

DPCデータを用いて論文を書こう (基礎編)

産業医科大学
公衆衛生学教室
松田晋哉

DPCデータとは何か

- 分析可能な全国統一形式の**患者臨床情報**
+ **診療行為**の電子データセット
- **患者臨床情報**
 - 患者基本情報
 - 病名、術式、各種のスコア・ステージ分類
- **診療行為情報**
 - 診療行為、医薬品、医療材料
 - 実施日、回数・数量
 - 診療科、病棟、保険種別

資料: 藤森研司

DPCデータから何が分かるか

- **患者の臨床情報**
 - 全国共通の「**簡易退院サマリ**」
- 「いつ」「何を」「どれ程」行ったか
 - (誰がオーダー、どの診療科・病棟の)
- **診療行為を時系列で把握**
 - レセプト情報から自動的、電子化
- **診療プロセスの可視化**
 - 平均像とバラツキ

資料: 藤森研司

DPCデータの分析は面白いけれど・・・

- 日常業務としての経営層へのデータ提供
 - どのくらい意味が分かっているのだろうか？
 - どのくらい活用されているのだろうか？
 - 他の施設ではどのような分析を行っているのか？
 - 自分のやっていることは正しいのか？
- データ分析をしていて気づいた「興味ある」知見を他の人と共有したい
- DPCデータを扱う者としての自分のこれからのキャリアパスは？

学会発表までは行くのだけれど・・・

- 学会発表は定期的に行っているけれど、どうも発展性がない
 - そもそも10分程度の発表では思いが伝えきれない
 - 質疑応答でアイデアをもらうのだけれど、そのまま対応ができていない

↓

論文にして、自分の思いを世に問うてみたらどうでしょう？

論文のネタはいたるところにある

脳外科部長曰く・・・
「Barthel Indexの改善度を医療評価指標に使うってホントかなあ。うちはさあ、重症患者が多いからBIの改善率悪くなっちゃうんだよねえ。DPCの一律評価は問題だよなあ...」
「森先生に指標の検討は慎重にねって言っとかなきゃな」

「重症って、どんな患者が多いんですか？」

「高齢者とか、合併症があったり意識障害の強い患者。不整脈系とか脳塞栓が多いでしょ。俺って腕がいいので有名じゃん。だから、救急隊も、重症例うちに送ってくるんだよねえ。頑張ってるんだから、もっと給料欲しいよなあ。院長に言っといて。」

「有名・・・？年齢や病型の違いはどのくらい在院日数や予後に影響してるの？」

臨床家の愚痴の中にも、データ解析のネタがある。

まず、データを準備する

J	A	B	C	D	E	F	G	H
1	PecID	zip	ambul	年齢	MDC	dpc_mdof	dpcod	los
2	095210074520081	0	77	06	060020	060200x04x0x	10	
3	095210074520084	1	65	06	060030	060300x03x0x	17	
4	095210074520094	0	74	06	060020	060200x08x0x	10	
5	095210074520095	0	20	11	110080	110080x07x00x	28	
6	0952100745210135	0	9	08	080070	080070x07xxxx	6	
7	0952100745230819	0	37	12	120020	120020xxxxxxx	10	
8	0952100745270041	0	65	06	060100	060100x02x0x	4	
9	0952100745290434	0	50	05	050050	050050x08f0x	2	
10	0952100745214214	0	10	04	040130	040130x08x00x	2	
11	0952100745290201	0	61	11	110080	110080x07x02x	6	
12	0952100745290607	0	72	03	03001x	03001x0f000x	7	
13	0952100745280012	0	61	12	120030	120030x0f000x	12	
14	0952100745270041	0	59	05	050050	050050x02x0x	18	
15	0952100745280014	0	55	11	110080	110080x08f0x	3	
16	0952100745230082	0	54	05	050070	050070x0f0x	4	
17	0952100745291147	0	39	06	060180	060180x0f0x	12	
18	0952100745230807	0	71	06	060040	060040x0f100x	18	
19	0952100745230812	1	69	06	060210	060210x08x00x	4	
20	0952100745294006	0	39	13	130040	130040x07x1xx	22	
21	0952100745490006	0	43	05	050050	050050x02x0x	19	
22	0952100745210121	0	27	12	120020	120020x08x40x	3	
23	0952100745280211	1	49	16	161070	161070xxxx00x	4	
24	0952100745290002	0	29	12	120020	120020x02xxxx	6	
25	0952100745200043	0	42	02	020160	020160x07xxxx0	10	
26	0952100745230822	0	55	07	070470	070470x08x00x	8	

様式1データにDPCコードを追加して、エクセルに取り込む。
年齢は生年月日と入院年月日を使って、在院日数は入院年月日と退院年月日から計算。

脳梗塞の患者だけ取り出す (DPC6桁=010060)

J	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	BasicID	hospcod	sex	purpose	status	admpath	urgenc	admission	admission	admission	admission	admission	admission	admission	admission
2	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
3	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
6	0	IC	2	0	0	0	0	0	0	0	0	0	0	0	0
7	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
8	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
9	0	IC	2	0	0	0	0	0	0	0	0	0	0	0	0
10	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
11	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
12	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
13	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
14	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
15	0	IC	2	0	0	0	0	0	0	0	0	0	0	0	0
16	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
17	0	IC	2	0	0	0	0	0	0	0	0	0	0	0	0
18	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
19	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
20	0	IC	2	0	0	0	0	0	0	0	0	0	0	0	0
21	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
22	0	IC	2	0	0	0	0	0	0	0	0	0	0	0	0
23	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
24	0	IC	2	0	0	0	0	0	0	0	0	0	0	0	0
25	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
26	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
27	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0
28	0	IC	1	0	0	0	0	0	0	0	0	0	0	0	0

ICD10を使うと脳梗塞を
アテローム血栓性脳梗塞 (I630, I633)、
ラクナ脳梗塞など (I638, G467)、
脳塞栓症 (I631, I634)
に分類できる。

最初のハードル: 学会発表や論文執筆のためには統計学的検討が不可欠...

- 統計学的分析は難しくありません。
 - 分析のためのソフトウェアはいいものがたくさんあります。
 - 一番重要なのは「生データをきちんとみること」、そして記述的な分析をきちんと行うこと(いきなり多変量解析をしらない)
 - 「どのようなデータを用いて、何を知らうとしているのか」によって用いる統計学的分析はおのずと決まります。

統計学的分析の意味

- 統計学的分析の結果は絶対的なものではなく、あくまで補助的なものととどまる。統計学的分析の結果の妥当性・精密性は対象の数、データの特性、等によって動く。
- 統計学的な結果はあくまで確率的なものであり、考察は常に当該学問における妥当性などと比較の上でなされなければならない。

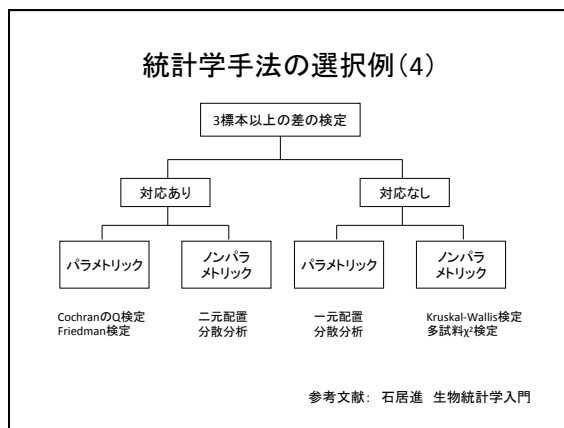
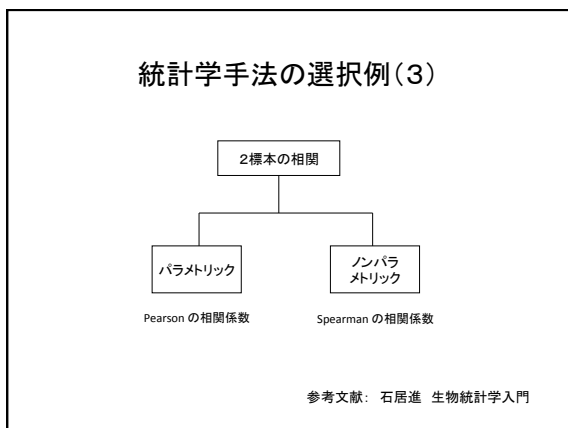
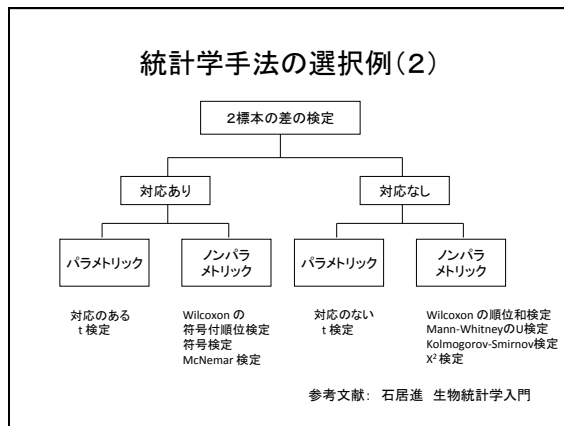
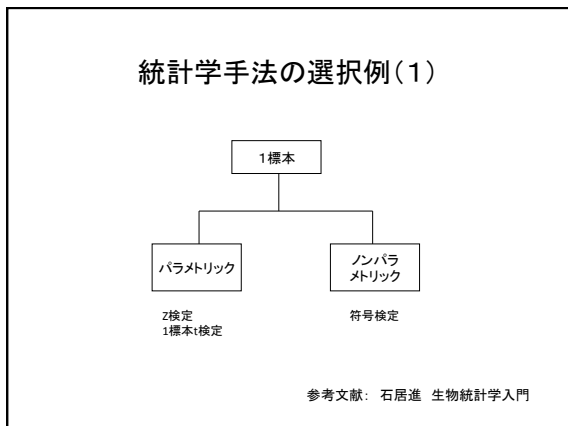
データの種類

- 分類(名義)尺度
 - 例: 性、都道府県、施設
- 順位尺度
 - 例: がんのStage分類
- 間隔尺度
 - 例: 連続的量的変数で差だけが意味を持つもの(手術年月日など)絶対的0点なし(マイナスの値が存在)
- 比率尺度
 - 例: 連続的量的変数で差と比が意味を持つもの(在院日数など)絶対的0点あり(マイナスの値が存在しない)

目的変数データの種類の検定手法

- 分類(名義)尺度
 - 分布を推計する母数(パラメーター)がない→パラメトリック検定ができない。ノンパラメトリック検定
- 順位尺度
 - 母数(パラメーター、例: 中央値と偏差)で分布が推計できる→パラメトリック検定ができる
- 間隔尺度
 - ただし、3,4でもN数が30未満の時はノンパラメトリックを使うのが一般的
- 比率尺度
 - パラメトリック検定が可能なデータは、ノンパラメトリック検定も可能。ただし、有意差の検出力が落ちる

一応の確認: $y = f(x) + \alpha$ 式におけるyが目的変数、xが説明変数



難しそうな統計学的分析をする前に

まず、単変量で分布を見る。
次いで目的変数と説明変数の関係を見る。

DPCデータを用いた分析例(1)

病型による在院日数の違い

病型	度数	平均値	標準偏差
アテローム血栓性脳梗塞	12,407	24.3	21.6
ラクナ梗塞	13,966	21.8	21.4
脳塞栓	6,337	29.5	25.8
合計	32,710	24.3	22.6

P<0.01; 一元配置分散分析

病型による年齢の違い

病型	度数	平均値	標準偏差
アテローム血栓性脳梗塞	12,407	72.8	11.7
ラクナ梗塞	13,966	72.2	12.5
脳塞栓	6,337	76.4	11.2
合計	32,710	73.2	12.1

P<0.01; 一元配置分散分析

DPCデータを用いた分析例(2)

病型による入院時Barthel Indexの違い

病型	度数	平均値	標準偏差
アテローム血栓性脳梗塞	7,934	50.2	40.2
ラクナ梗塞	8,744	53.8	39.6
脳塞栓	3,897	33.1	39.6
合計	20,575	48.5	40.6

P<0.01; 一元配置分散分析

病型による退院時Barthel Indexの違い

病型	度数	平均値	標準偏差
アテローム血栓性脳梗塞	7,397	72.4	36.7
ラクナ梗塞	8,014	75.8	35.3
脳塞栓	3,323	59.0	42.7
合計	18,734	71.4	37.8

P<0.01; 一元配置分散分析

様式1・E file・F file からの分析用データセット作成

F file

- F-1 施設コード
- F-2 データ識別番号
- F-3 退院年月日(西暦)
- F-4 入院年月日(西暦)
- F-5 データ区分
- F-6 病期番号
- F-7 行為明細番号
- F-8 病歴点数マスターコード
- F-9 レセプト電算処理システム用コード
- F-10 病期番号(基本)
- F-11 診断明細名称
- F-12 使用量
- F-13 基本単位
- F-14 行為明細点数
- F-15 行為明細薬剤
- F-16 行為明細材料
- F-17 付・点区分
- F-18 出来高実績点数
- F-19 出来高・点別フラグ

E file

- E-13 行為材料
- E-14 付・点区分
- E-15 行為明細
- E-16 病期番号
- E-17 レセプト種別コード
- E-18 退院年月日
- E-19 レセプト科区分

様式1

SQL

分析用ファイル

施設	データ識別番号	退院年月日	入院年月日	データ区分	病期番号	行為明細番号	病歴点数マスターコード	システム用コード	病期番号(基本)	診断明細名称	使用量	基本単位	行為明細点数	行為明細薬剤	行為明細材料	付・点区分	出来高実績点数	出来高・点別フラグ	
00000	450	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

DPCデータを用いた分析例(3)

病型による出来高換算点数(合計)の違い

病型	度数	平均値	標準偏差
アテローム血栓性脳梗塞	12,407	93,886.0	71,445.9
ラクナ梗塞	13,966	81,772.2	69,703.4
脳塞栓	6,337	116,624.5	87,976.0
合計	32,710	93,119.0	75,312.0

P<0.01; 一元配置分散分析

病型による出来高換算点数(注射)の違い

病型	度数	平均値	標準偏差
アテローム血栓性脳梗塞	12,407	17,359.5	14,986.2
ラクナ梗塞	13,966	13,581.2	14,743.2
脳塞栓	6,337	17,615.0	18,275.2
合計	32,710	15,800.1	15,694.7

P<0.01; 一元配置分散分析

DPCデータを用いた分析例(4)

病型による出来高換算点数(画像診断)の違い

病型	度数	平均値	標準偏差
アテローム血栓性脳梗塞	12,407	6,465.2	6,354.1
ラクナ梗塞	13,966	5,157.7	4,475.8
脳塞栓	6,337	6,922.2	5,496.4
合計	32,710	5,995.5	5,501.6

P<0.01; 一元配置分散分析

病型による出来高換算点数(リハその他)の違い

病型	度数	平均値	標準偏差
アテローム血栓性脳梗塞	12,407	9,766.8	15,568.5
ラクナ梗塞	13,966	7,946.7	13,586.8
脳塞栓	6,337	11,248.5	17,547.8
合計	32,710	9,276.8	15,231.3

P<0.01; 一元配置分散分析

「統計学的分析はやっぱり多変量回帰分析をやらないと論文にならないよね・・・」

これは誤解です。しかし、多くの人がこの誤解のもとで論文を書いているので「郷に入れば郷に従う」気持ちでやらないといけな現実もあります。

単変数での記述的分析

- 外れ値のチェック
- モデルに入れるべき変数の選択
- 多重共線性への配慮

適切なモデルでの多変量解析

何をやらうとしているのか再確認

- 探索的分析 or モデルの検証

多くの場合はこちら

多重共線性とは？

説明変数の間に強い相関関係が存在する場合、回帰分析により得られる結果に悪い影響が与えることがあります。

具体的には、

- 同時に用いる説明変数を変更すると回帰式の係数が大きく変化してしまう
- 決定係数(回帰式の説明力)が高い一方で各変数のt値が低く、解釈が難しい
- 想定していた符号と異なる結果がでる

などの現象が生じることがあります。これを「多重共線性」と言います。

多重共線性を見つける1つの方法はVIF(Variance Inflation Factor, 分散拡大要因)を計算することです。説明変数が x_1, x_2 という2変数の場合のVIFは以下の式で求められます。

$$VIF = 1 / (1 - r^2_{x_1, x_2})$$

VIFが大きいほど、多重共線性の影響があります。こゝで r_{x_1, x_2} は x_1 と x_2 の相関係数の2乗です。10より大きいVIFであれば、明らかに多重共線性が存在します。多重共線性がある場合は、どちらかの変数を回帰式から除くのが一般的です。

DPCデータを用いた分析例(5) BIの改善に関連する要因の分析 (“010060x099x3xx”リハあり症例のみ、重回帰分析：投入法)

	非標準化係数		標準化係数		t値	有意確率
	B	標準誤差	ベータ			
(定数)	104.480	2.060			50.712	0.000
sex	-3.585	0.610	-0.053		-5.875	0.000
入院時年齢	-0.594	0.026	-0.210		-22.951	0.000
BI前	-0.516	0.008	-0.609		-66.252	0.000
リハ日数	0.070	0.033	0.036		2.137	0.033
リハ開始日	-0.291	0.095	-0.029		-3.072	0.002
入院日数	-0.335	0.033	-0.179		-10.238	0.000
脳塞栓ダミー	-2.081	0.943	-0.019		2.206	0.027

従属変数: BI変化
説明変数: SEX 1=男、2=女
BI前 入院時のBarthel Index
脳塞栓ダミー 0=脳塞栓以外、1=脳塞栓

“010060x099x3xx”: 脳梗塞・手術なし・エダラボンあり

疫学とは？

「人間集団を対象¹⁾として、人間の健康およびその異常²⁾に関連する要因を、宿主、病因、環境の各方面から包括的に考究³⁾し、その増進と予防をはかる⁴⁾学問」

- 1) 対象は集団、したがって取り扱いには統計学的にならざるを得ない
- 2) 何を指標として計測を行うのか：統計学的手法を規定
- 3) 解釈は生物学的、社会科学的に合理性がなければならない。
- 4) 何を指標に評価を行うのか？

妥当な比較を行うための手法、それが疫学

では、論文を書きましょう(1)

- **まず、論文のアウトラインを設計します**
 - 目的と仮説の確認
 - 何を明らかにしたいのか？
 - 既存研究で何がわかっていて、何がわかっていないのか？
 - 対象及び方法
 - 上記目的・仮説を検証する上で妥当な対象と方法なのか、再確認します
 - 結果
 - 上記目的・仮説を検証する上で必要な図表を考えます
 - 考察
 - 6つの“C”をしっかりとおさえます。
- **投稿したい雑誌のフォーマットを確認します**

では、論文を書きましょう(2)

- **6つの“C”とは？**
 - 第1段落：“Clarify”
 - この論文の結果で一番大切なことをまとめる。
 - 第2段落：“Compare and contrast”
 - この論文の結果は他の研究、これまでの仮説とあっているか？
 - もしあっていなければ予想外の結果について考えられる理由を挙げて考察をします。
 - 第3段落：“Contemplate”
 - 結果について説明を加えます。臨床医学的、社会学的など、科学的理論から結果を解釈します。

参考: http://ryok.cocolog-nifty.com/mph/2007/10/post_3946.html



では、論文を書きましょう(2)

- **6つの“C”とは？**
 - 第4段落：“Contribution”
 - 臨床的意義、研究としての意義
 - 治療に対する貢献
 - 今後の研究に対する展望
 - あまり一般化しすぎないことが重要。どんなに、あなたが素晴らしい結果だと思っても、この研究ですべてが解決することはありません。
 - 第5段落：“Cons” バイアスについて考えます
 - 研究のデザインに問題はないか？
 - 研究対象者の代表性に問題はないか？
 - データ収集に問題はないか？
 - 第6段落：“Conclusion”
 - 明確なメッセージを伝えることが重要です。

参考: http://ryok.cocolog-nifty.com/mph/2007/10/post_3946.html

引用文献

- **引用文献は重要です。**
 - 医中誌やPubMedを使って重要な論文を検索し、考察に役立てます。
 - わからなければ、院内のDrに聞いてみましょう。

投稿する

- 投稿する前に
 - 分析結果から導けないことまで、書きすぎていないか、もう一度冷静になって推敲します。
 - 抄録は魅力的でしょうか？
 - フォーマットが投稿する雑誌にあっているか、もう一度確認します。
- そして・・・
 - 多くの場合、皆さんの思いを打ち砕くような、「辛い」査読結果が返ってきます。でも、くじけないでください。
 - 査読してくれたことに感謝し、コメントの一つずつ丁寧に答えます。辛い作業ですが、このやり取りが糧になります。
 - どうしてもダメだったら、論文の投稿先を変更します。ラストリゾート的な雑誌は必ずあります。
 - 印刷されたら、読んでほしい人に別刷りを配ります。

まとめ

- 日常業務の中でDPCデータを分析しているだけでは、ある種の「行き詰まり感」を覚えるようになります。
- みずからの能力開発のためにも、DPCデータを使って論文を書いてみましょう。
 - 最初は物まねでもいいかもしれません。
 - アイデアは皆さんの日常業務の中にあります。
 - 論文を書くという作業を通して、皆さんのスキルは必ず向上します。
- 皆さんが論文を書くことで、DPCデータの精度が向上し、そして制度の質も向上します。