DPCサマーセミナー
@産業医科大学

DPCデータ分析演習

横浜市立大学
データサイエンス研究科
ヘルスデータサイエンス専攻
清水沙友里

データ分析の前に…、よくある質問

Q.　PCが固まってしまいます

データの前処理(必要なデータに絞る、分割)、数式の最適化(ループ処理をしない、)データのリンク(外部データベースやCSVファイルからデータを直接読み込むことで、Excelシートにすべてをロードする必要を減らす)、マクロとVBA: 繰り返し行う作業は、マクロやVBA(Visual Basic for Applications)を使用して自動化してもよい。動作が重くなりやすい関数は以下のようなものがあります。

参照関数:VLOOKUP、HLOOKUP、INDEX、MATCH、OFFSETなど。大量のセルにわたって使用されると、計算に時間がかかる。
条件付き集計関数:SUMIFS、COUNTIFS、AVERAGEIFSなどは、多くの条件や大量のデータが関与する場合には、計算に時間がかかる
再帰的な参照:関数が他のセルを参照し、そのセルがさらに別の関数によって計算される場面では、計算量が急速に増加する可能性がある
自作関数(VBA):VBAで作成されたユーザー定義関数も、処理が複雑であると時間がかかる
外部データ参照:外部データベースやWebサービスへのクエリを含む関数は、データベースのパフォーマンスに依存するため、遅くなる可能性がある
関数のネスト:複数の関数をネスト(入れ子に)して使用すると、計算の複雑性が増し、パフォーマンスに影響を与える

パフォーマンスの問題を軽減するためには、計算の範囲を限定する、不要な計算を削除する、Excelの計算オプションを手動に設定する(自動計算をオフにする)などの方法があります。

Q. データをグラフ化する際のポイント①

データをグラフ化する際には、何を伝えたいのか、また誰に伝えたいのかというコンテキストが非常に重要です。

明確な目的:
グラフには明確な目的が必要です。何を伝えたいのかを考え、それに最も適したグラフの種類を選びます。

シンプルさ:
複雑なグラフは混乱を招きます。必要な情報だけを表示し、無駄を排除します。

色の使用:
色は強力な視覚的要素ですが、多用すると逆効果になる場合もあります。基本的には2～3色に抑え、強調したい点に色を使います。

ラベルと凡例:
グラフの各要素が何を意味するのかを明確にするために、適切なラベルと凡例を使用します。

スケールと軸:
スケールが不適切だと、データの解釈が誤りやすくなります。0から始まる軸や、等間隔のスケールを使用すると良いでしょう。

Q. データをグラフ化する際のポイント②

データをグラフ化する際には、何を伝えたいのか、また誰に伝えたいのかというコンテキストが非常に重要

棒グラフ（垂直または水平）:
用途: カテゴリ別に数値を比較する。月別売上、部門別利益など。

折れ線グラフ:
用途: 時間に沿ったデータのトレンドを表示する。株価の時間経過による変動、年度ごとの売上推移など。

円グラフ:
用途: 全体に対する各部分の割合を示す。分類が多すぎると読みにくくなる。通常は5～6個以下のカテゴリで使用する。

ヒストグラム:
用途: 一変数のデータセットの頻度分布を表示する。データをビン（範囲）に分け、各ビンに含まれるデータ点の数を棒グラフで表示

散布図:
用途: 二つの数値の関係性を探る。x軸とy軸にそれぞれ異なる変数を取り、各データ点をプロットします。
例: 年齢と収入、気温と売上など

箱ひげ図:
用途: データの四分位数と外れ値を視覚化する。
特長: 中央値、第一四分位数、第三四分位数などを一つのグラフで表示し、データのばらつきを理解しやすくします。

Q. 知っていると便利なexcel関数は？

VLOOKUP / HLOOKUP: 一方のシートのデータをもとに、別のシートのデータを検索します。変数のコード化によく使います
=VLOOKUP(A1, $D$1:$E$2, 2, FALSE)
D1:E2の範囲に変換テーブルがあり、A1の値（1または2）に基づいて”女”または”男”を返します。

XLOOKUP: 一つの列（または行）から特定の値を検索し、対応する値を別の列（または行）から返します
INDEX / MATCH: VLOOKUPよりも柔軟な検索が可能です。
SUMIF / COUNTIF: 特定の条件に一致するデータを集計します。
SUMPRODUCT: 複数の配列の要素同士を掛け合わせた後、その総和を計算します。これは、条件付きの集計にも使えます。
CONCATENATE / &: 文字列を結合します。
LEFT / RIGHT / MID: 文字列から特定の文字を抽出します。
IF / IFERROR: 条件に基づいて異なる計算を行い、エラーが発生した場合にはデフォルト値を返します。
=IF(A1=1, “女”, IF(A1=2, “男”, “不明”))
セルA1が1であれば”女”、2であれば”男”、それ以外であれば”不明”と表示されます。

AND / OR: 複数の条件を組み合わせて評価します。
LEN: 文字列の長さを返します。
TRIM: 文字列から余分なスペースを取り除きます。

Q. 知っていると便利な関数（Excel 365以降）

ダイナミック配列関数はExcel 365以降で使用できる新しいカテゴリの関数で、一つのセルに式を入力するだけで、複数の出力値を隣接するセルに自動的に「スピルオーバー」させることができます。

FILTER関数: 条件に合った行や列を抽出
例えば、A1:C10の範囲に「ID」「名前」「年齢」という列がある場合、年齢が20歳以上の人物だけを抽出したいとします。この場合、以下のようなFILTER関数を使用します。　=FILTER(A1:C10, C1:C10 >= 20)

SORT関数: データをソート。年齢でソートする場合、以下のようにします。
=SORT(A1:C10, 3, 1)　この式は、A1:C10の範囲を3列目（年齢）に基づいて昇順（1）でソートします。

UNIQUE関数: 重複を削除・名前列から重複を削除する場合、以下のようにします。
=UNIQUE(B1:B10)

SEQUENCE関数: 連続した数値を生成。1から10までの数値を生成する場合、以下のようにします。
=SEQUENCE(10)

関数を組み合わせる: SORT + FILTER　年齢が20以上の人物を年齢でソートする場合、以下のように関数を組み合わせます。
=SORT(FILTER(A1:C10, C1:C10 >= 20), 3, 1)

Q. データ量が少ない時に注意すべきことは？

- まずは自院の課題が何であるか、分析を目的を明確に持つことが重要です。やみくもにデータを分析をしても得られるものが少なくなってしまいます。分析を行った結果、どのような行動変容に繋げられるかということを考えて分析します。
- 記述統計（人数、平均、中央値など）を経時的に追っていくことで比較可能になります
- データを可視化することで、データの傾向や問題点を直感的に把握することができます。
- 小数サンプルの場合、外れ値があると解釈に影響を与える可能性が高くなります。外れ値が発生した要因を検討しましょう
- 別のグループや時間帯と比較することで、見落としていた傾向やパターンを発見することがあります。
- ドメイン知識を活かして、数名で分析結果を話し合うと新たな視点が生まれます

Q. 分析でよく使われる指標は？

病院の理念がなにより大切。

医療の質
■QI　国立病院機構を参考に
■患者満足度調査。ケアの質を評価します

患者の状態
■年齢構成、疾病構造、重症度(介護度)、

業務指標(KPI)
■LOS、病床利用率、紹介・逆紹介率、新患数と再来

競合と市場分析
■対象領域における市場占有率
■患者の地域分布　地域別に患者数を分析し、マーケティング戦略を考えます。
■競合病院の動向(公開データの活用)

職員のパフォーマンスと満足度
■　離職率: 高い離職率は職員の不満や疲労、効率の低下を示す可能性があります。
■　職員満足度調査: モチベーションや職場環境に問題がないかを定期的に確認します。

DPCデータを分析しよう！

医療系のさまざまなデータ

レセプト
- 医科（入・外）
- DPC
- 歯科
- 調剤
- 特定健診等

- 医療機関由来と保険者由来

DPC
- 急性期医療機関ベース

医療情報システム
- 電子カルテ
- 部門システム
- 地域連携システム
- 院内サーベイ

統計
- 人口動態調査
- 生命表
- 患者調査
- 病院報告
- 医療施設調査
- ３師調査
- 国民医療費
- 受領行動調査
- 伝染病統計
- 社会福祉施設等調査
- 介護サービス・事業所調査
- 社会医療診療行為別調査
- 国民生活基礎調査

…etc

調査
- 業務調査
- 疾患レジストリ
- 住民コホート
- バイオバンク
- がん登録
- 臨床試験
- 質問紙等調査

その他
- バイオバンク
- ウェアラブルデバイス
- 健康アプリ
- SNS

研究法に係わらず，実臨床のデータから
強く実践を志向した研究(real-world evidence)

DPCデータとは

ファイル名	内容
様式１ 患者単位で把握	簡易診療録情報 (カルテのサマリーのような情報)
様式３ 施設単位で把握	施設情報 (病床数、届け出入院料、加算の算定状況など)
様式４ 患者単位で把握	医科保険診療以外の診療の有無に関わる情報
Ｄファイル 患者単位で把握	診断群分類点数表により算定した患者にかかる診療報酬請求情報 (DPC レセプトの情報) ⇒ DPC 対象病院のみ提出
EF 統合ファイル 患者単位で把握	入院患者の医科点数表に基づく出来高点数算定情報 (出来高レセプトの情報)
外来 EF 統合ファイル 患者単位で把握	外来診療患者の医科点数表に基づく出来高点数算定情報（外来の出来高レセプトの情報）
Ｈファイル 患者単位で把握	一般病棟用、特定集中治療室用、ハイケアユニット用の重症度、医療・看護必要度に関する情報

- 様式1は退院時サマリーのイメージ。性別、年齢、身長、体重、病名、入院目的や入院経路、手術術式、重症度などを含む。患者がどのような状態で入院し、主にどのような手術を受け、何日間入院して、どのような状態で退院していったか、概略が把握できる
- EFファイルは診療報酬明細書情報。医科点数表に基づく出来高による診療報酬の算定分を範囲とする。手術、検査、処方等の診療区分のほか、実施年月日や行為回数、診療明細名称や薬剤の使用数量・基準単位などの情報が収載されている。「どのような診療行為を」「いつ」「どれだけ」実施したかが詳しく分かるため、診療プロセスを時系列で把握できる
- Hファイルは重症度、看護必要度に係る調査票を書く評価項目の点数。様式1と異なり、1日ごとに情報を入力する。産科等の患者についても作成する必要がある

DPCデータで医療の質を評価する

アウトカム分析
- 診療の成果を分析して医療の質を評価する方法
- 手術成功率、合併症発生率、死亡率など直接的な結果を評価する指標なので、医療の質の良し悪しが分かりやすく判断しやすいが、公平な評価が難しい指標
- 診療行為の明細情報が入ったEFファイルなどの診療データを用いて、診療の内容をより細かく比較する診療プロセス分析を行うことができる

プロセス分析
- 診療内容の詳細を分析して医療の質を評価する方法
- ガイドラインの遵守率などを評価
- 特にアウトカムと密接に関連するようなプロセス評価が重視されている

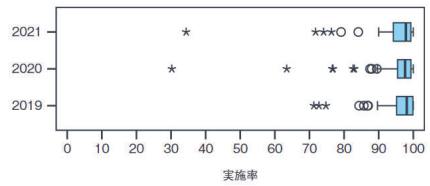
ストラクチャー分析
- 設備や人員体制の視点が主。診療報酬上の施設基準、法令に規定される人員、委員会等の整備等を評価する
- 医療機能の特性や目指している医療に合致する医療提供体制かどうか

DPCデータを活用して臨床指標を分析する

手術ありの患者の肺血栓塞栓症の予防対策の実施率を分析する(リスクレベルが高いリスク)

分子：分母のうち、肺血栓塞栓症の予防対策(弾性ストッキングの着用、間歇的空気圧迫装置の利用、抗凝固療法のいずれか、または2つ以上)を実施した患者数

分母：肺血栓塞栓症発症のリスクレベルが「高」の手術を施行した退院患者数



(年度)

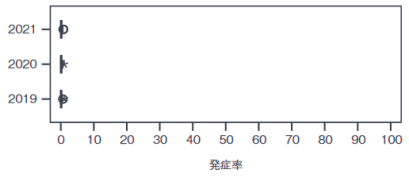
病院集計	2021	2020	2019
病院数	81	82	87
平均	95.2	95.1	96.2
標準偏差	8.9	9.4	5.7
中央値	97.9	97.6	98.1
25パーセンタイル	94.2	95.5	95.2
75パーセンタイル	99.3	99.3	99.9
目標値	95%以上		

DPCデータを活用して臨床指標を分析する

■75歳以上退院患者の入院中の予期せぬ骨折発症率

分子：分母のうち、当該入院の入院日から数えて2日目以降退院日までに骨折を発症した患者数

分母：75歳以上の退院患者数



(年度)

病院集計	2021	2020	2019
病院数	101	102	102
平均	0.2	0.2	0.1
標準偏差	0.2	0.2	0.2
中央値	0.1	0.1	0.1
25パーセンタイル	0.0	0.0	0.0
75パーセンタイル	0.3	0.3	0.2
目標値	0.2%以下		

第３版が発売されました！

第3版では2022年度診療報酬改定の内容に対応するとともに、今後ますます重要となるクオリティ・インディケーター（QI）の章を新たに追加し、さらにパワーアップしました。

★本書のポイント★

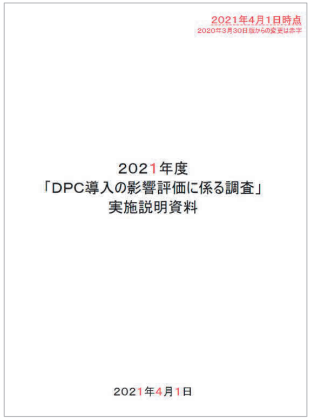
◎DPCデータの基礎知識からExcelを使ったデータ分析手法まで徹底指南
◎厚生労働省のオープンデータを用いた医療圏分析を解説
◎様式1やEFファイルなどのデモデータを用いて実際に手を動かしながら自院の診療機能や医療圏を分析できる
◎コード付き二次医療圏名などが入った「DPC調査対象病院リスト」など日常業務でも使える各種マスタがダウンロード可能
◎2022年度診療報酬改定の内容に対応。さらに医療の質を評価する「クオリティ・インディケーター（QI）」の章を追加

DPCデモデータの前処理（済）

- 様式1とEFファイル、Dファイルを連結した分析を行うため、1入院1患者となるような入院毎にユニークとなる分析用IDを作成します。様式1、EFファイル、Dファイルに分析用IDの列を追加します
  - ✓ 1人の患者が同年に複数回入院すると、複数のFF1データを持ちます
  - ✓ 様式1、EFファイル、Dファイル全てで同じ作業を実施
- DPCコード別の分析を行うため、DファイルからDPCコードを追加
  - ✓ 分析用IDを用いて、DファイルにあるDPCコードの情報を様式1に追加
- MDC（主要診断群分類）を追加
  - ✓ DPC14桁から頭2桁
- DPC6桁を追加
  - ✓ DPC14桁のうち上6桁が“DPC6”
- DPCコードから定義OPE（手術の有無）を追加
- 誕生年月日から年齢を計算
- “65歳以上高齢者”列を作成し、0と1でカテゴリー化
- 在院日数を計算
  - ✓ 様式1の開始日～終了日までの情報

変数 1	MDC、DPC6	「MDC」は主要診断群分類、「DPC6」は傷病分類に相当します。MDC は DPC コードの最初の 2 桁で、MDC を変数に加えることで診療科分類相当での集計が容易になります。DPC6 は DPC コードの最初の 6 桁で、この変数を加えておくと、病型による集計が容易になります。
変数 2	OPE（手術の有無）	手術のあり・なしは、DPC コードの上 9 桁目から 2 桁で分かります。この 2 桁が「99」なら手術なし、その他の番号は何かしらの手術が行われています。手術を行った場合は「1」、手術なしの場合は「0」のフラグを立てます。
変数 3	年齢	入院時点の年齢。生年月日と様式 1 開始日から算出します。
変数 4	年齢階級	10 歳刻みの年齢。この変数を加えることで、年代別の集計が容易になります。入院時の年齢を 10 歳刻みで分けます。
変数 5	65 歳以上高齢者	入院時の年齢が 65 歳以上の場合は「1」、65 歳未満の場合は「0」のフラグを立てます。
変数 6	在院日数	『様式 1』の「開始日」から「終了日」までの日数＊です。

実施説明資料をデータと照らし合わせて確認する



- 最初に見るのは変数定義のページ
- “入力要綱を参照”は詳細解説を別途見る
- 入力必須項目や条件指定を確認
- この資料は病院向け資料のため、実際に分析に使うデータが全てこの項目の通りに入力されているとは限らないことに注意

【ペイロード項目】

コード	ペイロード種別	レコード必須条件等	バージョン	連番	ペイロード番号	項目必須条件有	項目名	内容（入力様式等）
A000010	患者属性	○	ア	-	1	○	生年月日	0～9からなる 8 桁の数字　YYYYMMDD 例　1970 年 5 月 1 日→19700501
					2	○	性別	1. 男　2. 女
					3	○	患者住所地域の郵便番号	0～9からなる 7 桁の数字 例　〒100-8916→1008916　※前ゼロの必須
A000020	入院情報	○	ア	-	1	○	入院年月日	0～9からなる 8 桁の数字　YYYYMMDD 例　2021 年 4 月 1 日→20210401
					2	○	入院経路	入力要領を参照
					3	当A	他院よりの紹介の有無	0. 無　1. 有
					4	当A	救急の外来からの入院	0. 無　1. 有
					5	当A	予定・救急医療入院	入力要領を参照
					6	当A	救急車による搬送の有無	0. 無　1. 有
					7	当A	入院前の在宅医療の有無	0. 無　1. 当院が提供　2. 他施設が提供　9. 不明
					8	当B	自傷行為・自殺企図の有無	0. 無　1. 当院が提供　2. 他施設が提供　9. 不明 1. 服薬　2. 飛び降り・飛び込み　3. 服毒　4. 過量服薬　5. 刃物等による自傷（手首自傷を除く）　6. 手首自傷　7. その他　9. 無

・A000020　入院情報

① 入院年月日

ヘッダ項目の“入院年月日”と同じ値を入力する。

② 入院経路

入院にいたる経路について、該当する値を入力する。

子様式1で他病棟からの転棟があった場合は“0”院内の他病棟からの転棟を入力する。

値	区　分
0	院内の他病棟からの転棟
1	家庭からの入院
4	他の病院・診療所の病棟からの転院
5	介護施設・福祉施設に入所中
8	院内で出生
9	その他

EFファイルを使った分析

Q6 院内で急性心筋梗塞の診療ガイドラインがどれだけ遵守されているのか確認したいと考えています。何かいい方法はありませんか？

A6 基本的な対応策として、診療ガイドラインで推奨されている薬剤投与の施行率を調べてみるのは1つの手です。薬剤の投与実態を調べて、診療の適正化や質の改善につなげましょう

操作概要

ここでは、一例として、日本循環器学会がまとめた「ST 上昇型急性心筋梗塞の診療に関するガイドライン（2013 年改訂版）」で推奨されている HMG-CoA 還元酵素阻害薬（スタチン）の投与の施行率を調べてみることにします。付録 CD-ROM に収載した「スタチンマスタ」という Excel ファイルを活用して、まずは、『EF ファイル』で、スタチンが投与されているケースを特定します。その情報と『様式 1』のデータをリンクさせて、急性心筋梗塞患者への投与実態について集計します。

① スタチン**マスタ**を使って、EFファイルでスタチンが投与されている患者を特定する
② スタチンが投与された患者の分析用IDのリストを作成する
③ ②のリストを使って、同じ分析用IDの患者に1のフラグをたてる
④ ピポット機能を使って、急性心筋梗塞でスタチンが投与された患者の数を集計する

EFファイルを使った分析

主傷病ICD10コード	個数 / 分析用ID	個数 / スタチン患者	急性心筋梗塞患者数	スタチン投与患者数	割合
I210	17	12	30	19	63.3%
I211	9	4			
I214	1	1			
I219	3	2			
総計	30	19			

- ✓ 診療ガイドラインが遵守されているか、あるいはどれくらい逸脱した例があるかを検証する作業は、診療の適正化を図る上で、非常に重要な道筋と言える。心筋梗塞の再発防止に脂質管理を行うことは、高いエビデンスがあり、日本循環器学会がまとめた「ST 上昇型急性心筋梗塞の診療に関するガイドライン（2013 年改訂版）」では、「禁忌がなければ発症後早期からLDLコレステロール値にかかわらずHMG-CoA 還元酵素阻害薬（スタチン）を開始する」ことが推奨されている
- ✓ 今回の例では、主傷病名が急性心筋梗塞の患者に対するスタチン投与の施行率を集計するにとどめたが、スタチンが投与されなかった患者について院内調査を行い、非投与となった理由について調べ、以後、同様のケースで改善策が図れるかどうかなどもぜひ検討したい
- ✓ ただし、EF ファイルには、持参薬など、診療報酬の請求に関係のないデータは入力されていない点に注意が必要

分析時に参照する参考資料の例