医院A、薬局A、歯科医院Aのレセプト
- DPC病院A(外来と1回目入院)のレセプト

時系列医療データ

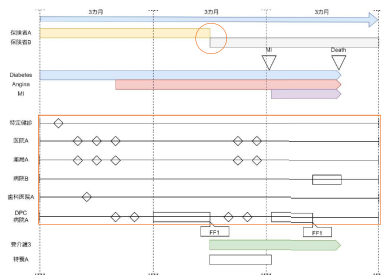
- 疫学の基本は、「だれが、いつ、どうしたか」
 - 「どこで」、「どれくらい」、も重要なことはある
- 「だれが、いつ、どうしたか」を明確にして、データを整理する。

Patients time flow とデータベース



保険者Bのデータベースでは、以下が含まれる。
- 医院A、薬局A、病院Bのレセプト
- DPC病院A(外来、2回目の入院)のレセプト
- 死亡転帰

Patients time flow とデータベース



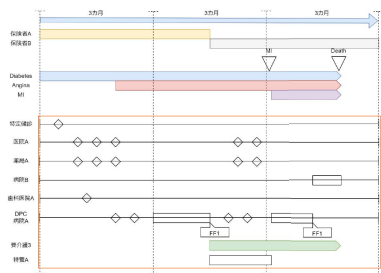
NDB データベースでは、以下が含まれる。

- 特定健診受診時の検査結果 (HbA1C等)
- 医院A, 薬局A, 歯科医院Aのレセプト
- DPC病院Aのレセプト

ハンドリングprocess の標準化の必要性

- データハンドリングをSQLで行う場合
- 記述内容が自由すぎる
 - レビュー困難でエラー発見が難しい
- プロセスが非効率
 - 長大なテーブルを無駄に全件スキャン
 - 長大なテーブルを無駄にコピー
- データハンドリングprocessの標準化は、システムの可用性・完全性につながる。

Patients time flow とデータベース



NDB データベースでは、以下が含まれる。

- 特定健診受診時の検査結果 (HbA1C等)
- 医院A, 薬局A, 歯科医院Aのレセプト
- DPC病院Aのレセプト

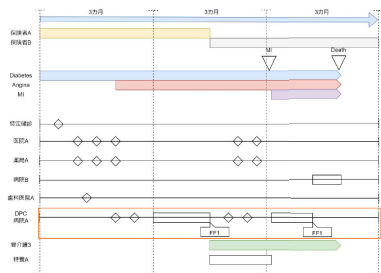
NDB データベースには以下の情報を紐づけられる

- 要介護度
- 特養Aの入所記録
- DPC 様式1 (退院時サマリ) 情報

クエリ設計

- 疫学研究に使用しやすい形に「だれが、いつ、どうしたか」を効率的に整理する。
 - だれが：解析単位
 - いつ：観察のタイムライン
 - どうしたか：マスタ整備

Patients time flow とデータベース



DPC データベースでは、以下が含まれる。

- DPC病院Aの外来レセプト情報
- DPC病院Aの入院情報 (FF1: 様式1情報)

解析単位を整理する

解析単位とは

解析用のテーブルの1行・一つの観察対象

- 個人
- 入院等、特定の期間で区切った観察期間

すべての解析単位は一つの解析タイムラインと一つのIndex date (Cohort entry date)を持つ。

なお、解析単位1つについて、複数のアンカーが存在しうる。

時系列医療データ

- 疫学の基本は、「だれが、いつ、どうしたか」
 - 「どこで」、「どれくらい」、も重要なことはある
- 「だれが、いつ、どうしたか」を明確にして、データを整理する。
- いずれのデータベースも、「だれが、いつ、どうしたか」が記録されている。
- 疫学研究に使用しやすい形に「だれが、いつ、どうしたか」を整理する。→データハンドリングprocess

演習：解析単位を整理する

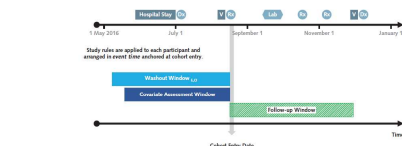
- New User Design
- 入院
 - 最初の入院だけを対象にする研究
 - 2回目以降の入院も対象にする研究
- 手術
 - 最初の手術だけを対象とする場合
 - 複数回手術をそれぞれ解析する場合
- 時間依存共変量モデル

演習：解析単位を整理する

- New User Design ptid
- 入院
 - 最初の入院だけを対象にする研究 **ptid**
 - 2回目以降の入院も対象にする研究 **ptid+dtadm+dtdisc**
- 手術
 - 最初の手術だけを対象とする場合 **ptid**
 - 複数回手術をそれぞれ解析する場合 **ptid+dtadm+dtdisc+オペ日**
- 時間依存共変量モデル **ptid+time_bgn+time_end**

演習：解析単位ごとに異なるアンカーの数は？

Figure 1. From transactional data to study implementation.



Individual patient data are documented as encounters from various sources and are arranged in calendar time. This work is licensed under CC BY, and the original versions can be found at www.repositorio.igp.org.br/projects.html. Dv = diagnosis; E = exposure; Lab = laboratory test; O = outcome; Rx = drug dispensing; V = visit.

Ann Intern Med. 2019;170:398-406. doi:10.7326/M18-2079

タイムラインを整理する

データベースによって、観察可能な時間の粒度は差がある。

- 月単位・日単位・分単位・秒単位・・・

タイムラインには、複数のアンカーがある。例えば、

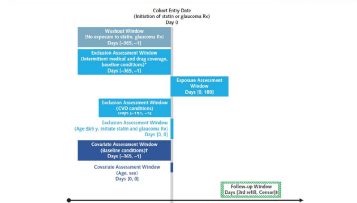
- **Index date**
- 退院後の再入院を検討する研究の場合は、**退院日**
- ICU退室後のイベントを検討する研究の場合は、**ICU退室日**

タイムラインには、アンカーを起点として複数のWindowがある。例えば、

- **Washout window**
- **Assessment window**
- **Follow-up window**

演習：解析単位ごとに異なるアンカーの数は？

Figure 4. Exposure-based cohort entry restricted to adverse cases.

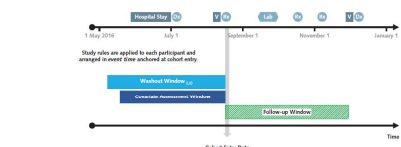


This work is licensed under CC BY, and the original versions can be found at www.repositorio.igp.org.br/projects.html. Dv = diagnosis; E = exposure; Lab = laboratory test; O = outcome; Rx = drug dispensing; V = visit.

Ann Intern Med. 2019;170:398-406. doi:10.7326/M18-2079

デザインダイアグラム

Figure 1. From transactional data to study implementation.



Individual patient data are documented as encounters from various sources and are arranged in calendar time. This work is licensed under CC BY, and the original versions can be found at www.repositorio.igp.org.br/projects.html. Dv = diagnosis; E = exposure; Lab = laboratory test; O = outcome; Rx = drug dispensing; V = visit.

Ann Intern Med. 2019;170:398-406. doi:10.7326/M18-2079

マスタを整備する

- データベース内には「どうしたか」がテーブルに保存されている。
 - 薬剤が処方された
 - 傷病の診断がついた
 - 死亡した
- 「どうしたか」を解析に使える形に定義しなおす。
- **テーブルの数=マスタ数**
- **コード：カテゴリの対応表**

デザインダイアグラム

- 論文投稿時に適当に書くものではない。
- タイムライン整理のツール
- ポイントは、解析単位ごとに異なるアンカー（可変アンカー）がどこにあるかを明確にする。

マスタを整備する

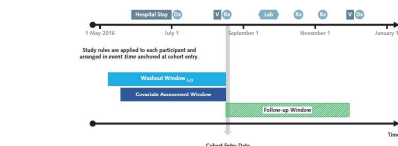
- 用いるデータベースに応じて用意するマスタは異なる。
- **JMDCデータベース**
 - 医薬品コードマスタ*医薬品、後_医薬品、歯_医薬品
 - 標準化材料コードマスタ*材料、後_材料、歯_材料
 - 標準化診療行為コードマスタ*医科診療行為、後_医科診療行為、歯_医科診療行為
 - 歯科診療行為コードマスタ*歯_歯科診療行為
 - 調剤行為コードマスタ*薬学管理KI、調剤情報CZ、調剤診療行為、後_調剤診療行為
 - 標準傷病コードマスタ*傷病、後_傷病

マスタを整備する

- 用いるデータベースに応じて用意するマスタは異なる。
- DeSCデータベース
 - 医薬品コードマスタ→receipt_drug
 - 標準化材料コードマスタ→receipt_specific_equipment
 - 標準化診療行為コードマスタ→receipt_ika_medical_practice_detail
 - 歯科診療行為コードマスタ→receipt_dental_practice
 - 歯科加算コードマスタ→receipt_dental_practice_addition
 - 調剤行為コードマスタ→receipt_dispensing
- 標準傷病コードマスタ→receipt_diseases
- 歯科歯式コードマスタ→receipt_tooth_type_diseases

演習： 解析単位ごとに異なるアンカーの数は？

Figure 1. From transactional data to study implementation.



テーブルの数のマスタを用意する

決してSQLのWHERE IN句内にハードコーディングしないこと。

- 保守性が低い
- 再利用しにくい
- 検索速度も遅い

マスタテーブルを作成してJoinする。

解析単位ごとに異なるアンカーを解析単位テーブルに紐づける

hospcd	ptid	Index_date	Fst_obs	End_obs
A	a	20201001	20200401	20221231
A	a	20201130	20200401	20221231
A	b	20220103	20200331	20221231
B	a	20190315	20180315	20220315

標準的な時系列データのクエリとは

- 解析単位を整理する。
 - Index Date 単位に行を持つ解析単位テーブルを用意
- タイムラインを整理する。
 - デザインダイアグラムを書く
 - 解析単位ごとに異なるアンカーを解析単位テーブルに紐づける
- マスタと、解析単位テーブルを時系列データに紐づけ
 - データベースに記録された時系列情報にマスタと解析単位テーブルをJOINする。
- Window単位で情報を絞り、解析単位×マスタカテゴリ単位で集計する。

マスタと解析単位テーブルを時系列情報に紐づける

hospcd	ptid	Fst_obs	End_obs	Index_date	実施日	reccd	Value	カテゴリー
A	a	20200401	20221231	20201001	20200831	aaaaaa		A
A	a	20200401	20221231	20201001	20201001	aaaaaa		A
A	a	20200401	20221231	20201001	20201201	aaaaaa		A
A	a	20200401	20221231	20201001	20201231	bbbbbb		B

解析単位テーブル

hospcd	ptid	Index_date
A	a	20201001
A	a	20201130
A	b	20220103
B	a	20190315

演習：以下のタイムラインをWindowに分ける

hospcd	ptid	Fst_obs	End_obs	Index_date	実施日	reccd	Value	カテゴリー	Window
A	a	20200401	20221231	20201001	20200831	aaaaaa		A	
A	a	20200401	20221231	20201001	20201001	aaaaaa		A	
A	a	20200401	20221231	20201001	20201201	aaaaaa		A	
A	a	20200401	20221231	20201001	20201231	bbbbbb		B	

演習：以下のタイムラインをWindowに分ける

hospcid	ptid	Fst_obs	End_obs	Index_date	実施日	reccd	Value	カテゴリー	Window
A	a	20200401	20221231	20201001	20200831	aaaaaa		A	Pre index window
A	a	20200401	20221231	20201001	20201001	aaaaaa		A	Date of Inclusion
A	a	20200401	20221231	20201001	20201201	aaaaaa		A	Outcome Window
A	a	20200401	20221231	20201001	20201231	bbbbbb		B	Outcome Window

解析単位の除外はいつするの？

- SQL の抽出過程で解析単位を除外しない。
- Patient Selection Flow** を書こうと思うと、**wide** 形式にしたうえで、件数をカウントし、解析コード内で除外を行うべき。

なぜこの手順が大事なのか？

- デザインダイアグラムとクエリ内容が一致する。
 - レビューがしやすい。
 - 他人の書いた意図のはっきりしない長大なクエリを読むのはつらい
 - 修正も比較的容易
- オリジナルのデータを読みに行く機会が減る
 - サーバーに無駄な負荷がかからない。
- 自動化への道筋ができる。
- 型を知ること、横展開がしやすい。
 - 時系列を持つ医療データ全般に適用できる。

切り出したデータが重くない？

- それを解決するために商用のソフト（Microsoft SQL SERVER）を使っている。
- CLUSTERED COLUMNSTORE INDEX** を使うと、集計速度が上がるうえに、テーブルサイズも10%くらいに抑えられる。

```
CREATE CLUSTERED COLUMNSTORE INDEX CIX_インデックス名 ON 切り出したテーブル;  
GO
```

集計内容は？

- そのWindow内にカテゴリがあったかなかったか？
- そのWindow内の最初のカテゴリ出現日・最後のカテゴリ出現日
- そのWindow内の投与量の最高値・最低値
- そのWindow内の・・・

Take Home Message

- 時系列データのハンドリングは、データを疫学研究に使用しやすい形に「だれが、いつ、どうしたか」を整理する。
 - だれが：解析単位
 - いつ：観察のタイムライン
 - どうしたか：マスタ整備
- データハンドリングprocessの標準化は、システムの可用性・完全性につながる。
- 研究としての再現性・再生性の観点からからも重要である。

集計した後は？

Long->Wide 変換で解析単位ごとのテーブルにする。
最終的に、解析単位テーブルに情報をLeft Join していく。

```
SELECT  
  解析単位,  
  MAX(CASE WHENカテゴリ = 'hoge' THEN 集計値 ELSE NULL END) AS hoge_集計値  
  , MAX(CASE WHENカテゴリ = 'fuga' THEN 集計値 ELSE NULL END) AS fuga_集計値  
FROM Window_集計テーブル  
GROUP BY 解析単位
```