

# Medical Data Privacy Handbook



Aris Gkoulalas-Divanis • Grigorios Loukides  
Editors

# Medical Data Privacy Handbook

*Editors*

Aris Gkoulalas-Divanis  
IBM Research - Ireland  
Mulhuddart  
Dublin, Ireland

Grigorios Loukides  
Cardiff University  
Cardiff, UK

ISBN 978-3-319-23632-2      ISBN 978-3-319-23633-9 (eBook)  
DOI 10.1007/978-3-319-23633-9

Library of Congress Control Number: 2015947266

Springer Cham Heidelberg New York Dordrecht London  
© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

*Dedicated to my parents, to Elena, and to the memory of my  
grandmother Sophia*  
*–Aris Gkoulalas-Divanis*

*Dedicated to the memory of my grandmother*  
*–Grigorios Loukides*



# Preface

The editors started working on medical data privacy in 2009, when they were postdoctoral researchers in the Health Information Privacy Laboratory, Department of Biomedical Informatics, Vanderbilt University. Their work on the topic involved understanding the privacy risks of medical data publishing and developing methods to prevent these risks. Protecting medical data privacy is a challenging problem, since a large volume of complex data must be protected in a setting that involves multiple parties (patients, physicians, carers, researchers, etc.). To address the problem, it is important to develop principled approaches that are specifically geared towards medical data. In addition, it is equally important to increase the awareness of all parties, involved in managing medical data, about privacy risks and approaches for achieving medical data privacy. Thus, the overarching aim of this book is to survey the field of medical data privacy and to present the state-of-the-art approaches to a wide audience.

The structure of the book closely follows the main categories of research works that have been undertaken to protect medical data privacy. Each such category is surveyed in a different part of the book, as follows. Part I is devoted to medical data *sharing*. Part II focuses on medical data privacy in *distributed and dynamic settings*. Following that, Part III examines privacy preservation in *emerging applications* featuring medical data, and Part IV discusses medical data privacy through *policy, data de-identification, and data governance*.

Privacy-preserving data sharing requires protecting the identity of patients and/or their sensitive information. For instance, attackers may use external data or background knowledge to learn patients' identity, even though attributes that directly identify patients (e.g., SSNs, phone numbers) have been removed. The problem has been studied extensively in the context of medical data, by the computer science, medical informatics, and statistics communities. However, there is no one-size-fits-all solution and various challenges remain. The purpose of Part I of this book is to survey the main research directions in the area of privacy-preserving medical data sharing and to present state-of-the-art approaches, including measures, algorithms, and software tools, that have been designed to solve this problem.

The protection of medical data privacy is particularly challenging, when multiple interrelated parties are involved. For example, medical data practitioners often need to link or exchange different parts of data about a patient, in the context of patient treatment. In addition, medical researchers or insurers may need to access patient information, according to the patient's privacy requirements. In this case, both the objectives of the parties accessing the data and the patient's requirements may change over time. Furthermore, data that are stored or processed in the cloud are vulnerable to a multitude of attacks, ranging from malicious access to intentional data modification. Part II of this book presents approaches focusing on privacy protection in such distributed and dynamic settings. These include approaches for linking data (record linkage), managing data access and patient consent, as well as exchanging health information. Furthermore, a comprehensive survey of privacy concerns and mitigation strategies for medical data in the cloud is presented.

Advances in medical devices and ubiquitous computing enable the collection and analysis of many complex data types, including genomic data, medical images, sensor data, biomedical signals, and health social network data. These data are valuable in a wide spectrum of emerging applications, either alone or in combination with data such as patient demographics and diagnosis codes, which are commonly found in Electronic Health Record (EHR) systems. For example, genomic studies have strong potential to lead to the discovery of effective, personalized drugs, and therapies. However, genomic data are extremely sensitive and must be privacy-protected. Part III of this book surveys privacy threats and solutions for all the aforementioned types of data that are central in emerging applications.

Parts I–III of this book focus on technical solutions that allow data owners (e.g., a healthcare institution) to effectively protect medical data privacy. On the other hand, Part IV focuses on the legal requirements for offering data privacy protection, as well as on the techniques and procedures that are required to satisfy this requirement. More specifically, this part examines key legal frameworks related to medical data privacy protection, as well as data de-identification and governance solutions, which are required to comply with these frameworks. A detailed presentation of the data protection legislation in the USA, EU, UK, and Canada is offered.

This book is primarily addressed to researchers and educators in the areas of computer science, statistics, and medical informatics who are interested in topics related to medical privacy. This book will also be a valuable resource to industry developers, as it explains the state-of-the-art algorithms for offering privacy. To ease understanding by nonexperts, the chapters contain a lot of background material, as well as many examples and citations to related literature. In addition, knowledge of medical informatics methods and terminology is not a prerequisite, and formalism was intentionally kept at a minimum. By discussing a wide range



of privacy techniques, providing in-depth coverage of the most important ones, and highlighting promising avenues for future research, this book also aims at attracting computer science and medical informatics students to this interesting field of research.

Dublin, Ireland  
Cardiff, UK  
July, 2015

Aris Gkoulalas-Divanis  
Grigorios Loukides



# Acknowledgements

We would like to thank all the authors, who have contributed chapters to this book, for their valuable contributions. This work would not have been possible without their efforts. A total of 63 authors who hold positions in leading academic institutions and industry, in Europe (France, Germany, Greece, Italy, Luxembourg, Switzerland, and UK), North America, Asia, Australia, and New Zealand, have contributed 29 chapters in this book, featuring more than 280 illustrations. We sincerely thank them for their hard work and the time they devoted to this effort.

In addition, we would like to express our deep gratitude to all the expert reviewers of the chapters for their constructive comments, which significantly helped towards improving the organization, readability, and overall quality of this handbook.

Last but not least, we are indebted to Susan Lagerstrom-Fife and Jennifer Malat from Springer, for their great guidance and advice in the preparation and completion of this handbook, as well as to the publication team at Springer for their valuable assistance in the editing process.



# Contents

<b>1</b>	<b>Introduction to Medical Data Privacy</b>	<b>1</b>
	Aris Gkoulalas-Divanis and Grigorios Loukides	
1.1	Introduction	1
1.1.1	Privacy in Data Sharing	2
1.1.2	Privacy in Distributed and Dynamic Settings	3
1.1.3	Privacy for Emerging Applications	3
1.1.4	Privacy Through Policy, Data De-identification, and Data Governance	4
1.2	Part I: Privacy in Data Sharing	5
1.3	Part II: Privacy in Distributed and Dynamic Settings	8
1.4	Part III: Privacy for Emerging Applications	9
1.5	Part IV: Privacy Through Policy, Data De-identification, and Data Governance	11
1.6	Conclusion	13
	References	13
 <b>Part I Privacy in Data Sharing</b>		
<b>2</b>	<b>A Survey of Anonymization Algorithms for Electronic Health Records</b>	<b>17</b>
	Aris Gkoulalas-Divanis and Grigorios Loukides	
2.1	Introduction	17
2.2	Privacy Threats and Models	19
2.2.1	Privacy Threats	19
2.2.2	Privacy Models	19
2.3	Anonymization Algorithms	21
2.3.1	Algorithms Against Identity Disclosure	21
2.4	Directions for Future Research	29
2.5	Conclusion	31
	References	31

<b>3</b>	<b>Differentially Private Histogram and Synthetic Data Publication ....</b>	<b>35</b>
	Haoran Li, Li Xiong, and Xiaoqian Jiang	
3.1	Introduction.....	35
3.2	Differential Privacy.....	36
3.2.1	Concept of Differential Privacy .....	36
3.2.2	Mechanisms of Achieving Differential Privacy .....	37
3.2.3	Composition Theorems.....	39
3.3	Relational Data .....	39
3.3.1	Problem Setting .....	39
3.3.2	Parametric Algorithms .....	42
3.3.3	Semi-parametric Algorithms .....	42
3.3.4	Non-parametric Algorithms .....	43
3.4	Transaction Data .....	48
3.4.1	Problem Setting .....	49
3.4.2	DiffPart.....	49
3.4.3	Private FIM Algorithms .....	50
3.4.4	PrivBasis .....	50
3.5	Stream Data.....	51
3.5.1	Problem Setting .....	51
3.5.2	Discrete Fourier Transform .....	52
3.5.3	FAST .....	52
3.5.4	w-Event Privacy .....	53
3.6	Challenges and Future Directions .....	54
3.6.1	Variety of Data Types .....	55
3.6.2	High Dimensionality .....	55
3.6.3	Correlated Constraints Among Attributes .....	55
3.6.4	Limitations of Differential Privacy .....	56
3.7	Conclusion.....	57
	References.....	57
<b>4</b>	<b>Evaluating the Utility of Differential Privacy: A Use Case</b>	
	<b>Study of a Behavioral Science Dataset .....</b>	<b>59</b>
	Raquel Hill	
4.1	Introduction.....	59
4.2	Background.....	62
4.2.1	Syntactic Models: $k$ -Anonymity .....	62
4.2.2	Differential Privacy: Definition .....	64
4.2.3	Applications.....	66
4.3	Methodology .....	67
4.3.1	Utility Measures .....	69
4.4	Results .....	70
4.4.1	Variable Distributions .....	71
4.4.2	Multivariate Logistic Regression .....	74
4.5	Discussion .....	79
4.6	Conclusion.....	80
	References.....	80

<b>5</b>	<b>SECRET: A Tool for Anonymizing Relational, Transaction and RT-Datasets</b>	<b>83</b>
	Giorgos Poulis, Aris Gkoulalas-Divanis, Grigorios Loukides, Spiros Skiadopoulos, and Christos Tryfonopoulos	
5.1	Introduction	84
5.2	Related Work	86
5.3	Overview of SECRET	87
5.3.1	Frontend of SECRET	87
5.3.2	Backend of SECRET	93
5.3.3	Components	98
5.4	Using SECRET	101
5.4.1	Preparing the Dataset	102
5.4.2	Using the Dataset Editor	103
5.4.3	The Hierarchy Editor	104
5.4.4	The Queries Workload Editor	104
5.4.5	Evaluating the Desired Method	105
5.4.6	Comparing Different Methods	106
5.5	Conclusion and Future Work	107
	References	108
<b>6</b>	<b>Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool</b>	<b>111</b>
	Fabian Prasser and Florian Kohlmayer	
6.1	Introduction	111
6.1.1	Background	112
6.1.2	Objectives and Outline	113
6.2	The ARX Data Anonymization Tool	114
6.2.1	Background	115
6.2.2	Overview	117
6.2.3	System Architecture	120
6.2.4	Application Programming Interface	123
6.2.5	Graphical User Interface	126
6.3	Implementation Details	133
6.3.1	Data Management	134
6.3.2	Pruning Strategies	136
6.3.3	Risk Analysis and Risk-Based Anonymization	138
6.4	Experimental Evaluation	139
6.5	Discussion	142
6.5.1	Comparison with Prior Work	142
6.5.2	Limitations and Future Work	144
6.5.3	Concluding Remarks	145
	References	145

<b>7</b>	<b>Utility-Constrained Electronic Health Record Data Publishing Through Generalization and Disassociation</b>	<b>149</b>
	Grigorios Loukides, John Liagouris, Aris Gkoulalas-Divanis, and Manolis Terrovitis	
7.1	Introduction	150
7.1.1	Identity Disclosure	150
7.1.2	Utility-Constrained Approach	152
7.1.3	Chapter Organization	154
7.2	Preliminaries	155
7.3	Generalization and Disassociation	156
7.4	Specification of Utility Constraints	159
7.4.1	Defining and Satisfying Utility Constraints	159
7.4.2	Types of Utility Constraints for ICD Codes	162
7.5	Utility-Constrained Anonymization Algorithms	163
7.5.1	Clustering-Based Anonymizer (CBA)	164
7.5.2	DISassociation Algorithm (DIS)	165
7.5.3	Comparing the CBA and DIS Algorithms	169
7.6	Future Directions	174
7.6.1	Different Forms of Utility Constraints	174
7.6.2	Different Approaches to Guaranteeing Data Utility	175
7.7	Conclusion	176
	References	176
<b>8</b>	<b>Methods to Mitigate Risk of Composition Attack in Independent Data Publications</b>	<b>179</b>
	Jiuyong Li, Sarowar A. Sattar, Muzammil M. Baig, Jixue Liu, Raymond Heatherly, Qiang Tang, and Bradley Malin	
8.1	Introduction	180
8.2	Composition Attack and Multiple Data Publications	181
8.2.1	Composition Attack	181
8.2.2	Multiple Coordinated Data Publications	183
8.2.3	Multiple Independent Data Publications	183
8.3	Risk Mitigation Through Randomization	185
8.4	Risk Mitigation Through Generalization	187
8.5	An Experimental Comparison	189
8.5.1	Data and Setting	190
8.5.2	Reduction of Risk of Composition Attacks	190
8.5.3	Comparison of Utility of the Two Methods	192
8.6	Risk Mitigation Through Mixed Publications	193
8.7	Conclusion	196
	References	198



<b>9</b>	<b>Statistical Disclosure Limitation for Health Data: A Statistical Agency Perspective .....</b>	<b>201</b>
	Natalie Shlomo	
9.1	Introduction .....	201
9.2	Statistical Disclosure Limitation for Microdata from Social Surveys .....	203
9.2.1	Disclosure Risk Assessment .....	204
9.2.2	Statistical Disclosure Limitation Methods .....	207
9.2.3	Information Loss Measures .....	211
9.3	Statistical Disclosure Limitation for Frequency Tables .....	213
9.3.1	Disclosure Risk in Whole Population Tabular Outputs ...	213
9.3.2	Disclosure Risk and Information Loss Measures Based on Information Theory .....	214
9.3.3	Statistical Disclosure Limitation Methods .....	217
9.4	Differential Privacy in Survey Sampling and Perturbation .....	219
9.5	Future Outlook for Releasing Statistical Data .....	222
9.5.1	Safe Data Enclaves and Remote Access .....	223
9.5.2	Web-Based Applications .....	224
9.5.3	Synthetic Data .....	226
9.6	Conclusion .....	228
	References .....	228

## **Part II Privacy in Distributed and Dynamic Settings**

<b>10</b>	<b>A Review of Privacy Preserving Mechanisms for Record Linkage ...</b>	<b>233</b>
	Luca Bonomi, Liyue Fan, and Li Xiong	
10.1	Introduction .....	233
10.2	Overview of Privacy Preserving Record Linkage .....	236
10.2.1	The PPRL Model .....	236
10.2.2	Taxonomy of Presented Techniques .....	238
10.3	Secure Transformations .....	244
10.3.1	Attribute Suppression and Generalization Methods .....	245
10.3.2	N-Grams Methods .....	246
10.3.3	Embedding Methods .....	248
10.3.4	Phonetic Encoding Methods .....	250
10.4	Secure Multi-Party Computation .....	251
10.4.1	Commutative Encryption Based Protocols .....	251
10.4.2	Homomorphic Encryption Based Protocols .....	252
10.4.3	Secure Scalar Product Protocols .....	254
10.5	Hybrid Approaches .....	256
10.5.1	Standard Blocking .....	257
10.5.2	Sorted Neighborhood Approach .....	258
10.5.3	Mapping .....	259
10.5.4	Clustering .....	259
10.6	Challenges and Future Research Directions .....	261

10.7	Conclusion .....	262
	References .....	262
<b>11</b>	<b>Application of Privacy-Preserving Techniques in Operational Record Linkage Centres .....</b>	<b>267</b>
	James H. Boyd, Sean M. Randall, and Anna M. Ferrante	
11.1	Introduction .....	267
11.1.1	Record Linkage Research Infrastructure .....	268
11.1.2	Privacy Challenges in Health Record Linkage .....	270
11.2	Data Governance .....	271
11.2.1	Legal Obligations .....	272
11.2.2	Information Governance .....	272
11.2.3	Separation of Data and Functions .....	273
11.2.4	Application and Approval Process .....	273
11.2.5	Information Security .....	274
11.3	Operational Models and Data Flows .....	274
11.3.1	Centralized Model .....	275
11.3.2	Separated Models .....	276
11.3.3	A Technique to Avoid Data Collusion .....	278
11.4	Privacy Preserving Methods .....	278
11.4.1	Privacy Preserving Models .....	279
11.4.2	Techniques for Privacy Preserving Linkage .....	279
11.4.3	Requirements of a Privacy Preserving Linkage Technique for Operational Linkage Centres .....	282
11.5	Conclusion .....	285
	References .....	285
<b>12</b>	<b>Privacy Considerations for Health Information Exchanges .....</b>	<b>289</b>
	Dalvin Hill, Joseph Walker, and John Hale	
12.1	Introduction .....	289
12.2	Health Information Exchanges .....	290
12.2.1	HIE Actors and Systems .....	290
12.2.2	HIE Models .....	293
12.2.3	HIPAA, HITECH and HIE Privacy Governance .....	294
12.3	Privacy Issues with HIEs .....	295
12.3.1	Patient Expectations and Concerns .....	296
12.3.2	Tension Between Functionality, Security and Privacy .....	297
12.3.3	Data Stewardship and Ownership .....	297
12.4	Principles and Practice of Privacy for HIEs .....	298
12.4.1	Guiding Principles .....	298
12.4.2	HIE Privacy in Practice .....	300
12.5	Emerging Issues .....	305
12.5.1	Big Data .....	305
12.5.2	m-Health and Telemedicine .....	306
12.5.3	Medical Devices .....	307

12.6	Conclusion .....	308
	References .....	308
<b>13</b>	<b>Managing Access Control in Collaborative Processes for Healthcare Applications .....</b>	<b>313</b>
	Xuan Hung Le and Dongwen Wang	
13.1	Introduction .....	314
13.2	Related Works .....	314
13.3	An Illustrative Example: New York State HIV Clinical Education Initiative .....	316
13.4	Development of the Enhanced RBAC Model .....	318
13.4.1	Overview of the Enhanced RBAC Model .....	319
13.4.2	Support Team Collaboration: Bridging Entities and Contributing Attributes .....	320
13.4.3	Extending Access Permissions to Include Workflow Contexts .....	322
13.4.4	Role-Based Access Delegation Targeting on Specific Objects: Providing Flexibility for Access Control in Collaborative Processes .....	322
13.4.5	Integration of Multiple Representation Elements for Definition of Universal Constraints .....	324
13.4.6	Case Studies to Encode Access Policies for CEI .....	326
13.5	System Framework for Implementation of Enhanced RBAC .....	329
13.5.1	System Architecture .....	330
13.5.2	Encoding of Access Policies .....	331
13.5.3	Interpretation of Access Control Policies .....	333
13.5.4	Application Layer .....	334
13.5.5	Demonstration Tool .....	334
13.6	Evaluation of the Enhanced RBAC Model .....	335
13.6.1	Selection of Study Cases .....	336
13.6.2	Access Permissions Computed with the Enhanced RBAC Model and the CEIAdmin System .....	339
13.6.3	Comparison Between the Enhanced RBAC Model and the CEIAdmin System .....	340
13.6.4	Development of the Gold-Standard .....	340
13.6.5	Measuring Effectiveness Based on Gold-Standard .....	342
13.6.6	Results .....	344
13.7	Discussion .....	345
13.7.1	Features of the Enhanced RBAC Model .....	345
13.7.2	System Framework for Implementation .....	349
13.7.3	Evaluation .....	350
13.7.4	Limitations .....	353
13.8	Conclusion .....	354
	References .....	355

<b>14 Automating Consent Management Lifecycle for Electronic Healthcare Systems</b> .....	361
Muhammad Rizwan Asghar and Giovanni Russello	
14.1 Introduction.....	361
14.2 Legal Background .....	363
14.2.1 Legal Framework for Consent .....	363
14.2.2 Consent in Healthcare Systems .....	365
14.2.3 Consent Limitations .....	366
14.3 A Case Study .....	368
14.4 Overview of Teleo-Reactive Policies .....	369
14.4.1 TR Policy Representation .....	369
14.4.2 TR Policy Evaluation.....	370
14.5 The ACTORS Approach .....	371
14.5.1 Authorisation Policies .....	373
14.5.2 Policy Templates.....	374
14.5.3 TR Policies .....	375
14.6 Managing Consent in Healthcare Scenarios .....	376
14.7 Related Work .....	382
14.8 Conclusion and Future Work .....	384
References.....	385
<b>15 e-Health Cloud: Privacy Concerns and Mitigation Strategies</b> .....	389
Assad Abbas and Samee U. Khan	
15.1 Introduction.....	389
15.2 An Overview of the e-Health Cloud.....	391
15.2.1 e-Health Cloud Benefits and Opportunities .....	391
15.2.2 Deployment Models for Cloud Based e-Health Systems .	393
15.2.3 Threats to Health Data Privacy in the Cloud .....	394
15.2.4 Essential Requirements for Privacy Protection.....	397
15.2.5 User/Patient Driven Privacy Protection Requirements....	399
15.2.6 Adversarial Models in the e-Health Cloud .....	399
15.3 Privacy Protection Strategies Employed in e-Health Cloud .....	400
15.3.1 Approaches to Protect Confidentiality in the e-Health Cloud.....	400
15.3.2 Approaches to Maintain Data Integrity in the e-Health Cloud.....	402
15.3.3 Approaches to Offer Collusion Resistance in the e-Health Cloud.....	406
15.3.4 Approaches to Maintain Anonymity in the e-Health Cloud.....	407
15.3.5 Approaches to Offer Authenticity in the e-Health Cloud.....	410
15.3.6 Approaches to Maintain Unlinkability in the e-Health Cloud.....	412
15.4 Discussion and Open Research Issues.....	416

15.5 Conclusion .....	417
References .....	418
 <b>Part III Privacy for Emerging Applications</b>	
<b>16 Preserving Genome Privacy in Research Studies .....</b>	<b>425</b>
Shuang Wang, Xiaoqian Jiang, Dov Fox, and Lucila Ohno-Machado	
16.1 Introduction .....	426
16.2 Policies, Legal Regulation and Ethical Principles of Genome Privacy .....	427
16.2.1 NIH Policies for Genomic Data Sharing .....	427
16.2.2 U.S. Legal Regulations for Genomic Data .....	430
16.2.3 Ethical Principles for Genome Privacy .....	432
16.2.4 Summary .....	433
16.3 Information Technology for Genome Privacy .....	433
16.3.1 Genome Privacy Risks .....	434
16.3.2 Genome Privacy Protection Technologies .....	434
16.3.3 Community Efforts on Genome Privacy Protection .....	436
16.4 Conclusion .....	437
References .....	438
 <b>17 Private Genome Data Dissemination .....</b>	<b>443</b>
Noman Mohammed, Shuang Wang, Rui Chen, and Xiaoqian Jiang	
17.1 Introduction .....	443
17.2 Literature Review .....	445
17.2.1 Privacy Attacks and Current Practices .....	445
17.2.2 Privacy Preserving Techniques .....	446
17.3 Problem Statement .....	447
17.3.1 Privacy Protection Model .....	448
17.3.2 Privacy Attack Model .....	448
17.3.3 Utility Criteria .....	449
17.4 Genomic Data Anonymization .....	449
17.4.1 Anonymization Algorithm .....	449
17.4.2 Privacy Analysis .....	453
17.4.3 Computational Complexity .....	453
17.5 Experimental Results .....	454
17.6 Conclusion .....	458
References .....	459
 <b>18 Threats and Solutions for Genomic Data Privacy .....</b>	<b>463</b>
Erman Ayday and Jean-Pierre Hubaux	
18.1 Threats for Genomic Privacy .....	463
18.1.1 Kin Genomic Privacy .....	465

18.2	Solutions for Genomic Privacy .....	470
18.2.1	Privacy-Preserving Management of Raw Genomic Data.....	470
18.2.2	Private Use of Genomic Data in Personalized Medicine .....	472
18.2.3	Private Use of Genomic Data in Research .....	477
18.2.4	Coping with Weak Passwords for the Protection of Genomic Data.....	481
18.2.5	Protecting Kin Genomic Privacy .....	484
18.3	Future Research Directions .....	487
18.4	Conclusion .....	490
	References .....	490
<b>19</b>	<b>Encryption and Watermarking for medical Image Protection .....</b>	<b>493</b>
	Dalel Bouslimi and Gouenou Coatrieux	
19.1	Introduction.....	493
19.2	Security Needs for Medical Data .....	495
19.2.1	General Framework.....	495
19.2.2	Refining Security Needs in an Applicative Context: Telemedicine Applications as Illustrative Example .....	497
19.3	Encryption Mechanisms: An <i>A Priori</i> Protection.....	498
19.3.1	Symmetric/Asymmetric Cryptosystems & DICOM .....	498
19.3.2	Block Cipher/Stream Cipher Algorithms.....	499
19.4	Watermarking: An <i>A Posteriori</i> Protection Mechanism .....	503
19.4.1	Principles, Properties and Applications .....	503
19.4.2	Watermarking Medical Images .....	506
19.5	Combining Encryption with Watermarking .....	512
19.5.1	Continuous Protection with Various Security Objectives: A State of the Art .....	512
19.5.2	A Joint Watermarking-Encryption (JWE) Approach .....	516
19.6	Conclusion .....	521
	References .....	521
<b>20</b>	<b>Privacy Considerations and Techniques for Neuroimages .....</b>	<b>527</b>
	Nakeisha Schimke and John Hale	
20.1	Introduction.....	527
20.2	Neuroimage Data .....	529
20.3	Privacy Risks with Medical Images .....	530
20.3.1	Neuroimage Privacy Threat Scenarios.....	530
20.3.2	Volume Rendering and Facial Recognition.....	532
20.3.3	Re-identification Using Structural MRI .....	534
20.4	Privacy Preservation Techniques for Medical Images .....	535
20.4.1	De-Identification Techniques .....	535
20.4.2	Privacy in Neuroimage Archives and Collaboration Initiatives .....	543

20.5	Conclusion .....	544
	References .....	544
<b>21</b>	<b>Data Privacy Issues with RFID in Healthcare .....</b>	<b>549</b>
	Peter J. Hawrylak and John Hale	
21.1	Introduction .....	549
21.1.1	RFID as a Technology .....	550
21.2	Dimensions of Privacy in Medicine .....	553
21.3	RFID in Medicine .....	556
21.3.1	Inventory Tracking .....	556
21.3.2	Tracking People .....	556
21.3.3	Device Management .....	557
21.4	Issues and Risks .....	558
21.5	Solutions .....	562
21.6	Conclusion .....	563
	References .....	564
<b>22</b>	<b>Privacy Preserving Classification of ECG Signals in Mobile e-Health Applications .....</b>	<b>569</b>
	Riccardo Lazzeretti and Mauro Barni	
22.1	Introduction .....	569
22.2	Plain Protocol .....	572
22.2.1	Classification Results .....	575
22.3	Cryptographic Primitives .....	575
22.3.1	Homomorphic Encryption .....	576
22.3.2	Oblivious Transfer .....	577
22.3.3	Garbled Circuits .....	578
22.3.4	Hybrid Protocols .....	579
22.4	Privacy Preserving Linear Branching Program .....	580
22.4.1	Linear Branching Programs (LBP) .....	580
22.4.2	ECG Classification Through LBP and Quadratic Discriminant Functions .....	584
22.4.3	ECG Classification Through LBP and Linear Discriminant Functions .....	586
22.4.4	Complexity Analysis .....	587
22.5	Privacy Preserving Classification by Using Neural Network .....	590
22.5.1	Neural Network Design .....	590
22.5.2	Quantized Neural Network Classifier .....	593
22.5.3	Privacy-Preserving GC-Based NN Classifier .....	595
22.5.4	Privacy-Preserving Hybrid NN Classifier .....	597
22.5.5	Comparison with the LBP Solution .....	598
22.6	Privacy Preserving Quality Evaluation .....	599
22.6.1	SNR Evaluation in the Encrypted Domain .....	599
22.6.2	SNR-Based Quality Evaluation .....	603
22.7	Conclusion .....	608
	References .....	609

<b>23</b>	<b>Strengthening Privacy in Healthcare Social Networks</b> .....	613
	Maria Bertsima, Iraklis Varlamis, and Panagiotis Rizomiliotis	
23.1	Introduction .....	613
23.2	Social Networks .....	615
23.2.1	On-line Social Networks .....	615
23.2.2	Healthcare Social Networks .....	616
23.3	Privacy .....	618
23.3.1	Background .....	618
23.3.2	Personal and Sensitive Data .....	619
23.3.3	Privacy Principles .....	621
23.3.4	Privacy Threats .....	622
23.4	Privacy Requirements for HSNs .....	627
23.4.1	Privacy as System Requirement .....	627
23.5	Enhancing Privacy in OSNs and HSNs .....	628
23.6	On-line Social Networks in the Healthcare Domain .....	631
23.6.1	Advice Seeking Networks .....	632
23.6.2	Patient Communities .....	632
23.6.3	Professional Networks .....	633
23.7	Conclusion .....	633
	References .....	634

#### **Part IV Privacy Through Policy, Data De-identification, and Data Governance**

<b>24</b>	<b>Privacy Law, Data Sharing Policies, and Medical Data: A Comparative Perspective</b> .....	639
	Edward S. Dove and Mark Phillips	
24.1	Introduction .....	639
24.2	Overview of Data Privacy Legal Frameworks .....	642
24.3	Data Privacy Laws and Guidelines .....	648
24.3.1	The OECD Privacy Guidelines .....	648
24.3.2	The Council of Europe Convention 108 .....	650
24.3.3	The European Union Data Protection Directive 95/46 .....	652
24.3.4	UK Data Protection Act 1998 .....	656
24.3.5	Canadian Privacy Legislation .....	658
24.3.6	The HIPAA Privacy Rule .....	659
24.4	Data Sharing Policies .....	664
24.4.1	US National Institutes of Health .....	665
24.4.2	Canadian Data Sharing Policies .....	666
24.4.3	Wellcome Trust (UK) .....	669
24.5	Towards Better Calibration of Biomedical Research, Health Service Delivery, and Privacy Protection .....	671
24.6	Conclusion .....	674
	References .....	674



<b>25</b>	<b>HIPAA and Human Error: The Role of Enhanced Situation Awareness in Protecting Health Information .....</b>	<b>679</b>
	Divakaran Liginlal	
25.1	Introduction.....	679
25.2	HIPAA, Privacy Breaches, and Related Costs .....	682
25.3	Situation Awareness and Privacy Protection .....	685
25.3.1	Definition of Situation Awareness .....	685
25.3.2	Linking Situation Awareness to Privacy Breaches .....	686
25.3.3	SA and HIPAA Privacy Breaches.....	688
25.4	Discussion and Conclusion .....	693
	References.....	695
<b>26</b>	<b>De-identification of Unstructured Clinical Data for Patient Privacy Protection .....</b>	<b>697</b>
	Stephane M. Meystre	
26.1	Introduction.....	697
26.2	Origins and Definition of Text De-identification .....	698
26.3	Methods Applied for Text De-identification .....	701
26.4	Clinical Text De-identification Application Examples .....	704
26.4.1	Physionet Deid.....	704
26.4.2	MIST (MITRE Identification Scrubber Toolkit) .....	705
26.4.3	VHA Best-of-Breed Clinical Text De-identification System .....	706
26.5	Why Not Anonymize Clinical Text? .....	708
26.6	U.S. Veterans Health Administration Clinical Text De-identification Efforts .....	709
26.7	Conclusion.....	713
	References.....	714
<b>27</b>	<b>Challenges in Synthesizing Surrogate PHI in Narrative EMRs .....</b>	<b>717</b>
	Amber Stubbs, Özlem Uzuner, Christopher Kotfila, Ira Goldstein, and Peter Szolovits	
27.1	Introduction.....	717
27.2	Related Work .....	719
27.3	PHI Categories .....	722
27.4	Data .....	724
27.5	Strategies and Difficulties in Surrogate PHI Generation .....	725
27.5.1	HIPAA Category 1: Names .....	726
27.5.2	HIPAA Category 2: Locations .....	728
27.5.3	HIPAA Category 3: Dates and Ages.....	729
27.5.4	HIPAA Category 18: Other Potential Identifiers .....	731
27.6	Errors Introduced by Surrogate PHI.....	732
27.7	Relationship Between De-identification and Surrogate Generation .....	732
27.8	Conclusion.....	733
	References.....	734

<b>28</b>	<b>Building on Principles: The Case for Comprehensive, Proportionate Governance of Data Access.....</b>	<b>737</b>
	Kimberlyn M. McGrail, Kaitlyn Gutteridge, and Nancy L. Meagher	
28.1	Introduction.....	737
28.2	Current Approaches to Data Access Governance.....	739
28.2.1	Existing Norms for Data Access Governance.....	739
28.2.2	The Preeminence of “Consent or Anonymize” as Approaches to Data Access Governance.....	740
28.2.3	Existing Data Access Governance in Practice.....	743
28.3	The Evolution of Data and Implications for Data Access Governance.....	744
28.3.1	Big Data.....	744
28.3.2	Open Data.....	745
28.3.3	The Ubiquity of Collection of Personal Information.....	745
28.3.4	The Limits of Existing Approaches to Data Access Governance.....	746
28.4	A Comprehensive Model for Governance: Proportionate and Principled.....	747
28.4.1	Proportionality.....	747
28.4.2	Principle-Based Regulation.....	748
28.4.3	Case Studies Using Proportionate and Principled Access.....	749
28.5	Building on the Present: A Flexible, Governance Framework.....	752
28.5.1	Science.....	754
28.5.2	Approach.....	754
28.5.3	Data.....	755
28.5.4	People.....	755
28.5.5	Environment.....	755
28.5.6	Interest.....	756
28.5.7	Translating Risk Assessment to Review Requirements...	756
28.5.8	Adjudication Scenarios.....	757
28.6	Conclusion.....	759
	References.....	760
<b>29</b>	<b>Epilogue.....</b>	<b>765</b>
	Aris Gkoulalas-Divanis and Grigorios Loukides	
29.1	Introduction.....	765
29.2	Topics and Directions in Privacy Preserving Data Sharing.....	766
29.3	Topics and Directions in Privacy Preservation for Distributed and Dynamic Settings.....	768
29.4	Topics and Directions in Privacy Preservation for Emerging Applications.....	769
29.5	Topics and Directions in Privacy Preservation Through Policy, Data De-identification, and Data Governance.....	771

29.6 Conclusion ..... 772

References ..... 772

**About the Authors** ..... 775

**Glossary** ..... 815

**Index** ..... 827



# List of Figures

Fig. 3.1	Example: released cell histogram ( <i>left</i> ) and subcube histogram ( <i>right</i> ), and $N_i$ is a random Laplace noise (see Sect. 3.2 for Laplace mechanism) .....	40
Fig. 3.2	Generate synthetic data via parametric methods .....	41
Fig. 3.3	Generate synthetic data via non-parametric methods .....	41
Fig. 3.4	Generate synthetic data via semi-parametric methods .....	41
Fig. 3.5	DExample of private quadtree: noisy counts ( <i>inside boxes</i> ) are released; actual counts, although depicted, are not released. Query Q ( <i>dotted red rectangle</i> ) could be answered by adding noisy counts of marked nodes (Color figure online) [6] .....	45
Fig. 3.6	Taxonomy tree of attributes [29] .....	48
Fig. 3.7	Tree for partitioning records [29] .....	48
Fig. 3.8	A context-free taxonomy tree of the sample data in Table 3.1 [5] .....	49
Fig. 3.9	The partitioning process of Fig. 3.1 [5] .....	50
Fig. 3.10	The FAST framework [16] .....	53
Fig. 4.1	Excerpt from doctor's notes .....	60
Fig. 4.2	Experiment flow chart .....	67
Fig. 4.3	Histogram of ages from original data ( <i>left</i> ) and using k-d tree algorithm with $\epsilon = 2.0$ , ET = 0.677 ( <i>right</i> ) .....	72
Fig. 4.4	Histogram of genders from original data ( <i>left</i> ) and using cell-based algorithm with $\epsilon = 2.0$ ( <i>right</i> ) .....	72
Fig. 4.5	Proportion of variable counts vs. $\epsilon$ for all algorithms (for the first reduced dataset) .....	73
Fig. 4.6	Proportion of variable counts vs. $\epsilon$ for all algorithms (for the second reduced dataset) .....	73
Fig. 4.7	Proportion of variable counts preserved vs. $\epsilon$ for k-d tree (for MART_rs1) .....	74

Fig. 4.8	Proportion of variable counts preserved vs. $\epsilon$ for k-d tree (for MART_rs2) .....	75
Fig. 4.9	Effect size versus $\epsilon$ for RS1 cell-based runs that were similar ....	75
Fig. 4.10	Logistic results for MART_final for k-d tree, effect size and proportion of good runs versus the DP $\epsilon$ parameter.....	76
Fig. 4.11	Logistic results for MART_rs1 for k-d tree, proportion of good runs versus the DP $\epsilon$ parameter.....	77
Fig. 4.12	Logistic results for MART_rs1 for k-d tree, effect size and proportion of good runs versus the DP $\epsilon$ parameter for entropy_threshold = 1.0 .....	77
Fig. 4.13	Logistic results for MART_rs2, proportion of good runs versus the DP $\epsilon$ parameter .....	78
Fig. 4.14	Logistic results for MART_rs2 for k-d tree, effect size and proportion of good runs versus the DP $\epsilon$ parameter for entropy_threshold = 1.0 .....	78
Fig. 5.1	System architecture of SECRETETA .....	88
Fig. 5.2	The main screen of SECRETETA .....	89
Fig. 5.3	Automatic creation of hierarchies. (a) Selecting the number of splits per level of the hierarchy and (b) Displaying the produced hierarchy .....	89
Fig. 5.4	The evaluation mode: method evaluation screen of SECRETETA ...	90
Fig. 5.5	The comparison mode: methods comparison screen of SECRETETA .....	91
Fig. 5.6	The experimentation interface selector .....	91
Fig. 5.7	Plots for (a) the original dataset, (b) varying parameters execution, and (c) the comparison mode .....	92
Fig. 5.8	An example of a hierarchy tree .....	94
Fig. 5.9	The dataset editor .....	103
Fig. 5.10	Frequency plots of the original dataset .....	103
Fig. 5.11	The hierarchy specification area .....	104
Fig. 5.12	Method parameters setup .....	105
Fig. 5.13	A messagebox with the results summary .....	106
Fig. 5.14	The data output area .....	106
Fig. 5.15	The plotting area.....	106
Fig. 5.16	The configurations editor .....	107
Fig. 6.1	Example cancer dataset: types of attributes and types of disclosure .....	115
Fig. 6.2	Generalization hierarchies for attributes age and gender .....	116
Fig. 6.3	Example search space .....	117
Fig. 6.4	High-level architecture of the ARX system .....	120
Fig. 6.5	Overview of the most important classes in ARX's core .....	121
Fig. 6.6	Overview of the most important classes in ARX's application programming interface .....	122
Fig. 6.7	Anonymization process implemented in ARX's GUI .....	127

Fig. 6.8	The ARX configuration perspective .....	128
Fig. 6.9	Wizard for creating a generalization hierarchy with intervals.....	129
Fig. 6.10	The ARX exploration perspective .....	130
Fig. 6.11	The ARX utility evaluation perspective .....	132
Fig. 6.12	The ARX risk analysis perspective .....	133
Fig. 6.13	Example of how data is encoded and transformed in ARX .....	134
Fig. 6.14	Example of how data snapshots are represented in ARX.....	135
Fig. 8.1	The average accuracy of the composition attack on the <i>Salary</i> and <i>Occupation</i> datasets .....	191
Fig. 8.2	The average query errors of the <i>Salary</i> and <i>Occupation</i> datasets with different methods .....	192
Fig. 8.3	Distance between the original dataset, the output of dLink, and several privacy budgets of differential privacy ( $\epsilon = 0.01, 0.05, 0.1$ ) .....	193
Fig. 8.4	An illustration of the mixed publication model .....	194
Fig. 8.5	An example of the mixed publication .....	195
Fig. 9.1	Confidential residual plot from a regression analysis on receipts for the Sugar Canes dataset. (a) Residuals by fitted values. (b) Normal QQ plot of residuals.....	226
Fig. 9.2	Univariate analysis of receipts for the Sugar Canes dataset .....	227
Fig. 10.1	The privacy-preserving record linkage (PPRL) model .....	237
Fig. 10.2	Bloom Filter representation for the names SMITH and SMYTH using 2-g. The map is obtained using one hash function and in total there are ten bits in A, and 11 bits in B set to 1. Only eight bits are shared among the Bloom filters, therefore the similarity measure between the original strings approximated with the Dice coefficient is $\frac{2 \cdot 8}{(10+11)} \approx 0.762$ (example from Schnell et al. [45]) .....	246
Fig. 10.3	Example of composite Bloom filter representation from Durham et al. [13]. (a) Transformation process. (b) Composite bloom filter .....	247
Fig. 10.4	Embedding example for the names SMITH and SMYT with an embedding base formed by the sets $S_1, S_2, S_3, S_4$ of randomly generated strings .....	249
Fig. 10.5	Example of bitwise encryption by Kuzu et al. [33] .....	253
Fig. 10.6	Performing blocking on datasets: (a) Original datasets $T$ and $V$ ; (b) Block decomposition using hyper-rectangles for $T$ ; (c) Block perturbation for $T$ ; (d) Block perturbation for $V$ . The candidate matching pairs tested in the SMC part are limited to the pair of overlapping blocks: $(T_1, V_1), (T_1, V_2), \dots, (T_5, V_5)$ (example from Inan et al. [24]) .....	258

Fig. 10.7	Private blocking via clustering (example from [28]) .....	260
Fig. 11.1	A centralized model: data providers give full datasets to the linkage unit, who link and then pass on the data to the researcher .....	275
Fig. 11.2	The data provider splits the data, sending the personal identifiers to the linkage unit and the clinical content to the client services team. The linkage unit then provides the linkage map to the client services team who join it to content data to create datasets for research and analysis .....	277
Fig. 11.3	In the absence of a repository of clinical data, this is supplied to the researcher by the data provider .....	278
Fig. 11.4	Basic data flow of three party privacy preserving linkage .....	280
Fig. 11.5	Numerous protocols attempt to reduce the variability between records of the same person, while maintaining variability between records belonging to different people .....	281
Fig. 11.6	Creating a statistical linkage key .....	281
Fig. 11.7	First and last name are phonetically encoded and concatenated with date of birth and sex, which is then hashed to form the Swiss Anonymous Linkage Code .....	282
Fig. 12.1	HIE actors and systems .....	292
Fig. 12.2	HIE models: (a) centralized, (b) decentralized, and (c) hybrid....	293
Fig. 12.3	HIPAA covered entities and business associates.....	295
Fig. 13.1	Workflow of a CEI training session (reprinted from [51], with permission from Elsevier) .....	318
Fig. 13.2	Enhanced RBAC model with universal constraints, workflow in permissions, and domain ontologies (reprinted from [51], with permission from Elsevier) .....	320
Fig. 13.3	Collaboration model with bridging entity and contributing attributes (reprinted from [51], with permission from Elsevier).....	321
Fig. 13.4	System architecture (reprinted from [49]).....	331
Fig. 13.5	Three-level access control policy encoding in Protégé (reprinted from [49]) .....	332
Fig. 13.6	An example of access policy for CEI (reprinted from [49]); (a) Access policy in first-order predicate logic, (b) Access policy in Protege SWRL .....	333
Fig. 13.7	A screenshot of the demo tool showing CEI access management (reprinted from [49]) .....	335
Fig. 13.8	Overall design of the evaluation study (reprinted from [52], with permission from Elsevier) .....	336



Fig. 13.9	Mappings of CEI Centers, system roles, and users (reprinted from [52], with permission from Elsevier) .....	337
Fig. 13.10	A screenshot of the online tool used by the judges to build the gold-standard (reprinted from [52], with permission from Elsevier).....	343
Fig. 13.11	Mapping the enhanced RBAC framework to XACML (reprinted from [49]) .....	350
Fig. 14.1	A layout of TR policies .....	369
Fig. 14.2	An example of a TR policy .....	370
Fig. 14.3	The ACTORS architecture for managing consent lifecycle .....	372
Fig. 14.4	An example of an authorisation policy .....	374
Fig. 14.5	An example of a policy template .....	375
Fig. 14.6	A TR policy for managing authorisation policy for providing consent to a GP.....	377
Fig. 14.7	A policy template for generating an authorisation policy for providing consent to a GP .....	377
Fig. 14.8	An authorisation policy for providing consent to a GP .....	378
Fig. 14.9	A TR policy for providing consent to a specialist .....	379
Fig. 14.10	A policy template for generating an authorisation policy for providing consent to a cardiologist.....	380
Fig. 14.11	An authorisation policy for providing consent to a cardiologist ..	380
Fig. 14.12	A policy template for generating an authorisation policy for providing consent to the emergency response team.....	381
Fig. 14.13	An authorisation policy for providing consent to the emergency response team .....	381
Fig. 15.1	Distinction among the EMR, PHR, and EHR.....	392
Fig. 15.2	An illustration of a private cloud in context of e-Health .....	394
Fig. 15.3	An illustration of a public cloud in context of e-Health .....	395
Fig. 15.4	An illustration of a hybrid cloud in context of e-Health .....	395
Fig. 15.5	Taxonomy of essential privacy requirements and patient-driven requirements .....	397
Fig. 16.1	HHS's new rules to address the risks of de-identified data that can be re-identified .....	431
Fig. 16.2	Privacy protection and number of released independent single nucleotide polymorphisms (SNPs) base on the report in [63].....	434

Fig. 16.3	Illustration of Homer’s attacks, where $ P_j - R_j $ and $ P_j - M_j $ measure how the person’s allele frequency $P_j$ differs from the allele frequencies of the reference population and the mixture, respectively. By sampling a large number of $N$ SNPs, the distance measure $D(P_j)$ will follow a normal distribution due to the central limit theorem. Then, a one-sample t-test for this individual over all sampled SNPs can be used to verify the hypothesis that an individual is in the mixture ( $T(P) > 0$ ) ..... 435
Fig. 17.1	Taxonomy tree of blocks ..... 451
Fig. 17.2	Tree for partitioning records ..... 452
Fig. 17.3	Boxplots of data utility of chr2 and chr10 data with different p-values ..... 456
Fig. 17.4	Privacy risk of chr2 and chr10 data. The <i>star</i> and <i>diamond</i> markers represent the test value of a specific individual in the case ( <i>left</i> ) or test ( <i>right</i> ) group, respectively. The <i>horizontal line</i> indicates the 0.95 confidence level for identifying case individuals that are estimated based on the test statistic values of test individuals ..... 457
Fig. 17.5	Comparison of data utility and privacy risk for chr2 and chr10 data with different privacy budget ..... 458
Fig. 18.1	Overview of the proposed framework to quantify kin genomic privacy [23]. Each vector $X^i$ ( $i \in \{1, \dots, n\}$ ) includes the set of SNPs for an individual in the targeted family. Furthermore, each letter pair in $X^i$ represents a SNP $x_j^i$ ; and for simplicity, each SNP $x_j^i$ can be represented using $\{BB, Bb, bb\}$ (or $\{0, 1, 2\}$ ). Once the health privacy is quantified, the family should ideally decide whether to reveal less or more of their genomic information through the genomic-privacy preserving mechanism (GPPM) ..... 466
Fig. 18.2	Family tree of <i>CEPH/Utah Pedigree 1463</i> consisting of the 11 family members that were considered. The notations <i>GP</i> , <i>P</i> , and <i>C</i> stand for “grandparent”, “parent”, and “child”, respectively. Also, the symbols ♂ and ♀ represent the male and female family members, respectively ..... 468

Fig. 18.3	Evolution of the genomic privacy of the parent (P5), with and without considering LD. For each family member, we reveal 50 randomly picked SNPs (among 100 SNPs in $S$ ), starting from the most distant family members, and the $x$ -axis represents the exact sequence of this disclosure. Note that $x = 0$ represents the prior distribution, when no genomic data is revealed .....	469
Fig. 18.4	Connections between the parties in the proposed protocol for privacy-preserving management of raw genomic data [4] .....	472
Fig. 18.5	Parts to be masked in the short reads for out-of-range content ....	473
Fig. 18.6	Parts to be masked in a short read based on patient's consent. The patient does not give consent to reveal the dark parts of the short read .....	473
Fig. 18.7	Proposed privacy-preserving disease susceptibility test (PDS) [6]	476
Fig. 18.8	Proposed system model for the privacy-preserving computation of the disease risk [5] .....	478
Fig. 18.9	System model for private use of genomic data in research setting [38]: participants (P), certified institution (CI), storage and processing unit (SPU), and medial units (MU) .....	481
Fig. 18.10	GenoGuard protocol [38]. A patient provides his biological sample to the CI, and chooses a password for honey encryption. The CI does the sequencing, encoding and password-based encryption, and then sends the ciphertext to the biobank. During a retrieval, a user (e.g., the patient or his doctor) requests for the ciphertext, decrypts it and finally decodes it to get the original sequence .....	483
Fig. 18.11	General protection framework. The GPPM [24] takes as inputs (i) the privacy levels of all family members, (ii) the genome of the donor, (iii) the privacy preferences of the family members, and (iv) the research utility. First, correlations between the SNPs (LD) is not considered in order to use combinatorial optimization. Note that we go only once through this box. Then, LD is used and a fine-tuning algorithm is used to cope with non-linear constraints. The algorithm outputs the set of SNPs that the donor can disclose ....	485
Fig. 19.1	A cryptosystem .....	498
Fig. 19.2	General scheme of the AES algorithm .....	501

Fig. 19.3	AES encryption in CBC mode. $B_i$ , $B_i^e$ and $K_e$ denote the plaintext block, the encrypted block and the encryption key, respectively. $iv$ is a random initialization vector .....	502
Fig. 19.4	Encryption/decryption processes of a stream cipher algorithm ...	502
Fig. 19.5	The principle of watermarking chain .....	504
Fig. 19.6	Example of two codebooks' cells in the mono-dimensional space (i.e. $x$ is a scalar value) considering an uniform quantization of quantization step $\Delta$ . Symbols $\circ$ and $\times$ denote cells' centers that encode 0 and 1, respectively. $d = \Delta/2$ establishes the measure of robustness to signal perturbations .....	509
Fig. 19.7	Insertion of a binary message using AQIM. $X_w$ represents the vector after the insertion of a bit equals to "1" within a host signal $X$ associated to a vector in the $N$ -dimensional space if $N$ pixels constitute $X$ . $\Delta$ is the quantization step, and circles and crosses represent the centers of the cells that encode the bits "0" and "1", respectively .....	510
Fig. 19.8	Basic principle of the histogram shifting modulation: (a) original histogram, and (b) histogram of the watermarked data .....	511
Fig. 19.9	General system architecture of a JWE algorithm. $M_e$ , $M_s$ , $K_e$ , $K_{ws}$ and $K_{we}$ are the message available in the encrypted domain, the message available in the spatial domain, the encryption and the watermarking keys in the spatial and the encrypted domain, respectively .....	517
Fig. 19.10	Examples of the images used for evaluation (using AES): (a) original PET image, (b) joint watermarked/ciphered image, (c) deciphered watermarked image, and (d) difference between the original image and the decrypted watermarked image .....	520
Fig. 20.1	Re-linkage using an imaging database .....	531
Fig. 20.2	Volume rendering from 3D Slicer .....	532
Fig. 20.3	Skull Stripping: 3dSkullstrip in AFNI ( <i>left</i> ); BET in FSL ( <i>middle</i> ); HWA in Freesurfer ( <i>right</i> ) [41] .....	536
Fig. 20.4	Defacing: Quickshear ( <i>top</i> ); MRI Defacer ( <i>bottom</i> ) [41] .....	537
Fig. 21.1	Passive HF RFID tags .....	552
Fig. 21.2	Passive UHF RFID tags .....	553
Fig. 21.3	Basic exchange between an RFID tag and reader .....	559
Fig. 21.4	Using the EPC number to retrieve additional information about the tag and associated asset .....	561
Fig. 22.1	Block diagram .....	572

Fig. 22.2	The decision graph leading to ECG segment classification. Given the array $y_1, \dots, y_6$ , the tree is traversed according to the result of the comparison of the values with 0 in each node, following the true ( $T$ ) or false ( $F$ ) edges .....	574
Fig. 22.3	Garbled circuit scheme .....	578
Fig. 22.4	Linear selection circuit (part of $C$ ) of a node .....	581
Fig. 22.5	Hybrid LBP protocol. For simplicity we assume that all the $y_i$ values can be packed in a single ciphertext .....	583
Fig. 22.6	Privacy-preserving ECG diagnosis .....	584
Fig. 22.7	Classification accuracy of dataset using 21 and 15 features .....	586
Fig. 22.8	Classification accuracy of dataset using 5 and 4 features .....	587
Fig. 22.9	A perceptron .....	591
Fig. 22.10	Transfer functions. (a) <b>tansig</b> . (b) <b>satlin</b> .....	592
Fig. 22.11	Classification accuracy as a function of the number of nodes in the hidden layer and <b>satlin</b> as activation function .....	593
Fig. 22.12	Neural network structure. In the ECG classification protocol $n = 4$ , $n_h = 6$ and $n_o = 6$ .....	593
Fig. 22.13	Classification accuracy as a function of $\ell^i$ , $\ell^h$ , $\ell^o$ .....	595
Fig. 22.14	Classification accuracy in function of $\ell^o$ , with $\ell^i = \ell^h = 13$ .....	595
Fig. 22.15	Hybrid implementation of the neural network .....	597
Fig. 22.16	Scheme to compute the SNR .....	600
Fig. 22.17	Sequence of steps performed to evaluate the quality of an ECG signal .....	604
Fig. 23.1	The participants of healthcare social networks .....	615
Fig. 24.1	Risks created by the lack of globally harmonisation data privacy standards .....	645
Fig. 24.2	Three main limitations to anonymisation of personal data .....	647
Fig. 24.3	Basic principles of national application, Part 2 of the OECD privacy guidelines [56] .....	649
Fig. 24.4	Basic principles of international application: free flow and legitimate restrictions, Part 4 of the OECD privacy guidelines [56] .....	650
Fig. 24.5	The UK data protection principles, Part 1 of Schedule 1 of the Data Protection Act 1998 [65] .....	657
Fig. 24.6	Conditions under which a covered entity is permitted to use and disclose PHI for research purposes [70] .....	661
Fig. 24.7	The seventeen HIPAA Privacy Rule de-identification fields (U.S. 2014: §164.514(b)(2)(i) [71]) .....	661
Fig. 24.8	HIPAA Privacy Rule limited dataset (U.S. 2014: §164.514(e)(2) [71]) .....	663
Fig. 24.9	Extract from the genome data release and resource sharing policy [31] .....	669

Fig. 24.10	Wellcome Trust policy on data management and sharing [76] .....	670
Fig. 25.1	Summary of OCR action on covered entities since 2008 .....	684
Fig. 25.2	Implications of Endsley's three-stage model on privacy protection .....	686
Fig. 25.3	A simplified illustration of information flow in a healthcare setting .....	689
Fig. 25.4	Example of SA errors in the registration process .....	690
Fig. 25.5	Level 1 SA error—failure to correctly perceive a situation or misperception of information .....	690
Fig. 25.6	Level 2 SA error—failure to comprehend situation or improper comprehension of information .....	691
Fig. 25.7	Level 3 SA error—incorrect projection of situation into the future .....	691
Fig. 26.1	U.S. HIPAA protected health information categories .....	700
Fig. 26.2	Examples of text de-identification with PHI tagging or resynthesis .....	701
Fig. 26.3	Workflow supported by MIST and its components .....	706
Fig. 26.4	The VHA BoB components and text processing pipeline .....	707
Fig. 26.5	Typical metrics for de-identification applications .....	711
Fig. 27.1	A fabricated sample medical record before and after surrogate generation .....	718
Fig. 27.2	Fabricated EMR prior to surrogate generation; PHI are delineated with XML tags .....	724
Fig. 27.3	Fabricated EMR after surrogate generation; PHI are delineated with XML tags .....	725
Fig. 27.4	Generic algorithm for replacing alphanumeric strings .....	726
Fig. 27.5	Algorithm for generating replacement names .....	728
Fig. 28.1	The focus on “identifiability” and “risk” in current data access processes .....	742
Fig. 28.2	An iterative approach to data access .....	757

# List of Tables

Table 2.1	Anonymization algorithms that protect from identity and attribute disclosure (adapted from [20], with permission from Elsevier) .....	22
Table 2.2	Algorithms that protect from identity disclosure based on demographics (Table adapted from [20], with permission from Elsevier) .....	26
Table 2.3	Algorithms that protect against identity disclosure based on diagnosis codes (Table adapted from [20], with permission from Elsevier) .....	27
Table 2.4	Algorithms for preventing attribute disclosure for demographics (Table adapted from [20], with permission from Elsevier) .....	29
Table 2.5	Algorithms for preventing attribute disclosure based on diagnosis codes (Table adapted from [20], with permission from Elsevier) .....	30
Table 3.1	A sample set-valued dataset [5] .....	49
Table 4.1	Original prescription database (derived from [11]) .....	63
Table 4.2	Disclosed prescription database ( $k = 2$ ) (derived from [11]) ....	63
Table 4.3	Odds ratios (OR) and statistical significance (SS) for males (M) and females (F) in original data set (kisbq18).....	70
Table 4.4	Odds ratios (OR) and statistical significance (SS) for males (M) and females (F) in original data set (kisbq20).....	71
Table 4.5	Logistic regressions for each dataset .....	76
Table 5.1	An example of an RT-dataset containing patient demographics and diseases .....	84
Table 5.2	(a) A 2-anonymous dataset with respect to relational attributes, (b) a $2^2$ -anonymous dataset with respect to the transaction attribute, and (c) a $(2, 2^2)$ -anonymous dataset...	85
Table 5.3	Comparison of data anonymization tools .....	87

Table 5.4	A $(2, 2^2)$ -anonymous dataset with privacy constraints $\mathcal{P} = \{Flu, Herpes\}$ and utility constraints $\mathcal{U} = \{\{Asthma, Flu\}, \{Herpes, Eczema\}\}$ .....	95
Table 5.5	Overview of the methods supported from SECRET A .....	101
Table 5.6	(a) A fictitious medical dataset presenting the Year-Of-Birth, Sex, Race, and Diagnosis codes of patients, (b) mapping of attribute Sex, and (c) mapping of attribute Race .....	102
Table 6.1	Example dataset and the result of applying the transformation $(1, 0)$ .....	117
Table 6.2	Datasets used in the experiments .....	140
Table 6.3	Runtime measures for risk-based anonymization .....	140
Table 6.4	Runtime measures for 5-anonymity .....	140
Table 6.5	Utility measures for risk-based anonymization .....	141
Table 6.6	Utility measures for 5-anonymity .....	141
Table 7.1	(a) Dataset comprised of diagnosis codes, and (b) diagnosis codes contained in the dataset of Table 7.1a and their description (reprinted from [15], with permission from Elsevier) .....	151
Table 7.2	(a) Utility requirements, and (b) anonymized dataset that <i>does not</i> satisfy the utility requirements (reprinted from [15], with permission from Elsevier).....	153
Table 7.3	Anonymized counterpart of the dataset in Table 7.1a, using the utility-constrained approach (reprinted from [15], with permission from Elsevier) .....	154
Table 7.4	Mappings between diagnosis codes and generalized terms, created by set-based generalization .....	157
Table 7.5	A possible dataset reconstructed from the dataset of Table 7.3 (reprinted from [15], with permission from Elsevier). .....	159
Table 7.6	Examples of hierarchy-based utility constraints .....	163
Table 7.7	Disassociation with a shared chunk (reprinted from [15], with permission from Elsevier) .....	169
Table 7.8	Sets of diagnosis codes that are added into the priority queue of CBA, and their support .....	170
Table 7.9	Anonymized dataset by CBA using the utility policy of Table 7.2a .....	171
Table 7.10	Disassociated dataset with a shared chunk (reprinted from [15], with permission from Elsevier) .....	172
Table 7.11	The result of functions $M_O$ and $M_A$ , for CBA and for a reconstructed dataset, produced by DIS .....	172
Table 7.12	MRE scores for each utility constraint in Table 7.2a .....	173



Table 7.13	Average percentage of records that are retrieved incorrectly, for workloads having different $\delta$ values and for: (a) $\mathcal{U}_1$ , (b) $\mathcal{U}_2$ , and (c) $\mathcal{U}_3$ .....	174
Table 8.1	An illustrative example of a composition attack; the shared common record is revealed by intersecting two corresponding equivalence classes .....	181
Table 8.2	An illustration of differential privacy based publications of datasets in Table 8.1. Anemia, Cancer, Migraine, Diabetes and Cough are all sensitive values in the datasets. The counts of sensitive values are noised and published with the equivalence class. It is difficult for an adversary to find true common sensitive values using noised counts. Note that the counts are small since we use the same datasets from Table 8.1 .....	186
Table 8.3	Comparison of risk of composition attack of two equivalence classes: Case 1 .....	188
Table 8.4	Comparison of risk of composition attack of two equivalence classes: Case 2 .....	189
Table 8.5	Domain size of different attributes .....	190
Table 9.1	Disclosure risk and information loss for the generated table .....	225
Table 10.1	Overview of the PPRL techniques reviewed in the chapter. ....	240
Table 10.2	Example of $k$ -anonymity, with $k = 2$ .....	242
Table 10.3	The Soundex encoding .....	250
Table 13.1	The responsibilities of the CEI centers .....	317
Table 13.2	The profile of the study cases and the selected sample to build the gold-standard (reprinted from [52], with permission from Elsevier) .....	338
Table 13.3	Assignment of sample cases to judges and results from the first round review <sup>a</sup> .....	342
Table 13.4	Comparison between the enhanced RBAC model and the CEIAdmin system .....	345
Table 13.5	Measuring the effectiveness of the enhanced RBAC model and the CEIAdmin system with a gold-standard .....	346
Table 15.1	Overview of the approaches employed to maintain confidentiality .....	403
Table 15.2	Overview of the approaches employed to maintain data integrity .....	405
Table 15.3	Overview of the approaches employed to offer collusion resistance .....	408
Table 15.4	Overview of the approaches employed to maintain anonymity .	411

Table 15.5	Overview of the approaches employed to maintain authenticity	413
Table 15.6	Overview of the approaches employed to maintain unlinkability	414
Table 16.1	Key aspects of GWAS and genomic data sharing (GDS) policies .....	429
Table 17.1	Raw genome data .....	447
Table 17.2	Genome data partitioned into blocks .....	450
Table 17.3	Anonymous data ( $\epsilon = 1, h = 2$ ) .....	452
Table 17.4	Data utility of chr2 data set with privacy budget of 1.0 and power of 0.01 .....	455
Table 17.5	Data utility of chr10 data set with privacy budget of 1.0 and power of 0.09 .....	455
Table 18.1	Frequently used notations .....	467
Table 20.1	Landmarks used in facial reconstruction [13, 15, 36, 52] and corresponding feature memberships.....	539
Table 21.1	The frequency band, RFID type, corresponding ISO standard, and typical applications.....	553
Table 22.1	Classification pattern (“*” means that the value of the variable does not influence the classification) .....	573
Table 22.2	QDF-based classification results (results from [1, Chap. 8]) ....	575
Table 22.3	QDF-based classification results obtained in the tests .....	575
Table 22.4	Protocols for secure evaluation of private LBPs .....	582
Table 22.5	Estimated communication complexity of LBP with QDF .....	588
Table 22.6	Performance of protocols for secure ECG classification through LBP .....	589
Table 22.7	Inputs to the GC implementation of the neural network .....	596
Table 22.8	Complexity of NN-based ECG classification protocol .....	597
Table 22.9	Number of bits necessary to represent the values obtained by a worst case analysis .....	602
Table 22.10	SNR protocol data transfer .....	603
Table 22.11	Maximum value and number of bits necessary for the magnitude representation of the variables involved in the computation by worst case analysis .....	607
Table 22.12	Bandwidth (bits) required by the protocol.....	607
Table 22.13	Performance of the protocol using the linear classifier or a single feature .....	608
Table 23.1	Identification techniques that may affect medical privacy (based on [10]) .....	619
Table 23.2	Privacy principles and their application is HSNs.....	623

Table 23.3	A list of privacy threats and associated risks in HSNs .....	623
Table 24.1	HIPAA Safe Harbor example de-identification of a simple medical dataset prior to de-identification and following de-identification (changed/removed data in bold) ....	662
Table 24.2	Deadlines for data submission and release in a supplement to the NIH Genomic Data Sharing Policy, which apply to all large-scale, NIH-funded genomics research .....	667
Table 25.1	The 12 largest privacy breaches identified by the OCR.....	683
Table 26.1	Advantages and disadvantages of methods used for text de-identification .....	702
Table 26.2	Sensitivity of machine learning-based text de-identification applications (partly reproduced from [19] .....	711
Table 26.3	Overall (micro-averaged) performance at the PHI level (partly reproduced from [19]) .....	712
Table 26.4	Sensitivity of de-identification applications when generalized to a different type of clinical notes .....	712
Table 27.1	HIPAA's list of PHI categories .....	723
Table 28.1	A comprehensive framework for proportionate governance: domains for adjunction and associated determination of risk .....	753
Table 28.2	The adjudication scenario 1 .....	758
Table 28.3	The adjudication scenario 2 .....	759