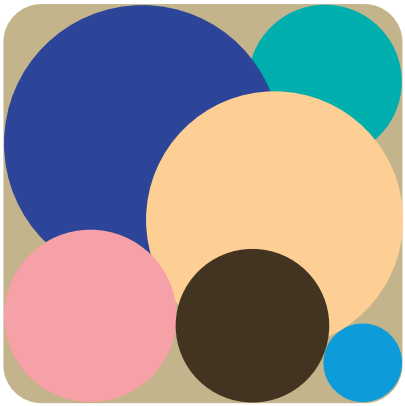
Real World Data

医療従事者のための リアルワールドデータの統計解析

はじめの一歩



奥田千恵子 著

Kinpodo

はじめに

本書を手にしていただいている方々の現在，あるいは，近い将来の職場である医療現場から日々生み出される膨大な量のデータは「日常診療データ」と呼ばれ，かつては「研究データ」とは別ものと考えられていました．臨床研究を志す医療従事者は多忙な日常臨床の合間に，「臨床家」から「研究者」へと頭を切り替えて研究プロトコールを作成し，研究費用を捻出し，研究テーマに沿って被験者を募り，同意を得た上で 1 例ずつコツコツとデータを収集しなければなりませんでした．

カルテの電子化，データベース化が進むにつれて，多くの医療機関で日常診療データが研究データとして二次利用されるようになってきました．現在では医学雑誌に投稿される論文のうち，電子カルテなどの既存のデータベースを利用した論文が 9 割近くを占めているとも言われています．このような流れの中で，リアルワールドデータ（RWD）という言葉が医療分野で頻繁に使われるようになりました．

RWD の定義は今でもそれほど明瞭ではありませんが，臨床試験などの実験的環境から得たデータに対して，カルテ由来の診療情報に加えて，診療報酬請求や疾患登録など，医療現場の情報をそのまま取得，整理したデータを総称することが多いと思われます．RWD を利用すれば，従来の臨床研究と比較して，データ収集のハードルを一気に下げることができます．規制当局が厳しい実施基準を設けている医薬品医療機器の開発においてさえも RWD を利活用しようという動きが出始めています．

RWD を利用する研究では，データ解析のハードルは逆に高くなります．例えば，ある疾患に対して治療法 A と治療法 B の有効性を比較する臨床試験（RCT）の場合，2 群に分けた被験者に，ランダムに治療法を割り付けてアウトカムを比較しますが，そのような方法が使えないデータベース研究では，患者の性別や年齢，併存疾患や重症度などの背景因子が，治療法の選択やアウトカムに影響（交絡）するため，いかにして交絡因子を調整し比較可能性を担保

するかが課題となります.

そのため，RWD の解析にはしばしば高度な統計モデルが用いられます．医学論文にはそのような統計モデルが日常的に登場するようになりましたが，現在のところ，初学者にも分かりやすく書かれた統計モデルの解説書は見当たりません．多くの医療従事者が臨床現場で疑問を持ち，エビデンスを求めて臨床研究を思い描き，その研究に必要な臨床データを身近に持ちながら，データ解析の壁に阻まれ，やむを得ず埋もれさせている現状があります.

本書では，統計モデルの解析法を習得する機会がない医療従事者を対象として，主に EZR のメニュー操作を用いてできるだけ分かりやすく解説し，さらに詳しい統計専門書や R プログラミング解説書へと橋渡しすることを目指しました．道筋が見えてさえいれば「千里の道も一歩から」です．まず，本書の数値例のデータファイルと EZR をダウンロードし，RWD 解析の「はじめの一歩」を踏み出しましょう.

執筆を終えるにあたり，金芳堂の名編集者，村上裕子氏に心より感謝申し上げます．私にとって初めての著書である「医薬研究者のためのケース別統計手法の学び方」（1999 年）の編集をしていただいたのが村上氏との出会いでした．それ以降 20 年に渡り，金芳堂より出版した全著書の編集を担当していただきました．定年退職されることとなり，本書が村上氏の最後の編集となったことにご縁の深さを感じます．編集の最終段階は一堂芳恵氏に引き継いでいただきました．改めて両編集者にお礼申し上げます.

2019 年秋

奥田千恵子

目　次

6 連続量データの解析

7 2値カテゴリデータの解析

8 さまざまな分布型のデータの解析

9 反復測定データの解析

1
リアルワールドデータの利活用

近年，多くの医療施設でカルテやレセプトの電子化が進み，検索機能によって必要なデータを抽出することができるようになったことにより，匿名化した上で研究データとして二次利用することが可能になりました．このような医療情報は特定の研究目的のために収集されたデータに対して，リアルワールドデータ（real world data, RWD）と呼ばれています．また，RWD から導き出されたエビデンスをリアルワールドエビデンス（real world evidence, RWE）と呼びます．

▶ RWD となる医療データベース

RWD とは，医療現場の情報をそのまま取得，整理したデータです．具体的には電子カルテ由来の診療情報や診療報酬請求（レセプト，receipt），疾患登録（レジストリ，registry）などを指します[注]．これらの RWD は以下のような特徴を持っています．

電子カルテ由来の診療情報

医療機関が保有する診療データであり，データ項目数が多く情報は豊富です

注：医療情報のデータベース等を用いた医薬品の安全性評価における薬剤疫学研究の実施に関するガイドライン，独立行政法人 医薬品医療機器総合機構，2014 （https://www.pmda.go.jp/files/000147250.pdf）

が，医療機関ごとにデータ項目やデータ形式が異なることがあるため，複数施設のデータを統合して研究に使用する場合にはデータ形式の標準化が大きな課題となります．現在，厚生労働省と独立行政法人医薬品医療機器総合機構（PMDA）により，複数の医療機関の診療情報データを標準化・統合して利用することの可能なネットワークの構築が進められています[注]．

診療報酬請求（レセプト）

医療機関や薬局が診療報酬または調剤報酬請求のために作成するものであり，医科，歯科，調剤および診断群分類（Diagnosis Procedure Combination, DPC）の 4 種類があります．研究に利用可能なデータセットとして提供，あるいは，そのデータを用いて解析業務を請け負う民間サービスもあります．また，厚生労働省が構築しているナショナルレセプトデータベース（National Database of Health Insurance Claims and Specific Health Checkups of Japan, NDB）は，国民の大半の医療保険診療の請求情報が集約された大規模なデータベースであり，発生が稀な副作用や疾患をアウトカムとする薬剤疫学研究などに利用されつつあります．

疾患登録（レジストリ）

医療従事者らにより自発的にがん登録や特定の手術に関連する情報の集積が行われています．個々の目的に沿って特定のデータ項目が収集されるため，ある研究のために作成したレジストリを二次利用する場合には，レジストリの作成目的と二次利用の研究目的が異なると，必要なデータ項目が含まれていないため追加の情報収集や他のデータとのリンケージの検討が必要となることがあります．

▶ RWD の研究デザイン

臨床研究の多くは因果関係（cause-and-effect relationship）を求める分析的

注：MID-NET は厚生労働省の医療情報データベース基盤整備事業によって構築された電子診療情報データベースとその解析システム．2018 年より協力医療機関 10 拠点に構築されたデータベースの PMDA による分析システムの運用が開始されている．

研究であり，実験的研究（experimental study）と観察的研究（observational study）に分けられます．RWD を利用した研究は後者に分類されます．

例えば，特定の治療法の効果を調べる場合，実験的研究では研究者が患者にどのような治療（原因）を行うかを制御して介入（intervention）できますからアウトカム（結果，outcome）との関係は明瞭です．一方，既存のデータベースを利用した研究では，患者が受けた治療は，研究を計画している研究者とは別の医療者によって行われており，研究者自身がその患者に介入したわけではありません．治療法は一般的な疫学研究における食事や喫煙などと同様に，疾患の快癒や増悪に影響する暴露因子（exposures）として扱われます．実験的研究のような制御ができない観察的研究では，原因が先，結果が後という時間的関係が判別できるデザインかどうかによってエビデンスレベルが異なります．

RWD を利用したデータベース研究の多くは，横断的（cross-sectional），または，縦断的（longitudinal）な後向きコホート研究（retrospective cohort study）です．この研究デザインで留意すべき点は適切な対照群が設定できるかどうかです．対照とは，通常は，同一の研究者によって同時に研究に組み込まれ観察される内部対照（internal control）を指します．適切な同時比較対照群を設定できない場合は，エビデンスレベルはやや低くなりますが，注目する

表　因果関係を求める臨床研究のデザインの分類　（表の下の方ほどエビデンスが強い）

A. 記述的研究（主として探索的研究として行われる，対照がない）
　　例：症例報告，症例集積研究，特定地域の健康調査など
B. 分析的研究（主として検証的研究として行われる，対照がある）
　1. 観察的研究（対象を制御せず，聞き取り調査や健康診断のみを行う）
　　a. 横断的研究（時間の要素がない）
　　　例：有病率や検査値の群間比較，相関関係など
　　b. 縦断的研究（時間の要素がある）
　　　1）後向き研究（スタート時点で「結果」が得られている）
　　　　例：後向きコホート研究，ケース・コントロール研究など
　　　2）前向き研究（スタート時点で「結果」が得られていない）
　　　　例：前向きコホート研究など
　2. 実験的研究（対象を制御し，薬剤の投与や処置などの介入を行う）
　　例：臨床試験など

要因を持つ群と年齢，性別，疾病，併用治療などが可能な限り似た集団や，過去のデータを歴史的対照（historical control）として比較することも可能です．

（参考文献 20）

▶ RWD の統計解析

因果関係を調べる研究において，原因と思われる因子（例，治療法）と，結果と思われる因子（例，治癒）の，両方と関わりのある因子を交絡因子（confounding factor）と呼びます．RWD を利用した研究では，何らかの方法で交絡因子の影響を除いておかないと，治療法と治癒との間に見せかけの因果関係が生じたり，逆に，あるはずの関係が検出できなくなったりしてしまいます．

回帰分析は交絡因子の影響を除く方法として広く用いられている統計手法です．データベース研究では，臨床検査値などの連続量データ（continuous data）をアウトカムとする場合には線形回帰分析，生／死や有効／無効などの 2 値カテゴリデータ（binary data）で表されるアウトカムには 2 項ロジスティック回帰分析を用いるのが定石となっています．

線形回帰分析や 2 項ロジスティック回帰分析では扱えないアウトカムもあります．また，RWD は同一患者から複数回測定され，欠測値が多く，測定時点が不ぞろいで，施設や治療者による偏り（クラスター化）があるなど，通常の解析手法では扱いにくいデータ構造をしています．そのため，RWD の統計解析には，一般化線形モデル（generalized linear model）や混合効果モデル（mixed effect model）などの高度なモデルが使われることもあります．

▶ 最近の統計ソフト事情

高度なモデルによる統計解析は今や医学論文には日常的に見られるようになってきましたが，SAS や SPSS などの汎用統計パッケージが必要になります．SAS は，製薬企業では米国 FDA（食品医薬品局）へ承認申請する際の事実上のスタンダードとなっている信頼度の高いソフトですが法人を対象とした高額なレンタル制をとっています．2014 年に SAS University Edition が無償で提供されるようになり，アカデミックな目的であれば個人のパソコンにダウン

ロードして有償の SAS に含まれる機能をほとんど利用することができるようになりました．SPSS は教育機関などでよく利用されていますが，高度な解析法にはいくつかのオプションシステムが必要なため，個人で所有するには高価なソフトです．どちらも職場にソフトがなかったり解説書が難解だったり，何となく敷居が高いと感じる医療従事者も少なくありません．

最近，医療分野でも利用者が増えてきている統計ソフト，R は国際共同研究プロジェクトで開発され，公開，配布されているオープンソースのフリーソフトウェアです．誰でも簡単に自分のパソコンにダウンロードすることができて，高価な汎用統計パッケージに匹敵する解析能力を持つソフトです．かつては製作者不詳の自作ソフトといった位置づけで，真偽の程はわかりませんが，R を使った論文を投稿するとプログラムの信頼性を保証するデータを要求されると言われていましたが，今や書店の棚に解説書があふれる人気ソフトとなっています．

しかし，R は自分でプログラミング（コンピュータに命令する用語を用いて記述）する必要があります．R プログラミングを系統的に習得するにはかなり時間がかかります．R の解説書の多くは，統計解析と R プログラミングの両方を同時にマスターしなければならないため，医療従事者にとって負担の大きいものになっています．

（参考文献 5）

▶ EZR

EZR（Easy R）という操作の簡易なメニュー型ソフトが 2012 年に神田善伸氏（自治医科大学附属さいたま医療センター血液科教授）により無償で提供されて以来，状況が一変しました．現在では EZR を使用した多くの論文が国際誌にも掲載されるに至っています．神田氏が著書に書いておられるように，「まずは簡単な EZR で統計に馴れてから，いずれ R のスクリプト入力（プログラミング）に挑戦する」ということも可能になります．

（参考文献 16）

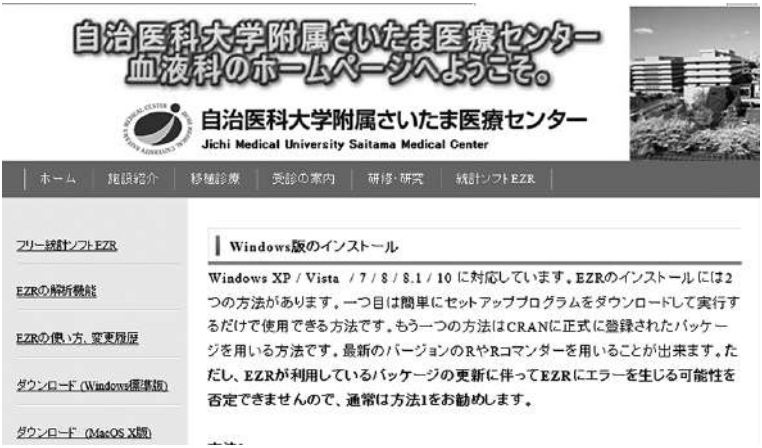
本書では，統計ソフトによる解析法を習得する機会がないまま，せっかくの貴重な臨床データを埋もれさせている医療従事者を対象として，主にEZRのメニュー操作を用いてできるだけ分かりやすく解説し，さらに詳しい統計専門書やRプログラミング解説書へと橋渡しすることを目指しました．

2
EZR のダウンロードとインストール

EZR は自治医科大学附属さいたま医療センター血液科ホームページ（http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/download.html）からダウンロードします．数分で終了する簡単な作業です．R 本体も同時にダウンロードされます．本書では Windows 版についてのみ説明しますが，ホームページには MacOS 版や LINUX 版も示されています．また，CRAN（R 本体の Web サイト）に正式に登録されたパッケージを用いる方法も示されています．

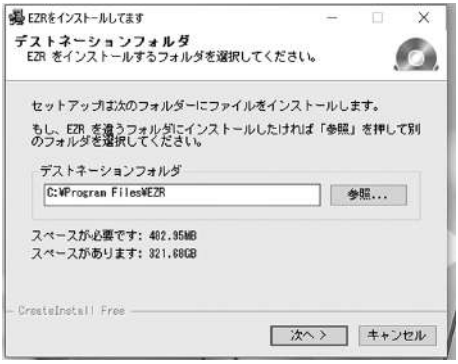
① ホームページのダウンロードをクリックする

セットアッププログラム（EZRsetup.exe）のダウンロードが始まる．Windows 版は WindowsXP/Vista/7/8/8.1/10 に対応している．



② セットアッププログラムが表示される

次へ＞をクリックすると，R および R コマンダー，その他の必要なパッケージを含めてすべてインストールされる（32 ビット版と 64 ビット版の両方の EZR がインストールされるが，32 ビット Windows では前者のみ使用可能．64 ビット Windows ではどちらを使ってもよい）．



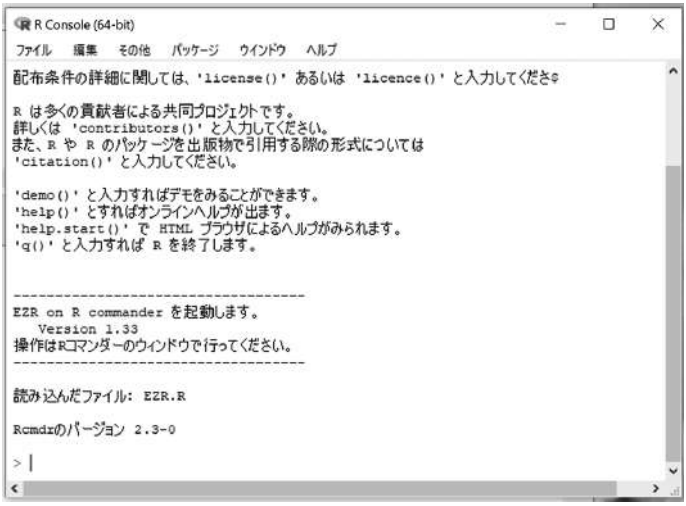
③　EZR のアイコンが生成する

Windows のスタートメニューに EZR（32-bit）と EZR（64-bit）のアイコンが作成されるので，使用する方のショートカットアイコンをデスクトップに作成しておく．

④　EZR のアイコンをクリックすると R コマンダーが表示される

R コマンダーには［R スクリプト］ウインドウ（上部）と，［出力］ウインドウ（下部）があり，最上部の［ファイル］，［編集］，［アクティブデータセット］などのメニュー，あるいは，その下のデータセットの 編集 ，表示 および 保存 により基本的な操作を行うことができる．

⑤　R Console も同時に表示されるが EZR の操作には使用しない．R を単独で使う時に利用（☞付録 1. R 本体の利用）．



（参考文献 16）

3
EZR のメニュー操作

EZR の R コマンダーのツールバーには 8 つのメニュー項目があります．本章では使用頻度の高い「ファイル」,「統計解析」,「グラフと表」および「標準メニュー」の使い方を説明します．

3.1 データセットの読み込みと保存

EZR はさまざまなファイル形式で保存したファイルを読み込むことができます．他のソフトで作成したファイルを，読み込み側で扱えるデータ形式に変換して読み込むことをインポート（import）と呼びます．本節では，excel 形式のデータファイルのインポート手順を説明します．

まず，本書の数値例に用いるデータセット data_x を excel で作成し，デスクトップに保存しておきます[注]．データセットの各変数の内容は後の章で説明します（☞ 5 データセットの解析計画）．

	A	B	C	D	E	F	G	H
1	ID	age	gender	treatment	conc	cure	freq	score
2	1	61	M	A	113	0	1	3
3	2	51	F	A	142	0	2	4
4	3	53	M	A	82	1	0	2
5	4	53	M	A	160	0	4	4
6	5	45	M	A	142	0	2	4
99	98	55	M	B	101	1	1	2
100	99	45	F	B	50	1	0	1
101	100	67	M	B	57	1	0	1

excel で作成したデータセット data_x

注：Excel 2007 以降のバージョンで作成すると，通常，Excel ブックというファイル形式（拡張子が .xlsx）で保存される．

A. excel 形式のデータファイルのインポート

① インポートするデータのファイル形式を選択する[注]

R コマンダー画面の最上部に並んだメニューの［ファイル］のプルダウンメニューから［データのインポート］→［Excel のデータをインポート］を選択.

② EZR で用いるファイル名をつける

データファイルをインポートする前に，EZR で用いるデータセット名を入力するダイアログボックスが表示される．適当な名前をつけて，OK をクリックする.

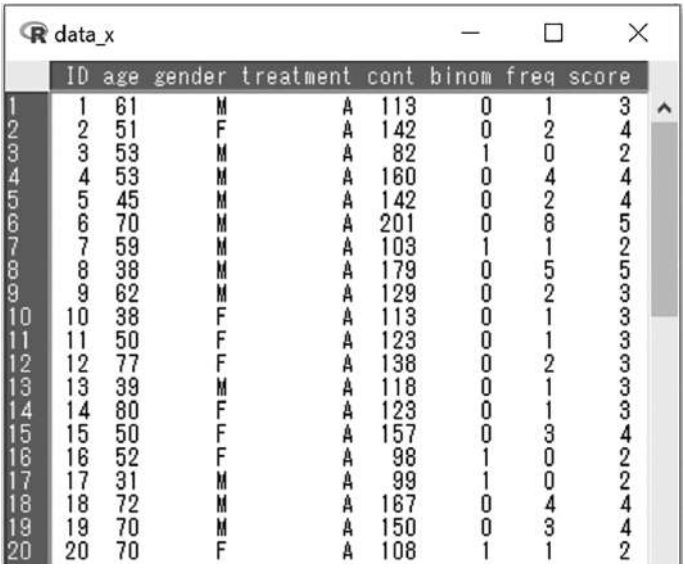
注：Excel，SPSS，Minitab，および Stata 以外のファイル形式（.txt や .csv など）のデータセットをインポートする時は，［ファイル］→［データのインポート］→［ファイルまたはクリップボード，URT からテキストデータを読み込む］を選択.

③　インポートを実行する

デスクトップに保存した data_x を選択し，開く をクリックすると EZR にインポートされる．

④　ファイルの内容を確認する

Rコマンダーの上部のデータセット：の 表示 をクリックすると data_x の内容が確認できる.



	ID	age	gender	treatment	cont	binom	freq	score
1	1	61	M	A	113	0	1	3
2	2	51	F	A	142	0	2	4
3	3	53	M	A	82	1	0	2
4	4	53	M	A	160	0	4	4
5	5	45	M	A	142	0	2	4
6	6	70	M	A	201	0	8	5
7	7	59	M	A	103	1	1	2
8	8	38	M	A	179	0	5	5
9	9	62	M	A	129	0	2	3
10	10	38	F	A	113	0	1	3
11	11	50	F	A	123	0	1	3
12	12	77	F	A	138	0	2	3
13	13	39	M	A	118	0	1	3
14	14	80	F	A	123	0	1	3
15	15	50	F	A	157	0	3	4
16	16	52	F	A	98	1	0	2
17	17	31	M	A	99	1	0	2
18	18	72	M	A	167	0	4	4
19	19	70	M	A	150	0	3	4
20	20	70	F	A	108	1	1	2

⑤　データセットを保存する

保存するフォルダを選び，保存 をクリックすると，Rのオリジナルのファイル形式（data_x.rda）として保存される.

[著者略歴]
奥田　千恵子　医学博士
1972 年　京都大学薬学部製薬化学科卒業
1986 年　京都府立医科大学麻酔学教室講師
1993 年　（財）ルイ・パストゥール医学研究センター基礎研究医療統計部門研究員
2011 年　横浜薬科大学教授
京都府立医科大学客員教授
2018 年　横浜薬科大学客員教授
[所属学会]
日本薬理学会，学術評議員
日本計算機統計学会
[著　書]
医薬研究者のためのケース別統計手法の学び方，金芳堂，京都，1999
医薬研究者のための統計ソフトの選び方（改 2），金芳堂，京都，2005
医薬研究者のための評価スケールの使い方と統計処理，金芳堂，京都，2007
医薬研究者のための研究デザインに合わせた統計手法の選び方，金芳堂，京都，2009
医薬研究者のための統計記述の英文表現（改 3），金芳堂，京都，2010
医薬研究者の視点からみた道具としての統計学（改 2），金芳堂，京都，2011
医療系はじめまして！統計学，金芳堂，京都，2015
親切な医療統計学（第 2 版），金芳堂，京都，2019
[訳　書]
たったこれだけ！医療統計学（改 2），金芳堂，京都，2015

医療従事者のためのリアルワールドデータの統計解析　はじめの一歩

2019年 12 月 10 日　第 1 版第 1 刷 ©

著　者　奥田千恵子　OKUDA, Chieko
発行者　宇山閑文
発行所　株式会社金芳堂
〒 606-8425 京都市左京区鹿ケ谷西寺ノ前町 34 番地
振替　01030-1-15605
電話　075-751-1111（代）
https://www.kinpodo-pub.co.jp/
印刷・製本　亜細亜印刷株式会社

落丁・乱丁本は直接小社へお送りください．お取替え致します．

Printed in Japan
ISBN978-4-7653-1802-0