Guide to the De-Identification of Personal Health Information

Khaled El Emam

CRC Press
Taylor & Francis Group
AN AUERBACH BOOK

Guide to the De-Identification of Personal Health Information

OTHER INFORMATION SECURITY BOOKS FROM AUERBACH

Asset Protection through Security Awareness
Tyler Justin Speed
ISBN 978-1-4398-0982-2

The CISO Handbook: A Practical Guide to Securing Your Company
Michael Gentile, Ron Collette, and Thomas D. August
ISBN 978-0-8493-1952-5

CISO's Guide to Penetration Testing: A Framework to Plan, Manage, and Maximize Benefits
James S. Tiller
ISBN 978-1-4398-8027-2

The Complete Book of Data Anonymization: From Planning to Implementation
Balaji Raghunathan
ISBN 978-1-4398-7730-2

Cybersecurity: Public Sector Threats and Responses
Kim J. Andreasson, Editor
ISBN 9781-4398-4663-6

Cyber Security Essentials
James Graham, Editor
ISBN 978-1-4398-5123-4

Cybersecurity for Industrial Control Systems: SCADA, DCS, PLC, HMI, and SIS
Tyson Macaulay and Bryan L. Singer
ISBN 978-1-4398-0196-3

Cyberspace and Cybersecurity
George Kostopoulos Request
ISBN 978-1-4665-0133-1

Defense Against the Black Arts: How Hackers Do What They Do and How to Protect against It
Jesse Varsalone and Matthew McFadden
ISBN 978-1-4398-2119-0

The Definitive Guide to Complying with the HIPAA/HITECH Privacy and Security Rules
John J. Trinckes, Jr.
ISBN 978-1-4665-0767-8

Digital Forensics Explained
Greg Gogolin
ISBN 978-1-4398-7495-0

Digital Forensics for Handheld Devices
Eamon P. Doherty
ISBN 978-1-4398-9877-2

Electronically Stored Information: The Complete Guide to Management, Understanding, Acquisition, Storage, Search, and Retrieval
David R. Matthews
ISBN 978-1-4398-7726-5

FISMA Principles and Best Practices: Beyond Compliance
Patrick D. Howard
ISBN 978-1-4200-7829-9

Information Security Governance Simplified: From the Boardroom to the Keyboard
Todd Fitzgerald
ISBN 978-1-4398-1163-4

Information Technology Control and Audit, Fourth Edition
Sandra Senft, Frederick Gallegos, and Aleksandra Davis
ISBN 978-1-4398-9320-3

Managing the Insider Threat: No Dark Corners
Nick Catrantzos
ISBN 978-1-4398-7292-5

Network Attacks and Defenses: A Hands-on Approach
Zouheir Trabelsi, Kadhim Hayawi, Arwa Al Braiki, and Sujith Samuel Mathew
ISBN 978-1-4665-1794-3

PRAGMATIC Security Metrics: Applying Metametrics to Information Security
W. Krag Brotby and Gary Hinson
ISBN 978-1-4398-8152-1

The Security Risk Assessment Handbook: A Complete Guide for Performing Security Risk Assessments, Second Edition
Douglas Landoll
ISBN 978-1-4398-2148-0

The 7 Qualities of Highly Secure Software
Mano Paul
ISBN 978-1-4398-1446-8

Smart Grid Security: An End-to-End View of Security in the New Electrical Grid
Gilbert N. Sorebo and Michael C. Echols
ISBN 978-1-4398-5587-4

Windows Networking Tools: The Complete Guide to Management, Troubleshooting, and Security
Gilbert Held
ISBN 978-1-4665-1106-4

AUERBACH PUBLICATIONS
www.auerbach-publications.com
To Order Call: 1-800-272-7737 • Fax: 1-800-374-3401
E-mail: orders@crcpress.com

Guide to the De-Identification of Personal Health Information

Khaled El Emam

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN AUERBACH BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20130204

International Standard Book Number-13: 978-1-4665-7908-8 (eBook - PDF)

Contents

Foreword

Personal health information comprises the most sensitive and intimate details of one's life, such as those relating to one's physical or mental health, and the health history of one's family. Intuitively, we understand the importance of protecting health information in order to ensure the confidentiality of such personal data and the privacy of the individual to whom it relates. Personal health information must also be accurate, complete, and accessible to health care practitioners in order to provide individuals with necessary health care. At a broader level, for secondary uses that go beyond the treatment of the individual, health-related data are needed for the benefit of society as a whole. These vitally important secondary uses include activities to improve the quality of care, health research, and the management of publicly funded health care systems.

As the information and privacy commissioner of Ontario, Canada, my role includes the oversight of health privacy legislation governing the collection, use, and disclosure of personal health information by organizations and individuals involved in the delivery of health care services. Ontario's Personal Health Information Protection Act (PHIPA) aims to respect an individual's right to privacy in relationship to his or her own personal health information while accommodating the legitimate need to access health information for well-defined purposes. PHIPA does this in part by establishing clear rules for the use and disclosure of personal health information for secondary purposes. The object of these rules is to maximize the benefits of both respecting personal privacy and making health information accessible for purposes that serve society as a whole.

My office has long championed a model that enables multiple goals. This process, which forms the basis of privacy by design, seeks to retire the traditional zero-sum paradigm, pitting individual privacy rights against broader societal interests, in favor of a doubly enabling positive-sum model in which both values are maximized. In the health sector, privacy by design addresses this issue by protecting the privacy of personal health information while at the same time making available quality health data for valuable secondary purposes. I am delighted to introduce this guide, which provides a practical, risk-based methodology for the de-identification of personal health information—an excellent example of the privacy by design approach in the health information context.

The de-identification of sensitive personal health information is one of our most valuable tools for protecting individual privacy. The routine de-identification or anonymization of personal health information can help health information custodians to comply with data minimization principles, incorporated in PHIPA, that require personal health information not to be collected, used, or disclosed if other information will serve the purpose, and that no more identified health information should be collected, used, or disclosed than is reasonably necessary to meet the purpose. Routine de-identification also helps to prevent privacy breaches in the case of loss, theft, or unauthorized access to personal health information.

At the same time, the de-identification of personal health information can enable the use of health data for important secondary purposes, such as health-related research. Done in a manner that significantly minimizes the risk of re-identification, while maintaining a level of data quality that is appropriate for the secondary purpose, the de-identification of personal health information embodies a privacy by design solution and rejects the traditional model's false dichotomy of privacy vs. data quality. By arguing persuasively for the use of de-identification as a privacy-enhancing tool, and setting out a practical methodology for the use of de-identification techniques and re-identification risk measurement tools, this book provides a valuable and much needed resource for all data custodians who use or disclose personal health information for secondary purposes. Doubly enabling, privacy-enhancing tools like these, that embrace privacy by design, will ensure the continued availability of personal health information for valuable secondary purposes that benefit us all.

Dr. Ann Cavoukian
Information and Privacy Commissioner of Ontario

Acknowledgments

Some of the materials in this book were based on peer-reviewed publications by myself and my colleagues. Where appropriate, I have referenced the originating publications. I also acknowledge and thank all of my coauthors on that earlier work.

This work was originally funded by the Office of the Privacy Commissioner of Canada under its contributions program. I thank the commissioner's office for this support, without which it would have been difficult to get this project started. Other sources of supportive funding for material included in this book are the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council, the Canada Research Chairs Program, and the Public Health Agency of Canada through a number of contracts.

I thank the following diverse individuals who have contributed to our research and implementation of the concepts and methods in this book, for reviewing earlier drafts, and providing critical input throughout the last seven years that has shaped our approach to de-identification (in alphabetical order): Luk Arbuckle, Sadrul Chowdhury, Fida Dankar, Ben Eze, Anita Fineberg, Lisa Gaudette, Elizabeth Jonker, Gunes Koru, Sarah Lyons, Grant Middleton, Angelica Neisa, Michael Power, Hassan Quereshi, Sean Rose, Saeed Samet, Morvarid Sedhakar, and John Wunderlich.

Also, it is important to acknowledge our development teams at the Electronic Health Information Laboratory and Privacy Analytics, whose implementation of the metrics and concepts in this book have allowed us to apply them in many different health data releases over the last few years. These practical applications have informed much of our work. Finally, I thank Brian Dewar for his help in formatting and arranging the manuscript.

Glossary (Abbreviations and Acronyms)

AHRQ: Agency for Healthcare Research and Quality
AIDS: Acquired immune deficiency syndrome
CA: Confidentiality agreement
CFR: Code of Federal Regulations
CIHR: Canadian Institutes of Health Research
CR: Capture-recapture model
CTO: Chief technology officer
DC: Data custodian
DF: De-identified file
DHHS: Department of Health and Human Services
DNA: Deoxyribonucleic acid
DSA: Data sharing agreement
DSMB: Data Safety Management Board
EHR: Electronic health record: EMR: Electronic medical record
ePHI: Electronic personal health information
EU: European Union
FOIA: Freedom of Information Act
FOIPPA: Freedom of Information and Protection of Privacy Act
GAPP: Generally accepted privacy principles
GIC: Group Insurance Commission
GPS: Global positioning system
HAF: HIPAA authorization form
HCFA: Health Insurance Claim Form (The Centers for Medicare and Medicaid Services)
HIA: Health Information Act
HIPAA: Health Insurance Portability and Accountability Act
HITECH: Health Information Technology for Economic and Clinical Health Act
HIV: Human immunodeficiency virus
HPV: Human papillomavirus

HR: Human resources
ICD: International Statistical Classification of Diseases and Related Health Problems (codes)
IM: Information management
IP: Internet protocol (address)
IPC: Information and Privacy Commissioner
IRB: Institutional Review Board
ISO/IEC: International Organization for Standardization/International Electrotechnical Commission
ITRC: Identity Theft Resource Center
MCI: Mitigating controls assessment instrument
MITS: Management of information technology security
MRC: Medical Research Council (UK)
MRN: Medical record number
NCVHS: National Committee on Vital and Health Statistics
NDA: Non-disclosure agreement
NIH: National Institutes of Health
NIST: National Institutes of Standards and Technology
OCR: Optical character recognition
OECD: Organization for Economic Cooperation and Development
OLA: Optimal lattice anonymization
OSSPS: Operational Security Standard on Physical Security
PHI: Personal health information
PHIA: Personal Health Information Act
PHIPA: Personal Health Information Protection Act (Ontario)
PHIPA: Personal Health Information Protection and Access Act (New Brunswick)
PIA: Privacy impact assessment
PIPEDA: Personal Information Protection and Electronic Documents Act
PPHI: Protection of personal health information
PPIA: Protection of Personal Information Act
REB: Research Ethics Board
RFID: Radio frequency identification
SAS: Statistical Analysis Software
SEER: Surveillance epidemiology and end results
SID: State inpatient database
SNP: Single nucleotide polymorphism
SPSS: Statistical Package for the Social Sciences
SSHRC: Social Sciences and Humanities Research Council of Canada
SSL: Single sockets layer (protocol for encrypting information over the Internet)
SSN: Social security number
THIPA: The Health Information Protection Act

TRA: Threat and risk assessment
URL: Universal resource locator
USA PATRIOT Act: Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act
VA: Vulnerability assessment

Chapter 1

Introduction

There is great demand for data. This may be financial data, health data, Internet transaction or clickstream data, or it may be travel/movement data. Large volumes of data can now be analyzed quite efficiently to gain new insights on the phenomenon being modeled. Some have called ours the age of the algorithm and have heralded the rise of data science.

The purposes for demanding such data vary widely. For example, the data may be used to develop new services and products or improve the efficiency and effectiveness of existing ones, for research and public health purposes, or to inform or even change the behavior of the public. Access to data also promotes transparency and provides the citizenry with the means to ensure accountability in government and public agencies.

Our focus here is on health data. While many of the methods described and conclusions drawn over the next few chapters may be relevant and valid for other types of data, the context and all of the examples will be on health data only.

At a time when our health care system is under serious fiscal strain, and the population is aging with multiple chronic conditions, it is incumbent on us to use the vast quantities of health data that are being collected to find ways to address system inefficiencies and improve patient outcomes and patient safety. In fact, one can argue that not to do so would be irresponsible and counter to what the public expects and the public interest.

There is little question that providing greater access to data will have many benefits to society (for instance, see the general examples in [1]). Therefore, with this as our starting assumption, we will not discuss the benefits part of the equation here. Our focus will be on how to make health data more accessible in a responsible way by protecting the privacy of patients and remaining compliant with current legislation and regulations.

De-identification of personal data is an effective way to protect the privacy of patients when their data are used or disclosed, and is the topic of this book. There are other ways to responsibly share health information while protecting patient privacy, and we will also discuss some of these. However, we argue that there are compelling legal and practical reasons why de-identification should be considered as one of the main approaches to use. We then present a risk-based methodology for the de-identification of health information.

Currently there are no complete, documented, and repeatable methodologies for the de-identification of health data. While there is a community of practice in this area, the sharing of practical details within that community tends to be limited. The methodology described here is intended to fill that gap, and it is based on our experiences with the de-identification of health information since 2005.

Primary and Secondary Purposes

Identifiable health data can be used for primary purposes, such as providing patient care. This is distinct from secondary purposes. Secondary purposes are defined as "non-direct care use of personal health information (PHI) including but not limited to analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities" [2].

When data are used for secondary purposes, this sometimes means that the data already exist. The data have been collected for a primary purpose and are now in a database, such as an electronic medical record (EMR) or in an integrated data warehouse, and there is a desire to use or disclose it for a secondary purpose, say, a research project. Data can also be collected for secondary purposes, for example, when a survey is conducted for a public health initiative.

Many primary purposes require that the data be identifiable, and therefore de-identification is not a realistic option. For instance, when providing care to a patient, it is not possible to hide the identity of that patient during the encounter.

On the other hand, using and disclosing health data for secondary purposes will often *not* require that the patients be identifiable. For example, for many health services research studies the identity of patients is not necessary to perform the analysis, and it may not be necessary to have identifiable patient data for training medical students or for evaluating health plan performance.

But there will be situations where identity is also important. For example, a research study may need to contact patients who meet certain criteria to collect additional information. In such a case the identity of the patients would be needed to contact the patients. Alternatively, there may be a need to re-identify patients in a de-identified database. For example, a public health analysis may detect that certain patients have been exposed to a virus, say, those who traveled to a certain country,

and would want to identify those individuals to contact them and perform follow-up interviews and tests, and for contact tracing. In such a case there needs to be a mechanism for re-identification under controlled conditions.

The Spectrum of Risk for Data Access

Data custodians can share health data with different degrees of access restrictions. At one extreme, they can make data available publicly with no access restrictions. For example, data can be made available by posting it on a website with no access controls. This option means that the data custodian imposes no constraints on who gets the data and what the data recipients do with those data. Data recipients may analyze the data by themselves or link them with other databases to create more detailed and richer data sets, which they may then also make publicly available. The custodian may not even know who has copies of the data at any point in time. Data recipients may be in the same country or halfway across the world, and they may be professional analysts or hobbyists and amateurs experimenting with the data.

An important caveat with the no-access-restrictions model is that the data custodian will not be able to manage the quality of the data analysis that is performed using the data. It will not be possible to ensure the verisimilitude of conclusions drawn by others from manipulations of the data—these conclusions may contest some of the custodian's own conclusions. These conclusions may put the data custodian in a negative light. While this is not a privacy issue, it often acts as a deterrent for the public disclosure of health information.

At the other extreme, the data custodian can disclose the health data under some restricted access regime. There are many ways in which this can be operationalized. The data recipient would have to sign a data sharing agreement and may have to go through regular audits to ensure that she has good security and privacy practices in place to handle the data. The audits may be conducted by the data custodian herself, or the data recipient may be required to conduct third-party audits and send the results of these audits to the data custodian on a regular basis.

Each of the above two approaches is suitable under different circumstances, and the option chosen will depend on factors such as the sensitivity of the data, as well as the data custodian's resources and their proclivity for risk. Clearly the former option allows much greater access to data and is the cheapest for the data custodian, as it does not require oversight of the data recipients. However, for data custodians that are risk averse and have resources to put into oversight, then the latter option may be more attractive.

Between these two ends of the spectrum there will be multiple possible options. The full spectrum of tradeoffs is illustrated conceptually in Figure 1.1. As shown, if we consider the data release as a transaction, the transaction risk does not have to be at one of the extremes. We will consider some examples to illustrate the spectrum.

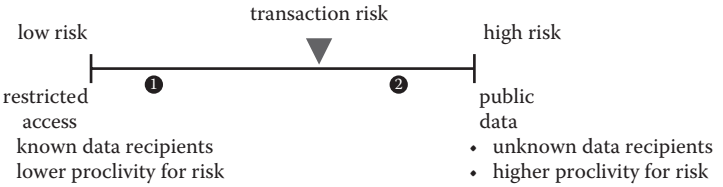

Figure 1.1 The tradeoffs for a data custodian in choosing the type of access to provide for data sets.

- A data custodian may be somewhat risk averse but does not have resources to provide significant oversight over data recipients. This means that she cannot really afford to conduct audits or enforce data sharing agreements. For instance, the data recipients may be researchers in other provinces/states/countries and the data custodian does not have the funds to conduct audits in other jurisdictions or halfway across the world. However, the data custodian may ask for third-party audits of the researchers' sites and practices, say, every three years. For such an organization transaction risk may be at point 1 in Figure 1.1.
- A data custodian may be running a public competition using her health data set. For example, this could be an analytics competition where the entrants have to build models using the data. However, the data custodian wants to create a deterrent against adversaries who may attempt to re-identify patients and contact them. Therefore all entrants must agree to terms of use for the data, and provide correct contact information. The data custodian is then prepared to legally pursue those who violate the terms rather than go and audit all data recipients (which would not be affordable because if the competition is successful, there could be thousands of entrants). In this case, the transaction risk would be around point 2 in Figure 1.1.

Once the transaction risk has been determined, even conceptually at this point, we need tools to manage that risk.

Managing Risk

De-identification methods provide a tool that allows data custodians to manage the transaction risk. This means that if a data custodian's transaction risk in Figure 1.1 is high, she can use de-identification methods to make it acceptable by reducing that risk. Even if the transaction risk is low, some de-identification would be needed because "low" does not mean "zero."

The amount of de-identification that is applied to a data set can be adjusted along a spectrum. As illustrated in Figure 1.2, the amount of de-identification can

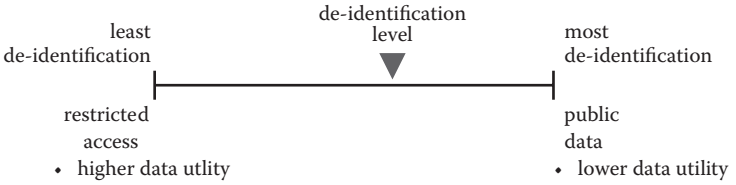

Figure 1.2 The use of de-identification to manage transaction risks.

be varied to manage the transaction risk. The rationale is easy to illustrate. One would not treat a public data release, on the web, the same way as the release to a trusted business or research partner. From a risk management perspective that would make no sense.

Consider Figure 1.3, which shows some of the tradeoffs. Following a balanced risk management approach, when the transaction risk is high, then more de-identification should be applied to the data to protect it, and when the transaction risk is low, then less de-identification is needed to protect the data.

A data custodian that always applied a lot of de-identification irrespective of the transaction risk would sometimes be conservative (quadrant 2) and sometimes have a risk-balanced outcome (quadrant 4). When he is conservative, it means that he unnecessarily incurs a high-cost burden to de-identify the data, and he will also unnecessarily use or disclose data that has a lower quality. The data custodian may also incur additional costs to ensure that the transaction risk is low, for example, by requiring a data sharing agreement with the data recipient. In quadrant 2 the data recipient has most likely invested in ensuring that its data management practices are strong, and so it will also incur some costs. Despite both the data recipient and the data custodian incurring a higher-cost burden, the data quality that is

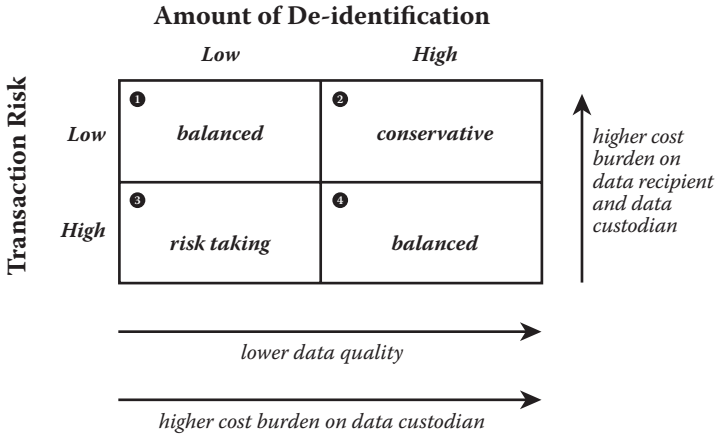

Figure 1.3 Tradeoffs from managing risk.

released is lower than it needs to be. This acts as a strong disincentive for the data recipient to invest in ensuring that the transaction risk is low.

Another common scenario is the data custodian who does not wish to incur any cost burden, and this puts him in quadrant 3. In that quadrant the data custodian is releasing high-quality data with a low amount of de-identification applied. There is also little cost incurred by the data recipient because he does not have to invest in improving their practices. However, the transaction risk is very high and it is not being managed. Operating in quadrant 3 can have significant negative legal, financial, and regulatory consequences on the data custodian. Let us consider some examples where a data release was in quadrant 3:

- With the intention of providing a real data set to be used by researchers working in the area of Internet search, AOL posted search queries from its clients on the web. *New York Times* reporters were able to re-identify one individual from these queries [3–5]. The bad publicity from this resulted in the CTO of the company resigning and the researcher who posted the data to lose his job.
- A Canadian national broadcaster aired a report on the death of a 26-year-old female taking a particular drug [6]. She was re-identified by the adverse drug reaction database released by Health Canada after matching with public obituary information. Subsequently, Health Canada restricted access to certain fields in the adverse drug event database that it releases, and litigation between the two organizations continued until 2008 in federal court.
- A re-identification attack on a movie ratings data set for a competition organized by Netflix [7] resulted in Netflix canceling a second competition and settling a class action lawsuit [8–10]. This had financial and reputational impacts on the organization.

A risk management approach would place the data custodian in quadrant 1 or 4. This concept of adjusting the level of de-identification to manage transaction risk will be explored in great detail and operationalized in the rest of the book.

There are two important implications from this risk-based approach to de-identification:

- The level of de-identification has an impact on data utility. The more de-identification that is applied, the lower the data utility. Here, reduced data utility means that the data have less information in them. Therefore, when the transaction risks are high, the data recipient will get lower utility level, and vice versa. Low data utility does not mean that analytics cannot be performed in the data. To the contrary, if de-identification is done properly, then the data are still useful for sophisticated data analysis. It is only when the de-identification is not optimized or not done adequately that data utility can be diminished significantly.

- The level of de-identification is not fixed for all data sets and for all data recipients. The same data set may be de-identified to a different extent depending on the transaction risk. For example, if the transaction risk is low, the amount of de-identification performed on the data may be quite small. This is because there are other factors that are in place to manage the overall risk, like data sharing agreements, audits to ensure that the data recipient can manage health information, and the most sensitive fields not being released to the data recipient. Therefore the totality of de-identification and other activities ensure that the risk is acceptable and that the data release is being done responsibly.

A risk-based approach to de-identification is not entirely new. Many organizations that have been disclosing and sharing data have been doing risk assessments for more than two decades. They may not have articulated the risk assessment precisely or formally, and the whole process may have been experiential. However, they did consider some of the same factors that we will be covering here. Our contribution is in formalizing this process, providing well-defined metrics and ways to interpret them, and by applying it specifically to health care data sets.

What Is De-Identification?

De-identification is, in general, intended to protect against inappropriate disclosure of personal information. In the disclosure control literature, there are two kinds of disclosure that are of concern: identity disclosure and attribute disclosure [11, 12]. The first type of disclosure is when an adversary can assign an identity to a record in a data set. For example, if the adversary would be able to determine that record number 10 belongs to patient Joe Bloggs, then this is identity disclosure. The second type of disclosure is when an adversary learns a sensitive attribute about a patient in the database with a sufficiently high probability without knowing which specific record belongs to that patient [11, 13]. For example, assume that in a specific data set all males born in 1967 had a creatine kinase lab test (a test often given to individuals showing symptoms of a heart attack), as illustrated in Table 1.1. Let's assume that an adversary knows that Joe Bloggs has a record in Table 1.1. This adversary does not need to know which record belongs to Joe Bloggs to know that he had that test if Joe was born in 1967. This is an example of attribute disclosure.

We only focus on identity disclosure and protections against that. There are three very pragmatic reasons why we do so: (1) Existing legislation and regulations only require protection against identity disclosure when a data set is de-identified, (2) there are no publicly known re-identification attacks that involve attribute disclosure, and (3) protections against attribute disclosure would destroy data utility for most analytical purposes. We examine each of these in turn.

All the analysis leading up to the U.S. Department of Health and Human Services (DHHS) issuing the (HIPAA) Privacy Rule de-identification standards

Table 1.1 Example of a Data Set with Lab Test Results

Sex	*Year of Birth*	*Lab Test*
Male	1959	Albumin, serum
Male	1967	Creatine kinase
Female	1955	Alkaline phosphatase
Male	1959	Bilirubin
Female	1942	BUN/creatinine ratio
Female	1975	Calcium, serum
Female	1966	Free thyroxine index
Female	1987	Globulin, total
Male	1959	B-type natriuretic peptide
Male	1967	Creatine kinase
Male	1968	Alanine aminotransferase
Female	1955	Cancer antigen 125
Male	1967	Creatine kinase
Male	1967	Creatine kinase
Female	1966	Creatinine
Female	1955	Triglycerides
Male	1967	Creatine kinase
Female	1956	Monocytes
Female	1956	HDL cholesterol
Male	1978	Neutrophils
Female	1966	Prothrombin time
Male	1967	Creatine kinase

has focused on protections against identity disclosure [14, 15]. HIPAA does not address identifiability risks from attribute disclosure. The case is similar for the different federal and provincial privacy and health privacy laws in Canada and the EU. Some of these definitions will be examined later on.

Known re-identification attacks of personal information that have actually occurred are all identity disclosures [19], for example:

1. Reporters figured out which queries belonged to a specific individual from a database of web search queries publicly posted by AOL [3–5].
2. Students re-identified individuals in the Chicago homicide database by linking it with the social security death index [16].
3. At least one individual was believed to be re-identified by linking his or her movie ratings in a publicly disclosed Netflix file to another public movie ratings database [7].
4. The insurance claims records of the governor of Massachusetts were re-identified by linking a claims database sold by the state employees' insurer with the voter registration list [17].
5. An expert witness re-identified most of the records in a neuroblastoma registry [18, 19].
6. A national broadcaster matched the adverse drug event database with public obituaries to re-identify a 26-year-old girl who died while taking a drug and filmed a documentary on the drug afterwards [6].
7. An individual in a prescriptions record database was re-identified by a neighbor [20].
8. The DHHS in the United States linked a large medical database with a commercial database and re-identified a handful of individuals [21].

In all of these cases the privacy breach was to assign individual identities to records that were ostensibly de-identified.

Finally, we illustrate the impact of attribute disclosure on the ability to perform analysis using an example. Consider Table 1.2, which is a data set showing whether daughters of parents with a particular religious affiliation are being vaccinated against HPV. HPV is a virus that is known to cause cervical cancer, and existing vaccines have been shown to provide effective protection. However, the data suggest that parents affiliated with religion A are not likely to vaccinate their daughters because they do not believe they will engage in sexual activity that would cause an infection, or put another way, if they vaccinate them, then they are admitting that they will engage in sexual activity (which is something they believe is not or should not be true). The relationship is statistically significant (chi-square test, $p < 0.05$).

However, the data set in Table 1.2 also has a big risk of attribute disclosure. We now know that individuals affiliated with religion A are not likely to vaccinate their daughters: The no vaccination rate is 89%. This may be stigmatizing information in

Table 1.2 Relationship between HPV Vaccination and Affiliation with Religion A

	HPV Vaccinated	*Not HPV Vaccinated*
Religion A	5	40
Religion B	40	5

Table 1.3 Relationship between HPV Vaccination and Affiliation with Religion A after Suppressing Some Records in the Data Set

	HPV Vaccinated	*Not HPV Vaccinated*
Religion A	5	6
Religion B	6	5

that it exposes those girls to a higher rate of infection and transmission of the virus. It is not necessary to know that Joe Bloggs is even in the data set to be able to draw this conclusion, let alone knowing which record belongs to Joe.

In Table 1.3 we have suppressed some records and removed them from the data set. Now it is not possible to draw a conclusion that there is a relationship between these two variables (the chi-square test is not statistically significant). This also eliminates the risk from attribute disclosure. But because attribute disclosure often represents the key relationships that we want to detect in a data set, the reduction in attribute disclosure also directly removes those relationships from the data. Eliminating interesting relationships in the data set would not be a desirable outcome of a de-identification exercise.

In the above example we are also able to make inferences about individuals who are not in the data, and if the relationship is robust, these inferences will likely be accurate. This is the basis of statistics and data analysis. If we eliminate the ability to draw inferences about the population at large, then we have crippled data analytics.

Therefore, as defined in this book, when a data set is de-identified, then the probability of assigning a correct identity to one or more records in the data is very small.

Attribute disclosure can still be considered a privacy breach under certain conditions. Based on our experiences, regulators do consider attribute disclosure issues when deciding whether a data set should be released or not. However, de-identification methods are not the appropriate tools to manage such risks. Other governance and regulatory mechanisms would need to be put in place. For example, in a research context ethics boards make decisions on whether a particular research question about a specific population is stigmatizing, and whether or how such research needs to be conducted and communicated. Outside the research world there are no such well-defined governance mechanisms, and this is a gap that would need to be addressed moving forward. However, it is not a gap that we are addressing here.

Learning Something New

The discussion of attribute disclosure brings us to a consideration of "learning something new" about the individuals in the data or the population at large. Figure 1.4 illustrates the possible scenarios that we consider. The columns indicate whether we

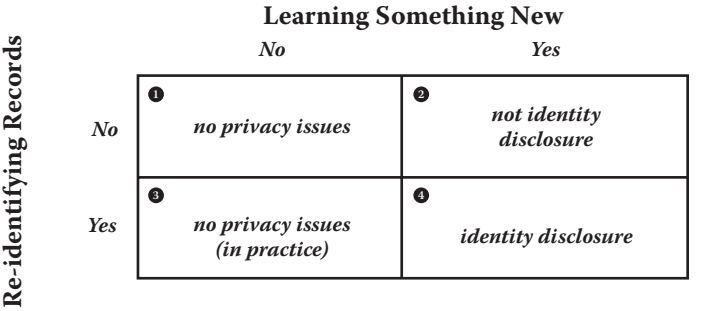

Figure 1.4 Situations where identity disclosure is relevant.

learn something new from the data, and the rows whether we can re-identify individuals in the data. For the latter, we do not make a distinction between whether new information learned is sensitive or not, or how sensitive it is. In quadrant 1 we do not learn anything new or re-identify individuals, so there would be no privacy concerns there, at least from the perspectives we are taking.

In quadrant 2 we learn something new about the individuals, but cannot re-identify their records. This is the classical attribute disclosure problem that we discussed above. While it is a potential privacy breach, it is not an identity disclosure issue.

The third quadrant is an interesting one. Here an individual is re-identified, but we do not learn something new about that individual. For example, consider a data set with two records about people who have had botulism, which is quite rare. Let's say that the data set has the age and gender of the individual, as well as the region of the country that they live in. Because it is rare and has a bad prognosis, these cases of botulism were also reported in the local media. An adversary then knows from the media that Joe Bloggs, who is male and 50, has botulism, and one of the records in the data set is for a 50-year-old male. In this case all of the information that exists in the data set is known by the adversary, and therefore the adversary learns nothing new. Strictly speaking, this is an identity disclosure problem because the adversary now knows that record number 2 in the data set is Joe Bloggs's, but the adversary does not learn anything new beyond what she used to re-identify the record. In practice, regulators have considered this not to be a disclosure of personal information. This is a pragmatic decision and makes a lot of common sense.

Finally, quadrant 4 is identity disclosure, and is the scenario that we focus on for the rest of this work.

The Status Quo

Today there are two distinct camps in the privacy community. One camp argues that current privacy protections through de-identification are inadequate and must

therefore be abandoned en masse. This extreme view is counterbalanced by another extreme view arguing that the status quo is perfectly fine and nothing needs to change. Both camps have a strong vested interest in their positions. The former have developed new privacy protection models and methods, and would like them adopted quickly; they are pushing out the old to bring in the new. The latter have been practicing current de-identification methods for some time and would not want their work undermined by new approaches that claim to be superior.

We have taken a third view, which falls in the middle of these two camps. Our argument is that some aspects of current de-identification practices need to be updated, but the basic premise of existing standards is still sound and should not be abandoned. This viewpoint is driven both by the existing evidence and by pragmatism.

We will start off by examining the status quo in terms of de-identification practices. This will help you understand why improvements to the status quo are needed, and why we cannot just continue as we were.

Current approaches that are used for the de-identification of health data are exemplified by the two standards in the HIPAA Privacy Rule: (1) the Safe Harbor standard and (2) the statistical method (also known as the expert determination method).

The Safe Harbor standard (henceforth "Safe Harbor") is a precise method for the de-identification of health information. It stipulates the removal or generalization of 18 variables from a data set (see Sidebar 1.1). The certainty and simplicity of Safe Harbor makes it quite attractive for health information custodians when disclosing data without patient consent [22], and it is used quite often in practice. The statistical method requires an expert to certify that "the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information" (see Sidebar 1.2).

Sidebar 1.1: The 18 HIPAA Safe Harbor Elements

The following identifiers of the individual, or of relatives, employers, or household members of the individual, are removed:

1. Names
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of a ZIP code if, according to the current publicly available data from the Bureau of the Census:

   a. The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people.
   b. The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, and date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
4. Telephone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web universal resource locators (URLs)
15. Internet protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

The catchall 18th element is "any other unique identifying number, characteristic, or code." Examples of interpretations of element 18 are clinical trial record numbers, unique keys that are derived from a date of birth or that are a hash value of a name and date of birth without a salt, and unique identifiers assigned to patients in electronic medical records.

Safe Harbor is relevant beyond the United States. For example, health research organizations in Canada choose to use the Safe Harbor criteria to de-identify data sets [23], Canadian sites conducting research funded by U.S. agencies need to comply with HIPAA [24], and international guidelines for the public disclosure of

clinical trials data have relied on Safe Harbor definitions [25]. Therefore the validity and strength of Safe Harbor is of importance to a broad international community.

Sidebar 1.2: The Statistical Method in the HIPAA Privacy Rule

A covered entity may determine that health information is not individually identifiable health information if a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

- Applying such principles and methods, determines that the risk is *very small* that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information
- Documents the methods and results of the analysis that justify such determination

The statistical method definition of what is considered de-identified data is consistent with definitions in other jurisdictions. For example, the Article 29 Data Protection Working Party in the EU notes that the term *identifiable* should account for "all means likely reasonably to be used either by the controller or by any other person to identify the said person" [26], and Ontario's Personal Health Information Protection Act states that identifying information means "information that identifies an individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify an individual."

As we illustrate below, Safe Harbor does not provide adequate protection in that it is possible to re-identify records from data sets that meet the Safe Harbor standard. The statistical method is a much better starting point for a de-identification methodology.

Safe Harbor-Compliant Data Can Have a High Risk of Re-Identification

Out of the 18 elements, only elements 2 and 3 would be included in data sets that are disclosed. The remaining elements would have to be either removed or replaced with pseudonyms. Any other information that is not specified in Sidebar 1.1 can

also be included in a data set, and the data set would be deemed compliant with the standard, for example, patient's profession, diagnosis codes, drugs dispensed, and laboratory tests ordered.

Below we describe common scenarios that can result in Safe Harbor-compliant data sets with a high risk of re-identification. To simplify these scenarios we assume that the data set is being disclosed publicly on the web. In such a case we assume that there is an adversary who will attempt a re-identification attack on the data.

The Adversary Knows Who Is in the Data

The Safe Harbor text does not explicitly state how re-identification risk is measured and what acceptable re-identification risk or the risk threshold is. However, the consultations and justifications provided by DHHS regarding Safe Harbor indicate that population uniqueness is the measure of re-identification risk that was being used in the analysis leading up to issuing the standard [14, 15]. Population uniqueness is defined as the proportion (or percentage) of individuals in the whole population who have unique values on the variables that can be used for re-identification.

There have been attempts at empirically measuring the actual re-identification risk of Safe Harbor data sets. One often cited analysis concluded that 0.04% of the U.S. population is unique in their gender, age in years, and first three digits of their ZIP code [27, 28]. Under an assumption that individuals in a data set are sampled with equal probabilities from the U.S. population, then a data set that meets the Safe Harbor requirements is expected to have a similar uniqueness value [29].

An important assumption being made here is that the adversary does not know who is in the data set. For example, if a particular data set is a random sample from the U.S. population, then it is reasonable to assume that an adversary would not know who is in that particular data set.

However, if the adversary does know who is in the data set, then the proportion of individuals that are unique can be much higher. Consider the small data set in Table 1.4 that is Safe Harbor compliant (it only includes age in years and the three digits of the ZIP code). All of the records in that data set are unique (i.e., 100% uniqueness) on age and ZIP3, which is quite different from the 0.04% value noted

Table 1.4 Example Data That Are Safe Harbor Compliant Where All Records Are Unique on Age and ZIP3

Gender	*Age*	*ZIP3*	*Clinical Information (e.g., diagnosis)*
M	55	112	Heart attack
F	53	114	Osteoporosis
M	24	134	Head injury

above. If the adversary knows that Tom is in this data set and knows Tom's age and the first three digits of his ZIP code, then Tom can be re-identified with certainty. In fact, the records of all three individuals, assuming that they are known to be in the data set, would be identifiable with certainty.

An adversary can know that an individual is in the data set under a number of different circumstances, for example:

1. Individuals may self-reveal that they are in the data set by mentioning that they are part of, say, a clinical trial to their colleagues or on their online social network.
2. It may be generally known whose records are in the data set, as in the case of an interview survey conducted in a company in which the participants missed half a day of work to participate. In such a case it is known within the company, and to an internal adversary, who is in the data set.
3. The data set may represent everyone in a particular group; for example, consider a registry on individuals with a rare and visible congenital anomaly in a particular state. If someone has that anomaly, then he would be in the registry with a high certainty, or if there is only one family doctor in a village, then all residents will be in a data set from that doctor's electronic medical record.

The Data Set Is Not a Random Sample from the U.S. Population

If a particular data set is a sample but it is not a random sample from the U.S. population, the percent of unique individuals can be quite high even for a data set that is Safe Harbor compliant. We will illustrate this with reference to the hospital discharge database for the state of New York for 2007. After cleaning, this database consists of approximately 1.5 million individuals who have been hospitalized. We took repeated 50% random samples from that data set. Any such sample drawn meets two criteria: (1) Hospitalized individuals in New York are not a random sample from the U.S. population, and (2) an adversary who knows that a patient has been hospitalized would not know if they are in the selected sample or not.

Our analysis shows that the expected percentage of unique individuals on gender, age in years, and the three-digit ZIP code was 0.71%, which is more than 17 times higher than the assumed Safe Harbor risk. If we restrict the cohort further to males over 65, then 0.91% are unique on these three variables, and males over 65 who were hospitalized for more than 14 days had a uniqueness of 4%, and those hospitalized more than 30 days had a uniqueness of 11.14%.

This example illustrates that if a particular cohort is not randomly selected from the U.S. population, the actual uniqueness can be dramatically higher than assumed, and it can be quite high in absolute terms as well. It would be fair to say that most health data sets are not random samples from the U.S. population, but rather they represent specific cohorts that can be quite different from the general population in their age and gender distribution.

Other Fields Can Be Used for Re-Identification

It is common for health data sets to have other kinds of fields that can be used for re-identification beyond the set included in Safe Harbor. Here we consider some examples of data elements that would pass the Safe Harbor standard but would still produce a data set with a high probability of re-identification.

First, Safe Harbor does not explicitly consider genetic data as part of the 18 fields to remove or generalize. There is evidence that a sequence of 30 to 80 independent single nucleotide polymorphisms (SNPs) could uniquely identify a single person [30]. There is also a risk of re-identification from pooled data, where it is possible to determine whether an individual is in a pool of several thousand SNPs using summary statistics on the proportion of individuals in the case or control group and the corresponding SNP value [31, 32].

Second, Safe Harbor does not consider longitudinal data. Longitudinal data contain information about multiple visits or episodes of care. For example, let us consider the state inpatient database for New York for the year 2007 again, which contains information on just over 2 million visits. Some patients had multiple visits and their ZIP code changed from one visit to the next. If we consider that the age and gender are fixed, and allow the three-digit ZIP code to change across visits (and the adversary knows those ZIP codes), then 1.8% of the patients are unique. If we assume that the adversary also knows the length of stay for each of the visits, then 20.75% of the patients are unique. Note that length of stay is not covered by Safe Harbor, and therefore can be included in the data set. Longitudinal information like the patient's three-digit ZIP code and length of stay may be known by neighbors, co-workers, relatives, and ex-spouses, and the public for famous people. As can be seen, there is a significant increase in uniqueness when the three-digit ZIP code is treated longitudinally, and a dramatic increase when other visit information is added to the data set.

Third, Safe Harbor does not deal with transactional data. For example, it has been shown that a series of International Statistical Classification of Diseases and Related Health Problems (ICD) diagnosis codes for patients makes a large percentage of individuals uniquely identifiable [33]. An adversary who is employed by the health care provider would have the diagnosis codes and patient identity, which can be used to re-identify records in a Safe Harbor-compliant data set.

Finally, other pieces of information that can re-identify individuals in free-form text and notes are not accounted for. The following actual example illustrates how the author used this information to re-identify a patient. In a series of medical records that had been de-identified using the Safe Harbor standard, there was one record about a patient with a specific (visible) injury. The notes mentioned the profession of the patient's father and hinted at the location of his work. This particular profession lists its members publicly. It was therefore possible to identify all individuals within that profession in that region. Searches through social networking sites allowed the identification of a matching patient (having the same surname) who

posted the details of the specific injury during that specific period matching exactly the medical record. The key pieces of information that made re-identification possible were the father's profession and region of work. Profession is not part of the Safe Harbor list, and the region was broad enough to meet the ZIP code restrictions.

By specifying a precise and limited set of fields to consider, the Safe Harbor standard provides a simple "cookie cutter" approach to de-identification. However, it also ignores the many other fields that can be used to re-identify individuals, reducing its effectiveness at providing meaningful universal protections for different kinds of data sets.

Moving Forward beyond Safe Harbor

We have argued that the application of Safe Harbor cannot provide meaningful guarantees about the probability of re-identification except in the following very narrow circumstances: (1) The adversary does not know who is in the data, (2) the data are a simple random sample from the U.S. population, and (3) all variables in the data that can potentially be used for re-identification are covered by the 18 elements. It would be prudent for data custodians to evaluate carefully whether their data disclosures are consistent with these narrow criteria before using the Safe Harbor standard. But these are very narrow criteria indeed, and would only be met in few circumstances. The Safe Harbor standard cannot be seen as a broad and generally applicable standard because of its narrow scope of application. Even if the criteria are met, there is evidence that Safe Harbor results in data sets that have diminished utility for some important purposes, such as research and comparative effectiveness evaluation [34, 35].

We encounter data custodians who insist on using the Safe Harbor standard because it provides certainty, even if it does not provide meaningful protection. For example, the Safe Harbor standard is being used for longitudinal data and for free-form text where an adversary would know who is in the data set. This is the danger with a prescriptive regulation that has unlimited scope—it encourages and embeds poor practices.

The second standard in HIPAA, the statistical method, provides a better basis for a general risk-based de-identification methodology. However, as formulated, it leaves considerable room for interpretation: The specification of the scale and value for "very small," or multiple context-dependent values. This means that different experts may provide different, and possibly inconsistent, answers. It has been argued that the statistical standard is not used often in practice because it is perceived as not precise and too complex [36]. Furthermore, there have been concerns about the liability of the statisticians should the data be re-identified at a later point [37]. Precise guidance on both of these items would go a long way to ensuring that the application of the statistical standard is more protective of privacy under different scenarios and is repeatable when followed by different organizations. Public

methodologies that are peer reviewed and open to scrutiny by the community can go a long way to ensure that the de-identification results are defensible.

In this book we provide a precise methodology for instantiating the statistical method.

Why We Wrote This Book

There are a number of reasons why we believed writing this book was important. These are enumerated as follows:

- There was a clear need to provide a more prescriptive process for the implementation of the "statistical method" in HIPAA. Given our critique of Safe Harbor, it was important to ensure that de-identification practices would meet certain quality standards through clear guidance to the community.
- There are examples of poor de-identification practices. Some we have already mentioned, such as AOL and Netflix in their public release of data [38]. While we do not claim that had they had this book, they would have done a better job, they at least would not have had an excuse for using poor de-identification practices.
- The Center for Democracy and Technology has promoted the idea of de-identification centers of excellence to serve as hubs of expertise and technology development to promote good practices in this area [39]. The material included here is intended to provide some practical information for such organizations.
- Many health information custodians are genuinely confused about what would be considered good practices for de-identification that are specific for health data. They would like to implement policies that will withstand scrutiny. The book can serve as the basis for developing and deploying risk-based de-identification policies.
- A group of experts convened by the Canadian Institute for Health Information and Canada Health Infoway developed a process guidance for de-identification that has been reviewed and found acceptable by large provincial data custodians and ministries [40]. This document, which is a *de facto* de-identification standard in Canada, covers basic concepts and metrics, but does not provide a detailed actionable process. Our book can serve as an implementation guide for that process standard.
- The disclosure control literature has good overviews of techniques that one can use [12, 41]. However, they do not provide a detailed methodology to follow in order to select and parameterize the techniques, many of the techniques they describe would not necessarily be adequate for health data, and many of them have not been applied on health data. For de-identification methods to be applied in practice, a coherent and repeatable methodology was needed that is known to work in practice in the health context.

Our concern was with the use and disclosure of individual-level health data. We discuss in detail the principles and methods that can be applied to ensure that the probability of assigning a correct identity to records in a data set is very small. We will also provide methods for determining what "very small" should be and what would be appropriate levels of access restrictions on these data.

Our approach is pragmatic and is based on our experiences with the use and disclosure of personal health information in Canada and the United States since 2005. We only present methods that work in practice and that we have found to be acceptable by the data analysis community. The objective is to give data custodians the tools to make decisions about the best way to use and disclose these data, but also ensure that the privacy of individuals is protected in a defensible way and that the resultant data can meet the analytic purpose of the use of disclosure.

We did not intend to write a literature review of the discipline of de-identification or disclosure control. There are already good reviews of statistical and computational disclosure control methods available [42, 43], and we did not wish to go over the same type of material. Rather, it is a selective assembly of practical information that can guide analysts in their efforts to create data sets with a known re-identification risk, and allow them to justifiably claim that the privacy concerns have been reasonably addressed. In practice, the issues that are not adequately addressed in the literature cause difficulties, such as how to measure re-identification risk in a defensible way and what is acceptable re-identification risk. These issues are covered here because they have to be in a real-world setting.

There were three audiences in mind when we wrote this book: (1) privacy professionals (for example, privacy officers, privacy lawyers, and those tasked with addressing privacy issues on research ethics boards), (2) policy makers and regulators, and (3) disclosure control professionals interested in risk measurement and management. To meet the needs of such a diverse group, the five sections of the book cover a broad set of legal, policy, and technical topics as follows:

Section I: The case for de-identifying personal health information. The first part of the book provides a detailed case for why de-identification is necessary and when it is advised to apply it when using and disclosing personal health information. This is essentially the business case for applying de-identification methods.

Section II: Understanding disclosure risks. In this part we situate and contextualize our risk-based methodology, and give a general overview of its steps. This part of the book gives important background.

Section III: Measuring re-identification risk. The measurement chapters explain in some detail how to measure re-identification risk. There are multiple dimensions to risk, and what is measured will depend on the assumptions made about re-identification attacks. Measurement is a critical foundation to our methodology.