Tidyverse を活用したデータ前処理の実践

静岡市立清水病院　呼吸器内科

森　　和貴

MORI　Kazutaka, M.D., Ph.D.

自己紹介

- 医師 21年め（臨床研修制度 1期生）
- 専門：呼吸器内科
- R遍歴：15年くらい
  - 2009年頃～
  - 2011年　　RStudio が登場、飛びつく
  - 2012年　　大学院の後輩から EZR を知る
  - 2017年頃～　Windows版 RStudioでの日本語対応に疲れ、仮想化に走る
    - ✓ そこから、解析の再現可能性に関心を持つようになりました
    - ✓ 2023年頃からRStudio Desktop版での日本語対応が改善し、Desktop版もまた使うようになりました

best subset selection procedure using Akaike's information criteria (AIC). All statistical analyses were performed using R version 2.11.1 (The R Foundation for Statistical Computing, Vienna, Austria, 2010). A value of $p < 0.05$ was considered significant for the results of all statistical analyses, and all tests were 2-sided.

はじめての論文
Mori K, *et al.* COPD. 2011

データの前処理とは

- 集計や分析に用いる生データを整えて加工すること全般をさす
- **「分析にかかる時間の8割は前処理の時間」**といわれる
- 主な工程：
  - 収集した生データを **コンピュータ／Rで処理できる状態** に整形する
  - **データクリーニング（クレンジング）**
    - ✓ 欠損値、外れ値などの確認・処理
  - **データ変換**
    - ✓ データの「型」を整える　（数値、順序変数、あり・なしの2値、など）
    - ✓ データの正規化・標準化　（表記のブレを揃える、など）
    - ✓ 匿名化　（年齢などを生の値→階級値に変換する操作も含む）

本日の内容

- **Tidyverse とパイプ演算子のおさらい**
- **データ整形の指針**

  ～本当はデータファイルを作る前に確認しておきたいこと～
- **初級編：最新の癌種別死亡数のグラフを作りたい**

  ～人口動態統計のCSVファイルを整形してみる～
- **中級編：臨床研究のまとめファイルがやってきた**

  ～“ネ申エクセル”と戦うための基礎知識～

本日の内容

- Tidyverse とパイプ演算子のおさらい
- データ整形の指針

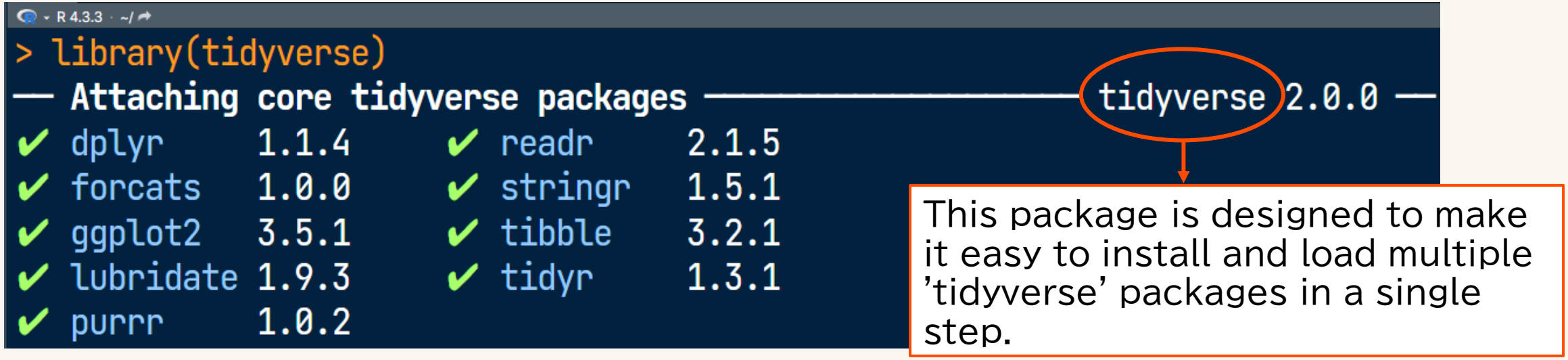
  ～本当はデータファイルを作る前に確認しておきたいこと～
- 初級編：最新の癌種別死亡数のグラフを作りたい

  ～人口動態統計のCSVファイルを整形してみる～
- 中級編：臨床研究のまとめファイルがやってきた

  ～“ネ申エクセル”と戦うための基礎知識～

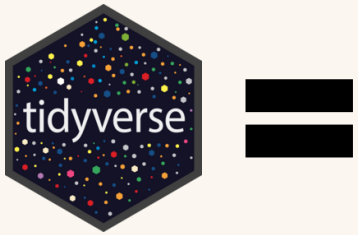
Tidyverse とは

- Posit（旧 RStudio）社の Hadley Wickham 氏らを中心に開発されているデータ処理のための R パッケージ群

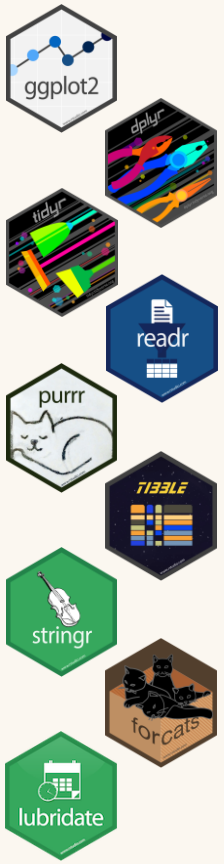


- **library(tidyverse)** で 9つの “core” パッケージ が読み込まれる。更に、インストールされているがこの時点では読み込まれていない多数の関連パッケージも控えている状態

Tidyverse の構成パッケージ



=



Core packages

パッケージ	用途
ggplot2	データ可視化
dplyr	データの加工、整理
tidyr	データフレームの整形
readr	データの読み書き
purrr	繰り返し処理
tibble	改良版のデータフレーム
stringr	文字列の処理
forcats	factor型変数(順序尺度)の処理
lubridate	日時データの処理

+

readxl
Excelファイルの読込み

googlesheets4
Googleスプレッドシートの読込み

haven
SPSS, Stata, SAS のデータファイルの読込み

magritter
特殊なパイプ

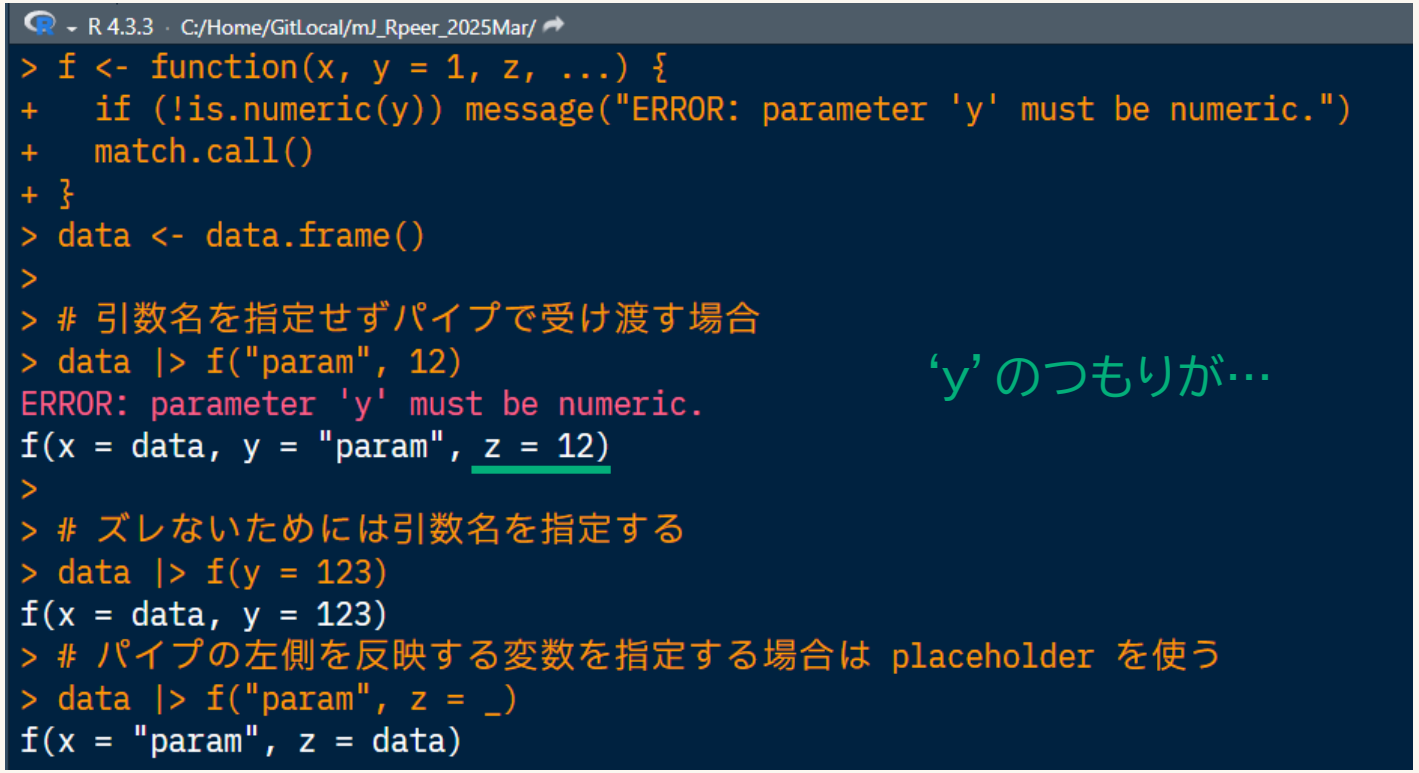
⋮

パイプ演算子

パイプの基本：左側を右側の関数の1つ目のパラメーター(引数)に挿入

```
data %>% function(param) ⇒ function(data, param)
```

```
data %>% f(p1) %>% g(p2) ⇒ g(f(data, p1), p2)
```



右側の関数に元々指定されていた引数がひとつずつ後ろにずれることに注意。
エラーの原因となりやすいので、特にパイプを使うときは**できるだけ引数名を省略しない**ようにした方が安全。

2つのパイプ演算子

- 現在、主に使われているパイプ演算子は2種類存在する
- 基本的な挙動はほぼ同様なので、他の用途で tidyverse を読み込んでいるかやコーディング規約などで選択、読み替えて使用可能

Magritter pipe	名称	Base pipe
%>%	使用する記号	**|>**
必要（tidyverse / magritter）	パッケージ読込み	不要（R 4.1以降）
関数以外も可（例: `x %>% .[1]`）	パイプの右側	関数のみ
`.` （複数可、引数名の省略可）	第1引数以外 への受け渡し （placeholder）	`param = _` （1個だけ、引数名の省略不可）
可能 （`%>% return()` でパイプを終了）	return() の使用	`|> return()` は使用できない

magritter の特別なパイプ（一部）

Tee pipe : %T>%

- 2つの関数にデータを受け渡す
- 途中経過の確認に有効

```
R 4.3.3 · ~/
> library(tidyverse)
> library(magrittr)
>
> tibble(var1 = rnorm(3), var2 = rnorm(3)) %T>%
+   print() %>%
+   summarise(var1 = sum(var1), var2 = sum(var2))
# A tibble: 3 × 2
     var1    var2
    <dbl>   <dbl>
1  0.0891 -0.377
2  1.20   -0.485
3 -0.558  -0.771
# A tibble: 1 × 2
   var1  var2
  <dbl> <dbl>
1 0.731 -1.63
```

tibble(…) %>% print() の結果

tibble(…) %>% summarise(…) の結果

Exposition pipe : %$%

- 目的の関数にデータフレームを受け取る機能がない場合に有効

```
R 4.3.3 · ~/
> diamonds %$% cor.test(price, carat)

        Pearson's product-moment correlation

data:  price and carat
t = 551.41, df = 53938, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9203098 0.9228530
sample estimates:
      cor
0.9215913
```

cor.test() 関数には data= がないので、通常は
cor.test(diamonds$price, diamonds$carat)

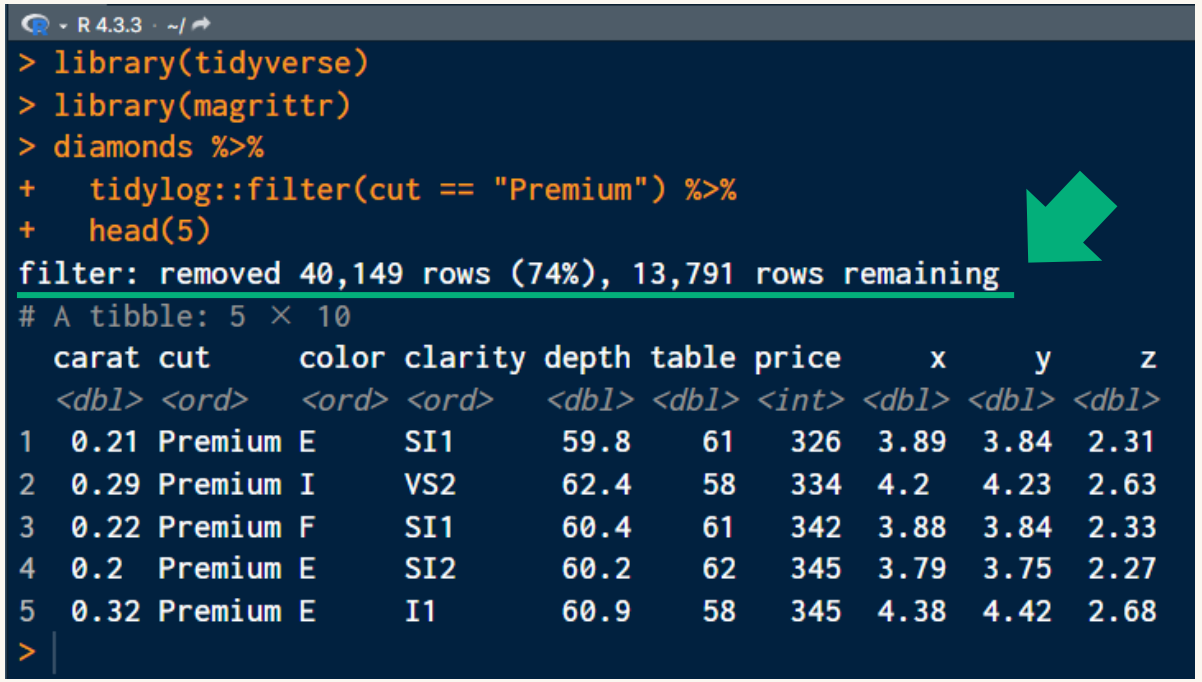
Tidyverse のお供におすすめ

- Posit社提供の **cheat sheet 集**

  https://posit.co/resources/cheatsheets/

  または、RStudio で Help – Cheat Sheets（上記の一部）

- **tidylog パッケージ**
  - dplyr, tidyr の主要な関数に簡単な処理結果の報告を追加
  - `library(tidylog)` でロードせず、`tidylog::function()` で確認後に必要に応じて `dplyr::function()` に戻す方が使い勝手が良い(私見)



- magritter パッケージの **Tee pipe** で `print()` や `View()` を活用

本日の内容

- Tidyverse とパイプ演算子のおさらい
- **データ整形の指針**

  ～本当はデータファイルを作る前に確認しておきたいこと～
- 初級編：最新の癌種別死亡数のグラフを作りたい

  ～人口動態統計のCSVファイルを整形してみる～
- 中級編：臨床研究のまとめファイルがやってきた

  ～“ネ申エクセル”と戦うための基礎知識～

データ整形の指針

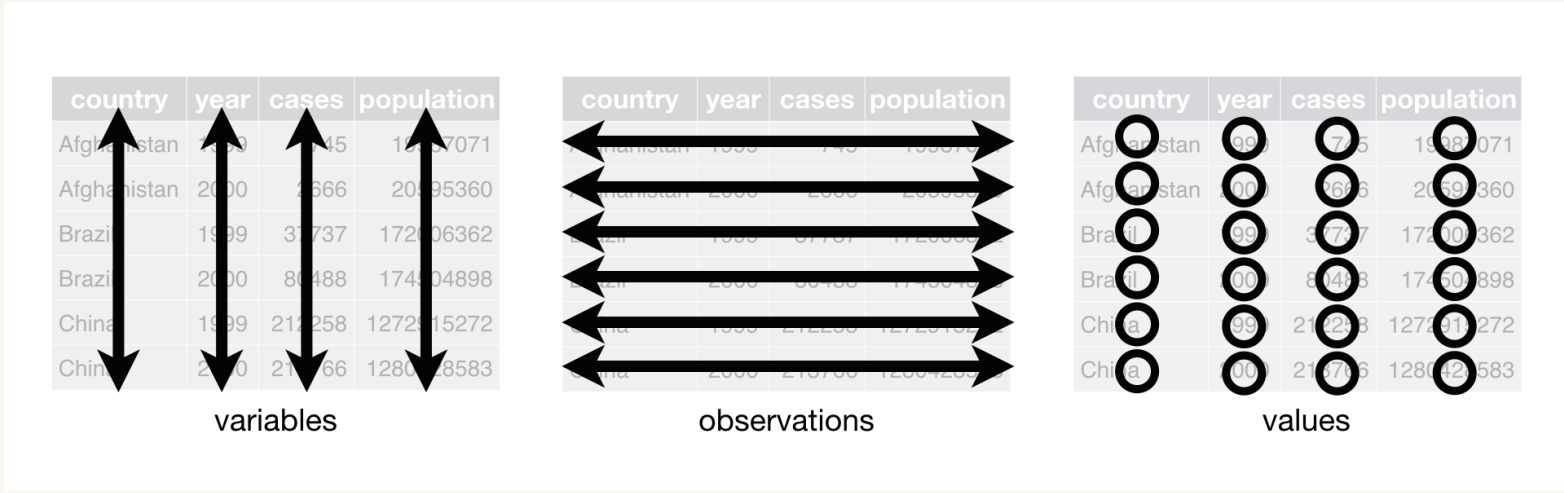
- 「見る」あるいは「見せる」ための表や雑然とした表を、コンピュータで認識できて（機械判読可能）、解析に使用できる表に変形・変換する
- 大まかな目標となる代表的なルール
  - 総務省「**統計表における機械判読可能なデータの表記方法の統一ルール**」
    https://www.soumu.go.jp/menu news/s-news/01toukatsu01 02000186.html
  - **整然データ** Tidy data
    - ✓ Wickham H. Journal of Statistical Software. 59: 1-23, 2014
    - ✓ Tidyverse はこの思想に基づいて設計されており、これらのパッケージをフル活用するためには tidy data を意識することが大切
- 元ファイルは極力編集せず、編集の内容や経過が後から確認できるようコードに残すようにする

総務省によるデータ表記方法の統一ルール

- 2020年に、政府統計の総合窓口(e-Stat)に掲載する統計表におけるデータ表記方法の統一ルールとして総務省が策定したもの
- Excelなど特定のソフトや政府統計に依存しない部分を抜粋すると
  - ✓ 1セル1データとなっている
  - ✓ 数値データは数値属性とし、文字列（注:単位、注釈など）を含まない
  - ✓ スペースや改行等で体裁を整えていない
  - ✓ 項目名等を省略していない （注:「薬剤A」「B」「C」は「薬剤A」「薬剤B」「薬剤C」とする）
  - ✓ データの単位を記載している（注:数値とは別のセルに記載している）
  - ✓ 機種依存文字を使用していない
  - ✓ データが分断されていない・1シートに複数の表が掲載されていない

整然データ Tidy data

- Tidyverseの作者が提唱した概念で、「データの構造（structure）」と「意味（semantic）」を一致させることを目指している
  - ✓ 個々の変数（variable）が1つの列（column）をなす
  - ✓ 個々の観測（observation）が1つの行（row）をなす
  - ✓ 個々の値（value）が1つのセル（cell）をなす
  - ✓ 個々の観測の構成単位の類型（type of observational unit）が1つの表（table）をなす



Wickham H., et al.
R for Data Science (2e).
https://r4ds.hadley.nz/

本日の内容

- Tidyverse とパイプ演算子のおさらい
- データ整形の指針

  ～本当はデータファイルを作る前に確認しておきたいこと～

- **初級編：最新の癌種別死亡数のグラフを作りたい**

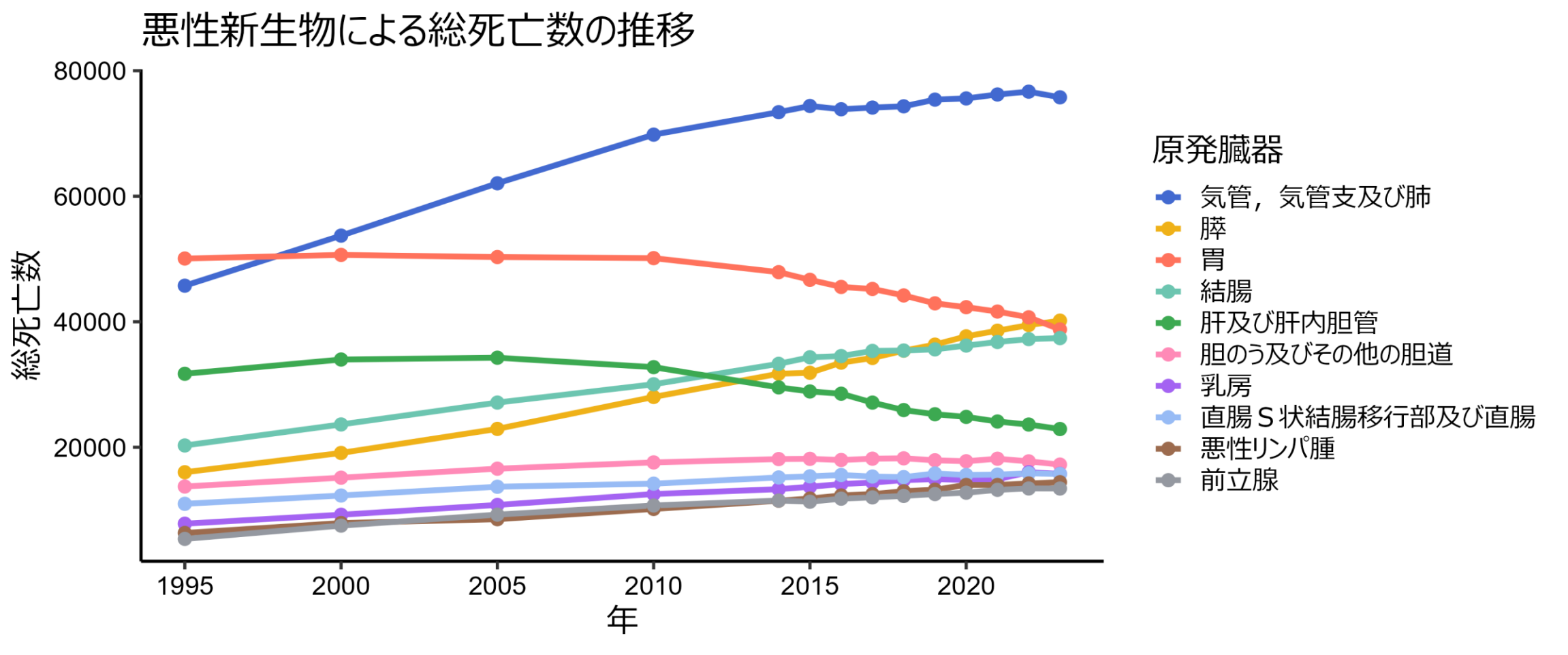
  ～人口動態統計のCSVファイルを整形してみる～

- 中級編：臨床研究のまとめファイルがやってきた

  ～“ネ申エクセル”と戦うための基礎知識～

初級編のゴール

厚労省 人口動態統計のCSVファイルを整理して、下の図を作成する

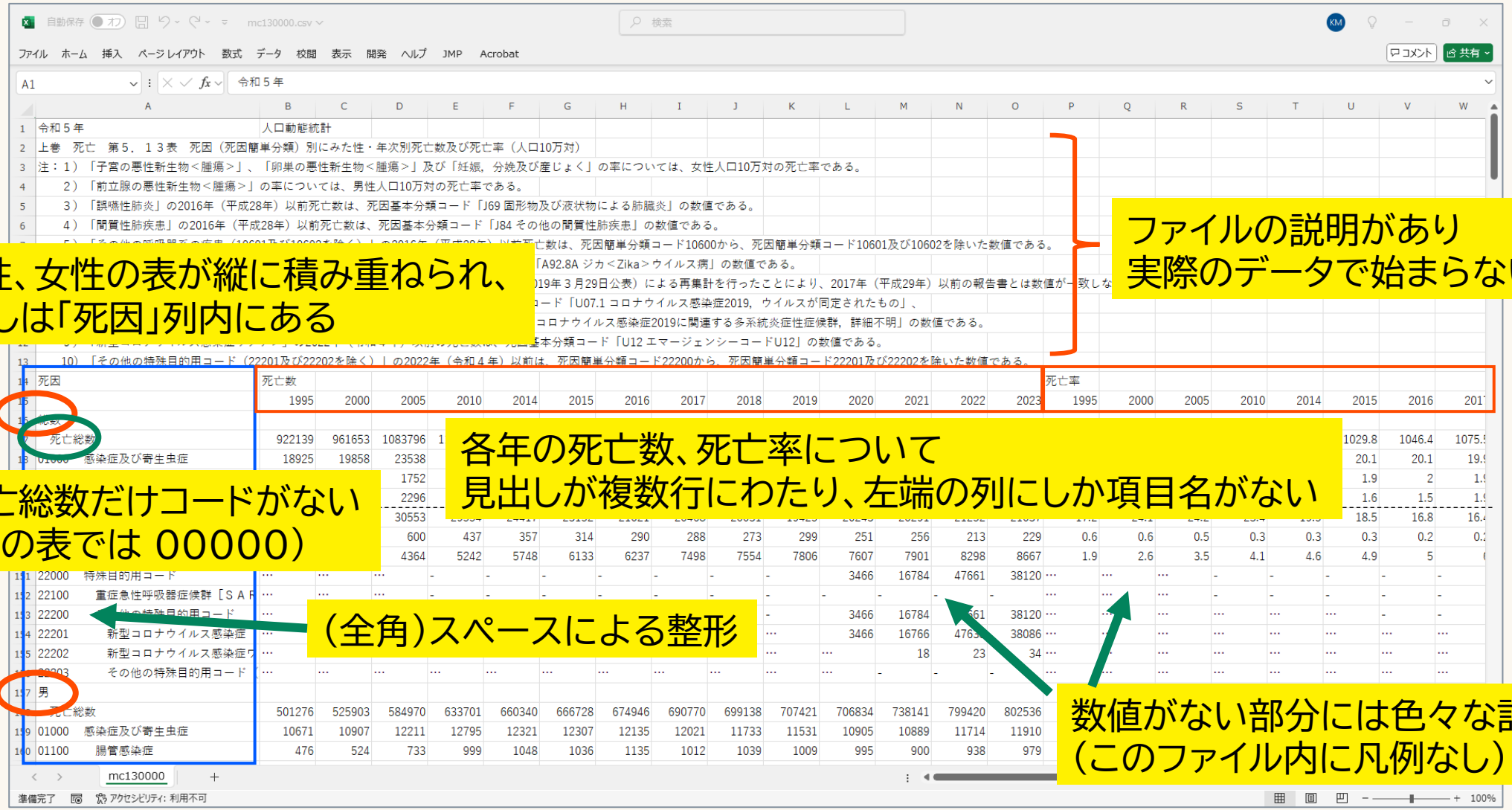


データの入手

- 厚労省 人口動態統計より、2023年の死因別死亡数のデータを使用
- 政府統計の総合窓口(e-Stat)の該当ページ
  https://www.e-stat.go.jp/stat-search/files?tclass=000001041646&cycle=7&year=20230
  より、**「5-13 死因(死因簡単分類)別にみた性・年次別死亡数及び死亡率(人口10万対)/ 2023年」** の CSVファイルを事前にご自身の Working Directory にダウンロードしておいて下さい
  - ダウンロードされるファイル名: mc130000.csv
  - 直接リンク: https://www.e-stat.go.jp/stat-search/file-download?statInfId=000040206118&fileKind=1

ファイル内容を確認

いきなり R で触る前に、まずは Excel などで開いてファイルの中身を俯瞰する



今回のデータ整形の方針

- 1ページの表に粗死亡数と人口10万人対の死亡率が一緒に掲載されているが、今回は「死亡数」の方を見ていくこととする
- 必要な処理としては以下のようなものが考えられる：
  1. CSVファイルを読み込むときに、先頭の解説行を読み飛ばす
  2. 列名（変数名）をRで扱いやすいように修正する
  3. 死因欄に性別の見出しも入っているので、性別と死因を分離する
  4. 死因欄の整形のためのスペースを取り除く
  5. 数値が入るべきセルにある記号を除去し、数値として扱えるように変換する
  6. （オプション）Tidy data に整形する

str_replace_all() 関数

stringr::str_replace_all(*string*, *pattern*, *replacement*)

- *string* 中に登場する *pattern* を *replacement* に置き換える
- 主に dplyr::mutate() の中で使用する

```
R 4.3.3 · ~/
> stringr::str_replace_all("ABCabc123", "abc", "***")
[1] "ABC***123"
> stringr::str_replace_all("ABCabc123", "[:lower:]", "*")
[1] "ABC***123"
> stringr::str_replace_all("ABCabc123", "\\d", "*")
[1] "ABCabc***"
> stringr::str_replace_all("ABCabc123", "([:lower:]+)", "_\\1_")
[1] "ABC_abc_123"
> stringr::str_replace_all("ABCabc123", "^[:graph:]", "*")
[1] "*BCabc123"
> stringr::str_replace_all("ABCabc123", "[:graph:]$", "*")
[1] "ABCabc12*"
>
```

任意の文字列を置き換える（基本形）

正規表現（ワイルドカード）

( ) で囲んだ pattern に合致した部分は replacement で引用できる

- \+ 1文字以上の繰り返し
- ^ 先頭
- $ 末尾

🔥 String manipulation with stringr  Cheat sheet を片手に試行錯誤 🔥

本日の内容

- Tidyverse とパイプ演算子のおさらい
- データ整形の指針

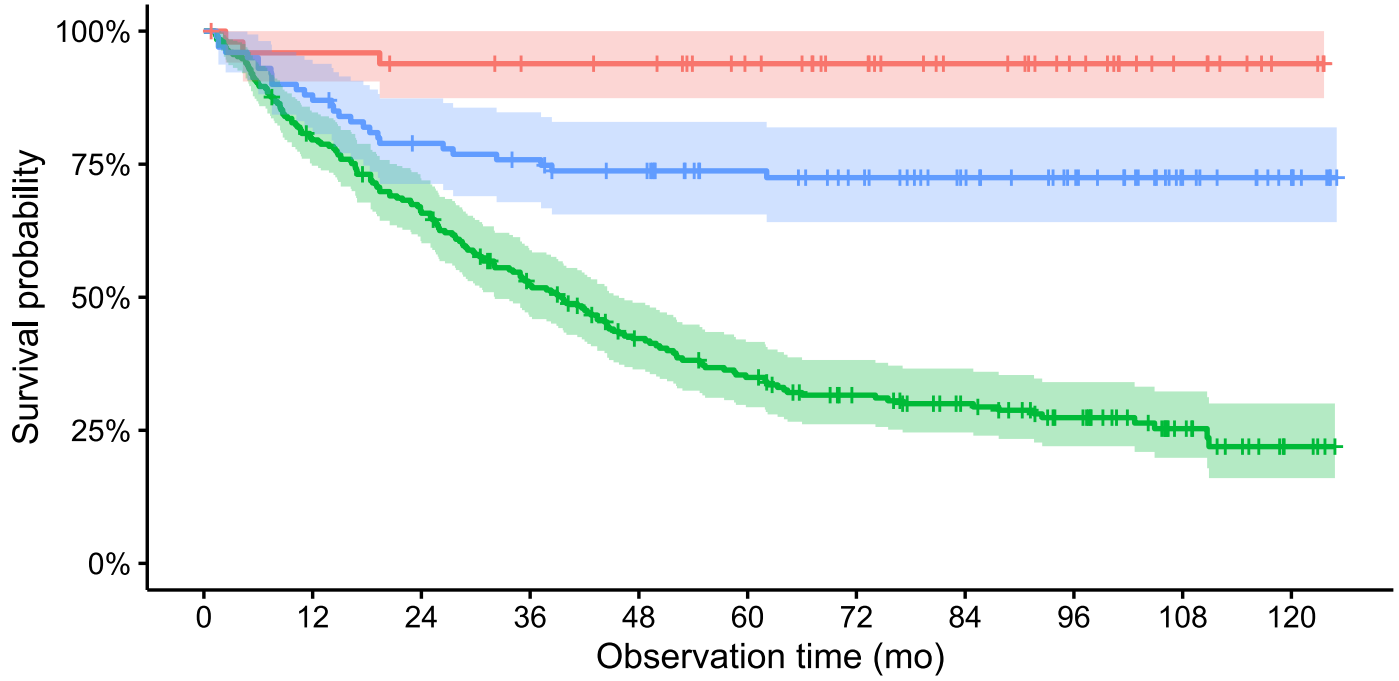
  ～本当はデータファイルを作る前に確認しておきたいこと～
- 初級編：最新の癌種別死亡数のグラフを作りたい

  ～人口動態統計のCSVファイルを整形してみる～
- **中級編：臨床研究のまとめファイルがやってきた**

  ～“ネ申エクセル”と戦うための基礎知識～

中級編のゴール

臨床研究データを解析に使用できる形に整形し、以下の図表を作成する



Characteristic	HR[1]	95% CI[1]	p-value[2]
dx			
COP	—	—	
IPF	23.7	3.29, 171	0.002 **
NSIP	10.4	1.37, 79.0	0.024 *
sex			
女	—	—	
男	1.53	1.07, 2.19	0.020 *
age_enroll	1.05	1.02, 1.08	0.002 **
delta_FVC_1y	0.17	0.10, 0.30	<0.001 ***

[1] HR = Hazard Ratio, CI = Confidence Interval

[2] *p<0.05; **p<0.01; ***p<0.001

データの入手

データが集まりました

From: 共同研究者 <eraihito@daigaku.ac.jp>

To: 自分, ボス ▼

お疲れ様です。症例登録担当の xxx です。
臨床研究のデータが集まりました。相談した解析をお願いします。
シート「症例登録票」が症例登録のFileMakerからエクセル出力したものです。
不足データは各施設に確認して分かる範囲で記入しました。1年後以降の結果はシート「アウトカム」で一緒に入れてあります。あと、登録時と経過のデータが分かりにくかったので見出しを追加してあります。
個人情報保護のため、患者氏名は削除してありますので安心して下さい。
パスワードはいつも通りです。それではよろしく。

症例登録票_3 250220 Final-2 生存ﾃ?ｰﾀつき Rev.1 解析用.xlsx

※フィクションです！！

承知しました……（とりあえずパスワード外して別名保存しよう）

ファイルを開く前に頭をよぎる可能性

データが集まりました
From: 共同研究者 <eraihito@daigaku.ac.jp>
To: 自分, ボス ▼

お疲れ様です。症例登録担当の xxx です。
臨床研究のデータが集まりました。相談した解析をお願いします。
シート「症例登録票」が症例登録のFileMakerからエクセル出力したものです。
不足データは各施設に確認して分かる範囲で記入しました。1年後以降の結果
はシート「アウトカム」で一緒に入れてあります。あと、登録時と
分かりにくかったので見出しを追加してあります。
個人情報保護のため、患者氏名は削除してありますので安心して下さい。
パスワードはいつも通りです。それではよろしく。

症例登録票_3 250220 Final-2 生存データ

チェックリスト ⇒ 1セルに複数データがありそう

複数行の見出しかセル結合がありそう

違う形式で入力されているかも

そのままではRで開けないな…

2つのデータの結合が必要だけど、連結に使える情報は残っているかな

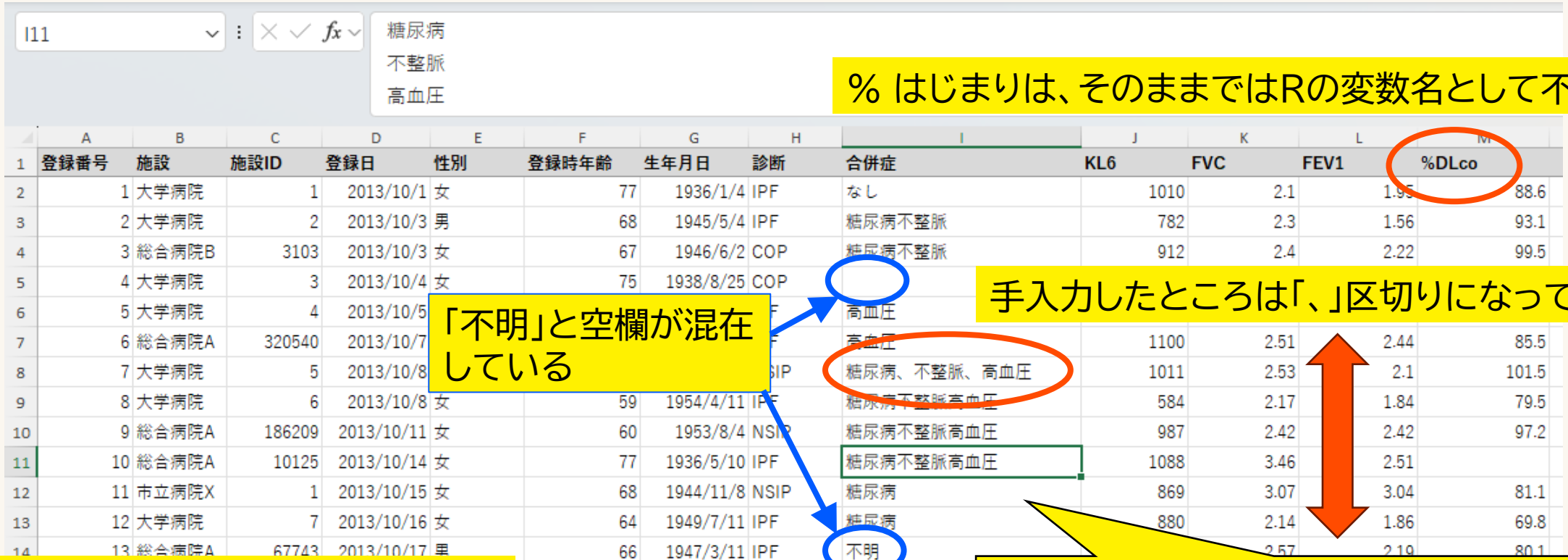
先方は mac かな。こっちは Windows だけど文字化けしないと良いな

あえずパスワード外して別名保存しよう）

※フィクションです！

ファイル内容を確認（1）

シート1「症例登録票」



I11

糖尿病
不整脈
高血圧

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	登録番号	施設	施設ID	登録日	性別	登録時年齢	生年月日	診断	合併症	KL6	FVC	FEV1	%DLco
2	1	大学病院	1	2013/10/1	女	77	1936/1/4	IPF	なし	1010	2.1	1.95	88.6
3	2	大学病院	2	2013/10/3	男	68	1945/5/4	IPF	糖尿病不整脈	782	2.3	1.56	93.1
4	3	総合病院B	3103	2013/10/3	女	67	1946/6/2	COP	糖尿病不整脈	912	2.4	2.22	99.5
5	4	大学病院	3	2013/10/4	女	75	1938/8/25	COP					
6	5	大学病院	4	2013/10/5					高血圧				
7	6	総合病院A	320540	2013/10/7						1100	2.51	2.44	85.5
8	7	大学病院	5	2013/10/8				SIP	糖尿病、不整脈、高血圧	1011	2.53	2.1	101.5
9	8	大学病院	6	2013/10/8	女	59	1954/4/11	IPF		584	2.17	1.84	79.5
10	9	総合病院A	186209	2013/10/11	女	60	1953/8/4	NSIP	糖尿病不整脈高血圧	987	2.42	2.42	97.2
11	10	総合病院A	10125	2013/10/14	女	77	1936/5/10	IPF	糖尿病不整脈高血圧	1088	3.46	2.51	
12	11	市立病院X	1	2013/10/15	女	68	1944/11/8	NSIP	糖尿病	869	3.07	3.04	81.1
13	12	大学病院	7	2013/10/16	女	64	1949/7/11	IPF	糖尿病	880	2.14	1.86	69.8
14	13	総合病院A	67743	2013/10/17	男	66	1947/3/11	IPF	不明		2.57	2.19	80.1
						68	1945/9/11	NSIP	糖尿病不整脈高血				
						63	1950/7/13	NSIP	糖尿病				
						71	1942/7/30	IPF					

% はじまりは、そのままではRの変数名として不適当

「不明」と空欄が混在している

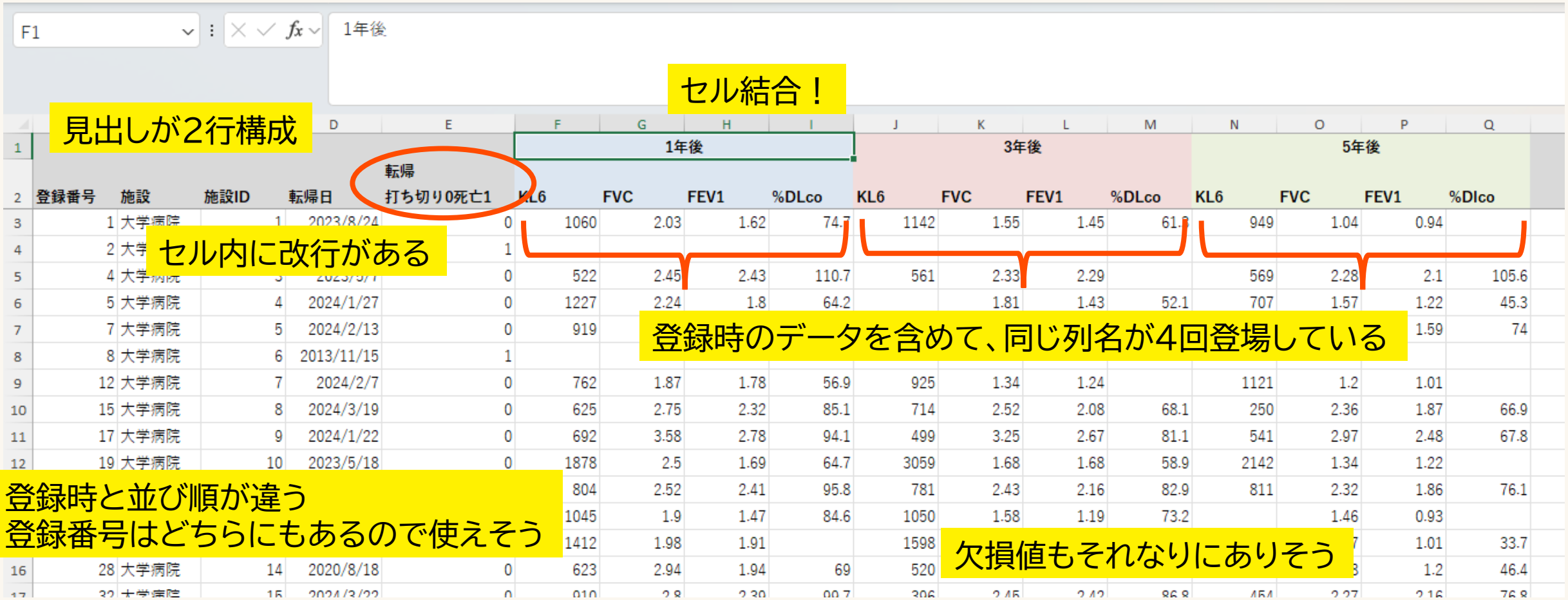
手入力したところは「、」区切りになっている

施設IDは形式がバラバラ
登録番号は両方にあれば使えそう

FileMaker で複数選択チェックリストだった項目は改行区切りで1セルに入っている

ファイル内容を確認（2）

シート2「アウトカム」



	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1			1年後				3年後				5年後			
2	登録番号	施設	施設ID	転帰日	転帰 打ち切り0死亡1	KL6	FVC	FEV1	%DLco	KL6	FVC	FEV1	%DLco	KL6	FVC	FEV1	%DIco
3	1	大学病院	1	2023/8/24	0	1060	2.03	1.62	74.7	1142	1.55	1.45	61.8	949	1.04	0.94	
4	2	大学			1												
5	4	大学病院	3	2023/5/7	0	522	2.45	2.43	110.7	561	2.33	2.29		569	2.28	2.1	105.6
6	5	大学病院	4	2024/1/27	0	1227	2.24	1.8	64.2		1.81	1.43	52.1	707	1.57	1.22	45.3
7	7	大学病院	5	2024/2/13	0	919										1.59	74
8	8	大学病院	6	2013/11/15	1												
9	12	大学病院	7	2024/2/7	0	762	1.87	1.78	56.9	925	1.34	1.24		1121	1.2	1.01	
10	15	大学病院	8	2024/3/19	0	625	2.75	2.32	85.1	714	2.52	2.08	68.1	250	2.36	1.87	66.9
11	17	大学病院	9	2024/1/22	0	692	3.58	2.78	94.1	499	3.25	2.67	81.1	541	2.97	2.48	67.8
12	19	大学病院	10	2023/5/18	0	1878	2.5	1.69	64.7	3059	1.68	1.68	58.9	2142	1.34	1.22	
13						804	2.52	2.41	95.8	781	2.43	2.16	82.9	811	2.32	1.86	76.1
14						1045	1.9	1.47	84.6	1050	1.58	1.19	73.2		1.46	0.93	
15						1412	1.98	1.91		1598						1.01	33.7
16	28	大学病院	14	2020/8/18	0	623	2.94	1.94	69	520						1.2	46.4
17	32	大学病院	15	2024/3/22	0	910	2.8	2.39	90.7	396	2.45	2.42	86.8	454	2.27	2.16	76.8

ファイル内容を確認（3）

シート2「アウトカム」

```
readxl::read_xlsx("ip_registry_data.xlsx", sheet = 2) %>% View()
```

セル結合の範囲の左上端に読み込まれ、他は空白扱い

Filter

	...1	...2	...3	...4	...5	1年後	...7	...8	...9	3年後	...11
1	登録番号	施設	施設ID	転帰日	転帰 打ち切り0死亡1	KL6	FVC	FEV1	%DLco	KL6	FVC
2	1	大学病院	1	45162	0	1060	2.0299999999999998	1.62	74.7	1142	1.55
3	2	大学病院	2	41647	1	NA	NA	NA	NA	NA	NA
4	4	大学病院	3	45053	0	522	2.4500000000000002	2.4300000000000002	110.7	561	2.33
5	5	大学病院	4	45318	0	1227	2.2400000000000002	1.8	64.2	NA	1.81
6	7	大学病院	5	45335	0	919	2.39	1.96	NA	1005	2.11
7	8	大学病院	6	41593	1	NA	NA	NA	NA	NA	NA
8	12	大学病院	7	45329	0	762	1.87	1.78	56.9	925	1.34
9	15	大学病院	8	45370	0	625	2.75	2.3199999999999998	85.1	714	2.52
10	17	大学病院	9	45313	0	692	3.58	2.78	94.1	499	3.25
11	19	大学病院	10	45064	0	1878				059	1.68
12	21	大学病院	11	43657	0	804				81	2.4300000000
13	22	大学病院	12	44129	0	1045				050	1.58
14	26	大学病院	13	44366	0	1412				598	2.04
15	28	大学病院	14	44061	0	623	2.94	1.94	69	520	2.7

2行目の見出しは
データの1行目扱い

RとExcelの数字の丸めの違いで
数値が変わっている
（臨床データとしてはほぼ誤差レベル）

日付データが整数値になってしまっている

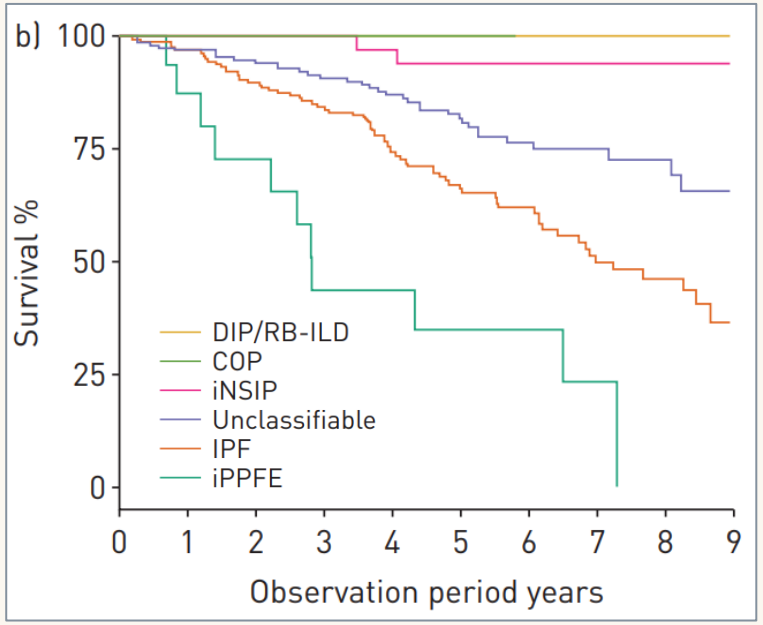
データ整形の方針

生存分析に使用することを考え、観察期間や各種変数を整理する：

- 列名はRで扱いやすい英単語ベースに改名する
  - 4回登場する KL6, FVC, FEV1, %DLco は時期を区別できるよう工夫する
- 2つのシートは「登録番号」をキーにして連結する
- 日付データは整数値（Excelのシリアル値）から日付に再変換する
- 観察期間がないので「転帰日 − 登録日」で求める
- 合併症は、「糖尿病」「不整脈」「高血圧」それぞれの有無に分解する
- 誤差が生じている数値があれば、小数点以下の桁数を揃えて丸める
- 入力間違いによる外れ値がないか確認

配布データについて（ドメイン知識）①

- 間質性肺炎の多施設レジストリ研究をイメージした模擬データ（n = 400）
- 観察期間は10年、診断（病型）は以下の3つに簡略化
  - IPF　（特発性肺線維症）
    - ✓ 特発性（他に原因がない）間質性肺炎の中で最多
    - ✓ 難治性で予後不良
    - ✓ 喫煙の影響あり、男性に多い
  - NSIP（非特異性間質性肺炎）
    - ✓ 治療反応性・予後などは雑多
    - ✓ 膠原病に関連する例も多く、女性に多い
  - COP（特発性器質化肺炎）
    - ✓ 治療反応性良く、予後良好



Fujisawa T, *et al.* Eur Respir J. 2019
（実際のレジストリ研究）より引用改変

- 努力肺活量の低下、高齢、男性などが間質性肺炎における既知の予後不良因子

配布データについて（ドメイン知識）②

これらの値まで取り扱う時間がないですが、toy data として残しておきます。

- 努力肺活量

  通常の呼吸から目一杯息を吸い込んだ後、力いっぱい一気に吐き出す検査

  - 最初の1秒間で吐き出した空気の量が「1秒量 $FEV_{1.0}$」
  - 最後まで吐き切った総量が「努力肺活量 FVC」
    - ✓ したがって必ず $FEV_{1.0}$ < FVC となり、それなりの相関関係がみられる
- 肺拡散能 DL_{CO}（$\%DL_{co}$ は年齢体格等からの予測値に対する値）
  - 測定手技の関係で、最低 1.5L（基本的には 2.0L以上）の肺活量が必要
    - ✓ それ以下では実際の拡散能によらず計測できない ⇒（FVC低値例では）MAR型の欠測が生じる

※もちろん病態の悪化により呼吸機能全体が実施不能となる場合もあるので、実際にはMNAR型の欠測も含まれます

- KL-6
  - 血液検査項目。代表的な肺組織障害（≒間質性肺炎の病勢）のバイオマーカー

まとめ

- データの前処理では、データクリーニングや変換操作の前に生データをコンピュータで認識できて解析に使用できる（機械判読可能な）表に変形・変換する過程が重要かつ時間がかかる部分
- Tidyverse の各関数を用いてデータを操作する際には、
  - 一息に操作を行わず、少しずつ結果を確認しながら進める
  - 作業のお供に
    - ✓ Posit社提供の各種チートシート
    - ✓ tidylog パッケージ
    - ✓ magritter の Tee pipe
    - ✓ AI アシスタント