

Contents

Part I Setting the Stage: Rationale Behind and Challenges to Health Data Analysis

1	Objectives of the Secondary Analysis of Electronic Health Record Data	3
1.1	Introduction	3
1.2	Current Research Climate	3
1.3	Power of the Electronic Health Record	4
1.4	Pitfalls and Challenges	5
1.5	Conclusion	6
	References	7
2	Review of Clinical Databases	9
2.1	Introduction	9
2.2	Background	9
2.3	The Medical Information Mart for Intensive Care (MIMIC) Database	10
2.3.1	Included Variables	11
2.3.2	Access and Interface	12
2.4	PCORnet	12
2.4.1	Included Variables	12
2.4.2	Access and Interface	13
2.5	Open NHS	13
2.5.1	Included Variables	13
2.5.2	Access and Interface	13
2.6	Other Ongoing Research	14
2.6.1	eICU—Philips	14
2.6.2	VistA	14
2.6.3	NSQUIP	15
	References	16

3	Challenges and Opportunities in Secondary Analyses of Electronic Health Record Data	17
3.1	Introduction	17
3.2	Challenges in Secondary Analysis of Electronic Health Records Data	17
3.3	Opportunities in Secondary Analysis of Electronic Health Records Data	20
3.4	Secondary EHR Analyses as Alternatives to Randomized Controlled Clinical Trials	21
3.5	Demonstrating the Power of Secondary EHR Analysis: Examples in Pharmacovigilance and Clinical Care	22
3.6	A New Paradigm for Supporting Evidence-Based Practice and Ethical Considerations	23
	References	25
4	Pulling It All Together: Envisioning a Data-Driven, Ideal Care System	27
4.1	Use Case Examples Based on Unavoidable Medical Heterogeneity	28
4.2	Clinical Workflow, Documentation, and Decisions	29
4.3	Levels of Precision and Personalization	32
4.4	Coordination, Communication, and Guidance Through the Clinical Labyrinth	35
4.5	Safety and Quality in an ICS	36
4.6	Conclusion	39
	References	41
5	The Story of MIMIC	43
5.1	The Vision	43
5.2	Data Acquisition	44
5.2.1	Clinical Data	44
5.2.2	Physiological Data	45
5.2.3	Death Data	46
5.3	Data Merger and Organization	46
5.4	Data Sharing	47
5.5	Updating	47
5.6	Support	48
5.7	Lessons Learned	48
5.8	Future Directions	49
	References	49
6	Integrating Non-clinical Data with EHRs	51
6.1	Introduction	51
6.2	Non-clinical Factors and Determinants of Health	51
6.3	Increasing Data Availability	53
6.4	Integration, Application and Calibration	54

6.5	A Well-Connected Empowerment.	57
6.6	Conclusion	58
	References.	59
7	Using EHR to Conduct Outcome and Health Services Research.	61
7.1	Introduction	61
7.2	The Rise of EHRs in Health Services Research	62
7.2.1	The EHR in Outcomes and Observational Studies.	62
7.2.2	The EHR as Tool to Facilitate Patient Enrollment in Prospective Trials	63
7.2.3	The EHR as Tool to Study and Improve Patient Outcomes.	64
7.3	How to Avoid Common Pitfalls When Using EHR to Do Health Services Research	64
7.3.1	Step 1: Recognize the Fallibility of the EHR.	65
7.3.2	Step 2: Understand Confounding, Bias, and Missing Data When Using the EHR for Research	65
7.4	Future Directions for the EHR and Health Services Research	67
7.4.1	Ensuring Adequate Patient Privacy Protection	67
7.5	Multidimensional Collaborations.	67
7.6	Conclusion	68
	References.	68
8	Residual Confounding Lurking in Big Data: A Source of Error	71
8.1	Introduction	71
8.2	Confounding Variables in Big Data	72
8.2.1	The Obesity Paradox	72
8.2.2	Selection Bias	73
8.2.3	Uncertain Pathophysiology	74
8.3	Conclusion	77
	References.	77
 Part II A Cookbook: From Research Question Formulation to Validation of Findings		
9	Formulating the Research Question.	81
9.1	Introduction	81
9.2	The Clinical Scenario: Impact of Indwelling Arterial Catheters.	82
9.3	Turning Clinical Questions into Research Questions.	82
9.3.1	Study Sample	82

9.3.2	Exposure	83
9.3.3	Outcome	84
9.4	Matching Study Design to the Research Question	85
9.5	Types of Observational Research	87
9.6	Choosing the Right Database	89
9.7	Putting It Together	90
	References.	91
10	Defining the Patient Cohort	93
10.1	Introduction	93
10.2	PART 1—Theoretical Concepts	94
10.2.1	Exposure and Outcome of Interest.	94
10.2.2	Comparison Group	95
10.2.3	Building the Study Cohort.	95
10.2.4	Hidden Exposures	97
10.2.5	Data Visualization	97
10.2.6	Study Cohort Fidelity	98
10.3	PART 2—Case Study: Cohort Selection.	98
	References.	100
11	Data Preparation	101
11.1	Introduction	101
11.2	Part 1—Theoretical Concepts	102
11.2.1	Categories of Hospital Data.	102
11.2.2	Context and Collaboration	103
11.2.3	Quantitative and Qualitative Data	104
11.2.4	Data Files and Databases.	104
11.2.5	Reproducibility	107
11.3	Part 2—Practical Examples of Data Preparation	109
11.3.1	MIMIC Tables.	109
11.3.2	SQL Basics	109
11.3.3	Joins	112
11.3.4	Ranking Across Rows Using a Window Function	113
11.3.5	Making Queries More Manageable Using WITH	113
	References.	114
12	Data Pre-processing.	115
12.1	Introduction	115
12.2	Part 1—Theoretical Concepts	116
12.2.1	Data Cleaning	116
12.2.2	Data Integration.	118
12.2.3	Data Transformation	119
12.2.4	Data Reduction	120

12.3	PART 2—Examples of Data Pre-processing in R	121
12.3.1	R—The Basics	121
12.3.2	Data Integration	129
12.3.3	Data Transformation	132
12.3.4	Data Reduction	136
12.4	Conclusion	140
	References	141
13	Missing Data	143
13.1	Introduction	143
13.2	Part 1—Theoretical Concepts	144
13.2.1	Types of Missingness	144
13.2.2	Proportion of Missing Data	146
13.2.3	Dealing with Missing Data	146
13.2.4	Choice of the Best Imputation Method	152
13.3	Part 2—Case Study	153
13.3.1	Proportion of Missing Data and Possible Reasons for Missingness	153
13.3.2	Univariate Missingness Analysis	154
13.3.3	Evaluating the Performance of Imputation Methods on Mortality Prediction	159
13.4	Conclusion	161
	References	161
14	Noise Versus Outliers	163
14.1	Introduction	163
14.2	Part 1—Theoretical Concepts	164
14.3	Statistical Methods	165
14.3.1	Tukey’s Method	166
14.3.2	Z-Score	166
14.3.3	Modified Z-Score	166
14.3.4	Interquartile Range with Log-Normal Distribution	167
14.3.5	Ordinary and Studentized Residuals	167
14.3.6	Cook’s Distance	167
14.3.7	Mahalanobis Distance	168
14.4	Proximity Based Models	168
14.4.1	k-Means	169
14.4.2	k-Medoids	169
14.4.3	Criteria for Outlier Detection	169
14.5	Supervised Outlier Detection	171
14.6	Outlier Analysis Using Expert Knowledge	171
14.7	Case Study: Identification of Outliers in the Indwelling Arterial Catheter (IAC) Study	171
14.8	Expert Knowledge Analysis	172

14.9	Univariate Analysis	172
14.10	Multivariable Analysis	177
14.11	Classification of Mortality in IAC and Non-IAC Patients	179
14.12	Conclusions and Summary	181
	Code Appendix	182
	References	183
15	Exploratory Data Analysis	185
15.1	Introduction	185
15.2	Part 1—Theoretical Concepts	186
15.2.1	Suggested EDA Techniques	186
15.2.2	Non-graphical EDA	187
15.2.3	Graphical EDA	191
15.3	Part 2—Case Study	199
15.3.1	Non-graphical EDA	199
15.3.2	Graphical EDA	200
15.4	Conclusion	202
	Code Appendix	202
	References	203
16	Data Analysis	205
16.1	Introduction to Data Analysis	205
16.1.1	Introduction	205
16.1.2	Identifying Data Types and Study Objectives	206
16.1.3	Case Study Data	209
16.2	Linear Regression	210
16.2.1	Section Goals	210
16.2.2	Introduction	210
16.2.3	Model Selection	213
16.2.4	Reporting and Interpreting Linear Regression	220
16.2.5	Caveats and Conclusions	223
16.3	Logistic Regression	224
16.3.1	Section Goals	224
16.3.2	Introduction	225
16.3.3	2×2 Tables	225
16.3.4	Introducing Logistic Regression	227
16.3.5	Hypothesis Testing and Model Selection	232
16.3.6	Confidence Intervals	233
16.3.7	Prediction	234
16.3.8	Presenting and Interpreting Logistic Regression Analysis	235
16.3.9	Caveats and Conclusions	236
16.4	Survival Analysis	237
16.4.1	Section Goals	237
16.4.2	Introduction	237

16.4.3	Kaplan-Meier Survival Curves	238
16.4.4	Cox Proportional Hazards Models	240
16.4.5	Caveats and Conclusions	243
16.5	Case Study and Summary	244
16.5.1	Section Goals	244
16.5.2	Introduction	244
16.5.3	Logistic Regression Analysis	250
16.5.4	Conclusion and Summary	259
	References	261
17	Sensitivity Analysis and Model Validation	263
17.1	Introduction	263
17.2	Part 1—Theoretical Concepts	264
17.2.1	Bias and Variance	264
17.2.2	Common Evaluation Tools	265
17.2.3	Sensitivity Analysis	265
17.2.4	Validation	266
17.3	Case Study: Examples of Validation and Sensitivity Analysis	267
17.3.1	Analysis 1: Varying the Inclusion Criteria of Time to Mechanical Ventilation	267
17.3.2	Analysis 2: Changing the Caliper Level for Propensity Matching	268
17.3.3	Analysis 3: Hosmer-Lemeshow Test	269
17.3.4	Implications for a ‘Failing’ Model	269
17.4	Conclusion	270
	Code Appendix	270
	References	271

Part III Case Studies Using MIMIC

18	Trend Analysis: Evolution of Tidal Volume Over Time for Patients Receiving Invasive Mechanical Ventilation	275
18.1	Introduction	275
18.2	Study Dataset	277
18.3	Study Pre-processing	277
18.4	Study Methods	277
18.5	Study Analysis	278
18.6	Study Conclusions	280
18.7	Next Steps	280
18.8	Connections	281
	Code Appendix	282
	References	282

19	Instrumental Variable Analysis of Electronic Health Records	285
19.1	Introduction	285
19.2	Methods	287
19.2.1	Dataset.	287
19.2.2	Methodology	287
19.2.3	Pre-processing	290
19.3	Results	291
19.4	Next Steps	292
19.5	Conclusions	293
	Code Appendix	293
	References.	293
20	Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project	295
20.1	Introduction	295
20.2	Dataset and Pre-preprocessing.	297
20.2.1	Data Collection and Patients Characteristics	297
20.2.2	Patient Inclusion and Measures	297
20.3	Methods	299
20.3.1	Prediction Algorithms	299
20.3.2	Performance Metrics	301
20.4	Analysis	302
20.4.1	Discrimination	302
20.4.2	Calibration.	303
20.4.3	Super Learner Library	305
20.4.4	Reclassification Tables.	305
20.5	Discussion.	308
20.6	What Are the Next Steps?	309
20.7	Conclusions	309
	Code Appendix	310
	References.	311
21	Mortality Prediction in the ICU	315
21.1	Introduction	315
21.2	Study Dataset	316
21.3	Pre-processing.	317
21.4	Methods	318
21.5	Analysis	319
21.6	Visualization	319
21.7	Conclusions	321
21.8	Next Steps	321
21.9	Connections	322
	Code Appendix	323
	References.	323

22	Data Fusion Techniques for Early Warning of Clinical Deterioration	325
22.1	Introduction	325
22.2	Study Dataset	326
22.3	Pre-processing	327
22.4	Methods	328
22.5	Analysis	330
22.6	Discussion	333
22.7	Conclusions	335
22.8	Further Work	335
22.9	Personalised Prediction of Deteriorations	336
	Code Appendix	337
	References	337
23	Comparative Effectiveness: Propensity Score Analysis	339
23.1	Incentives for Using Propensity Score Analysis	339
23.2	Concerns for Using Propensity Score	340
23.3	Different Approaches for Estimating Propensity Scores	340
23.4	Using Propensity Score to Adjust for Pre-treatment Conditions	341
23.5	Study Pre-processing	343
23.6	Study Analysis	346
23.7	Study Results	346
23.8	Conclusions	347
23.9	Next Steps	347
	Code Appendix	348
	References	348
24	Markov Models and Cost Effectiveness Analysis:	
	Applications in Medical Research	351
24.1	Introduction	351
24.2	Formalization of Common Markov Models	352
24.2.1	The Markov Chain	352
24.2.2	Exploring Markov Chains with Monte Carlo Simulations	353
24.2.3	Markov Decision Process and Hidden Markov Models	355
24.2.4	Medical Applications of Markov Models	356
24.3	Basics of Health Economics	356
24.3.1	The Goal of Health Economics: Maximizing Cost-Effectiveness	356
24.3.2	Definitions	357
24.4	Case Study: Monte Carlo Simulations of a Markov Chain for Daily Sedation Holds in Intensive Care, with Cost-Effectiveness Analysis	359

24.5	Model Validation and Sensitivity Analysis for Cost-Effectiveness Analysis.	364
24.6	Conclusion	365
24.7	Next Steps	366
	Code Appendix	366
	References.	366
25	Blood Pressure and the Risk of Acute Kidney Injury in the ICU: Case-Control Versus Case-Crossover Designs.	369
25.1	Introduction	369
25.2	Methods	370
25.2.1	Data Pre-processing	370
25.2.2	A Case-Control Study	370
25.2.3	A Case-Crossover Design	372
25.3	Discussion.	374
25.4	Conclusions	374
	Code Appendix	375
	References.	375
26	Waveform Analysis to Estimate Respiratory Rate	377
26.1	Introduction	377
26.2	Study Dataset	378
26.3	Pre-processing.	380
26.4	Methods	381
26.5	Results	384
26.6	Discussion.	385
26.7	Conclusions	386
26.8	Further Work	386
26.9	Non-contact Vital Sign Estimation	387
	Code Appendix	388
	References.	389
27	Signal Processing: False Alarm Reduction	391
27.1	Introduction	391
27.2	Study Dataset	393
27.3	Study Pre-processing.	394
27.4	Study Methods	395
27.5	Study Analysis	397
27.6	Study Visualizations	398
27.7	Study Conclusions	399
27.8	Next Steps/Potential Follow-Up Studies	400
	References.	401

28	Improving Patient Cohort Identification Using Natural Language Processing	405
28.1	Introduction	405
28.2	Methods	407
28.2.1	Study Dataset and Pre-processing	407
28.2.2	Structured Data Extraction from MIMIC-III Tables	408
28.2.3	Unstructured Data Extraction from Clinical Notes	409
28.2.4	Analysis	410
28.3	Results	410
28.4	Discussion.	413
28.5	Conclusions	414
	Code Appendix	414
	References.	415
29	Hyperparameter Selection.	419
29.1	Introduction	419
29.2	Study Dataset	420
29.3	Study Methods	420
29.4	Study Analysis	423
29.5	Study Visualizations	424
29.6	Study Conclusions	425
29.7	Discussion.	425
29.8	Conclusions	426
	References.	427
	Erratum to: Secondary Analysis of Electronic Health Records	E1

<http://www.springer.com/978-3-319-43740-8>

Secondary Analysis of Electronic Health Records

MIT Critical Data

2016, XXI, 427 p. 108 illus., 100 illus. in color.,

Hardcover

ISBN: 978-3-319-43740-8