Anonymizing Health Data
Case Studies and Methods to Get You Started

Khaled El Emam and Luk Arbuckle
# Table of Contents

Preface ................................................................. ix

1. Introduction ......................................................... 1
   To Anonymize or Not to Anonymize ............................... 1
   Consent, or Anonymization? ....................................... 2
   Penny Pinching .................................................... 3
   People Are Private ................................................ 4
   The Two Pillars of Anonymization ............................... 4
   Masking Standards ................................................ 5
   De-Identification Standards ...................................... 5
   Anonymization in the Wild ....................................... 8
   Organizational Readiness ......................................... 8
   Making It Practical ............................................... 9
   Use Cases ................................................................ 10
   Stigmatizing Analytics ............................................ 12
   Anonymization in Other Domains ............................... 13
   About This Book .................................................. 15

2. A Risk-Based De-Identification Methodology ...................... 19
   Basic Principles .................................................... 19
   Steps in the De-Identification Methodology .................. 21
   Step 1: Selecting Direct and Indirect Identifiers ............ 21
   Step 2: Setting the Threshold .................................... 22
   Step 3: Examining Plausible Attacks ............................ 23
   Step 4: De-Identifying the Data .................................. 25
   Step 5: Documenting the Process ............................... 26
   Measuring Risk Under Plausible Attacks ...................... 26
   T1: Deliberate Attempt at Re-Identification ................... 26
   T2: Inadvertent Attempt at Re-Identification ................. 28
3. Cross-Sectional Data: Research Registries. ........................................ 43
   Process Overview 43
   Secondary Uses and Disclosures 43
   Getting the Data 46
   Formulating the Protocol 47
   Negotiating with the Data Access Committee 48
   BORN Ontario 49
   BORN Data Set 50
   Risk Assessment 51
   Threat Modeling 51
   Results 52
   Year on Year: Reusing Risk Analyses 53
   Final Thoughts 54

4. Longitudinal Discharge Abstract Data: State Inpatient Databases. ............ 57
   Longitudinal Data 58
   Don't Treat It Like Cross-Sectional Data 60
   De-Identifying Under Complete Knowledge 61
   Approximate Complete Knowledge 63
   Exact Complete Knowledge 64
   Implementation 65
   Generalization Under Complete Knowledge 65
   The State Inpatient Database (SID) of California 66
   The SID of California and Open Data 66
   Risk Assessment 68
   Threat Modeling 68
   Results 68
   Final Thoughts 69

   The Heritage Health Prize 71
   Date Generalization 72
8. Free-Form Text: Electronic Medical Records
Not So Regular Expressions
General Approaches to Text Anonymization
Ways to Mark the Text as Anonymized
Evaluation Is Key
  Appropriate Metrics, Strict but Fair
  Standards for Recall, and a Risk-Based Approach
  Standards for Precision
Anonymization Rules
Informatics for Integrating Biology and the Bedside (i2b2)
i2b2 Text Data Set
Risk Assessment
  Threat Modeling
  A Rule-Based System
Results
Final Thoughts

9. Geospatial Aggregation: Dissemination Areas and ZIP Codes
Where the Wild Things Are
Being Good Neighbors
  Distance Between Neighbors
  Circle ofNeighbors
  Round Earth
  Flat Earth
Clustering Neighbors
  We All Have Boundaries
  Fast Nearest Neighbor
Too Close to Home
  Levels of Geoproxy Attacks
  Measuring Geoproxy Risk
Final Thoughts

10. Medical Codes: A Hackathon
Codes in Practice
Generalization
  The Digits of Diseases
  The Digits of Procedures
  The (Alpha)Digits of Drugs
Suppression
Shuffling