

SpringerBriefs in Electrical and Computer Engineering

For further volumes:

<http://www.springer.com/series/10059>

Aris Gkoulalas-Divanis • Grigorios Loukides

Anonymization of Electronic Medical Records to Support Clinical Analysis

Aris Gkoulalas-Divanis
IBM Research - Ireland
Damastown Industrial Estate
Mulhuddart, Ireland

Grigorios Loukides
Cardiff University
The Parade
Cardiff
United Kingdom

ISSN 2191-8112

ISBN 978-1-4614-5667-4

DOI 10.1007/978-1-4614-5668-1

Springer New York Heidelberg Dordrecht London

ISSN 2191-8120 (electronic)

ISBN 978-1-4614-5668-1 (eBook)

Library of Congress Control Number: 2012948006

© The Author(s) 2013

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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book grew out of the work of the authors on the problem of preserving the privacy of patient data, which started when they were postdoctoral researchers in the Health Information Privacy Laboratory, Department of Biomedical Informatics, Vanderbilt University. Part of their work was to understand the privacy threats that disseminating clinical data entails and to develop methods to eliminate these threats. The use of data derived from the Electronic Medical Record (EMR) system of the Vanderbilt University Medical Center enabled the authors to realize and appreciate the complexity of the problem and to gain valuable insights that led to developing practical solutions.

The structure of the book closely follows the order in which the authors undertook this research, and some of their works formed the basis of the material presented in this book. We started by realizing that disseminating EMR data requires addressing many important, and often unexplored, privacy issues. One of the main issues was to examine the re-identifiability of diagnosis codes. Towards this goal, we studied a range of attacks that may lead to patient re-identification and performed empirical studies to demonstrate the feasibility of these attacks. Using several large EMR datasets, we showed that the attacks we considered pose a serious privacy threat, which cannot be addressed by popular approaches. The work won a Distinguished Paper Award from the American Medical Informatics Association (AMIA) Annual Symposium in 2009 and appeared in a revised and extended form in the Journal of the American Medical Informatics Association (JAMIA). Having realized the importance of guaranteeing both data privacy and the usefulness of data in biomedical applications, we designed the first approach which guarantees that the disseminated data will be useful in validating Genome-Wide Association Studies. These studies attempt to find clinically meaningful associations between patients' diagnosis and genetic variations, and are considered as the holy grail of personalized medicine. The work was published in the Proceedings of the National Academy of Sciences in 2010 and was reported by the National Human Genome Research Institute (NHGRI) among the important advances in the last 10 years of genomic research (Eric D. Green, et al. in *Nature*, vol. 470, 2011). We also gained useful insights on the problem and future directions, when we were preparing two

tutorials that were presented at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) in 2011, and at the SIAM International Conference on Data Mining (SDM) in 2012. The slides of these tutorials serve as a helpful companion to the book and can be found at <http://www.zurich.ibm.com/medical-privacy-tutorial/> and http://www.siam.org/meetings/sdm12/gkoulalas_loukides.pdf, respectively.

This book is primarily addressed to computer science researchers and educators, who are interested in data privacy, data mining, and information systems, as well as to industry developers, and we believe that the book will serve as a valuable resource to them. Knowledge of data mining or medical methods and terminology is not a prerequisite, and formalism was kept at a minimum to enable readers with general computer science knowledge to understand the key challenges and solutions in privacy-preserving medical data sharing and to reflect on their relation with practical applications. By discussing a wide spectrum of privacy techniques and providing in-depth coverage of the most relevant ones, the book also aims at attracting data miners with little or no expertise in data privacy. The objective of the book is to inform readers about recent developments in the field of medical data privacy and to highlight promising avenues for academic and industrial research.

Dublin, Ireland
Cardiff, UK

Aris Gkoulalas-Divanis
Grigorios Loukides

Acknowledgements

This work would not have been possible without the support of several people. The authors would like to thank Bradley Malin and Joshua Denny for many useful discussions, as well as Sunny Wang for providing access to the data we used in the case study. We would also like to thank Hariklia Eleftherohorinou, Efi Kokiopoulou, Jianhua Shao and Michail Vlachos for their insightful comments that helped the presentation of this work.

We are also grateful to Hector Nazario, Susan Lagerstrom-Fife, Jennifer Maurer and Melissa Fearon from Springer and to the publication team at SpringerBriefs, for their great support and valuable assistance in the preparation and completion of this work. Their editing suggestions were valuable to improving the organization, readability and appearance of the manuscript.

Part of this research was funded by grant U01HG004603 of the National Human Genome Research Institute (NHGRI), National Institutes of Health. Grigorios Loukides' research is also supported by a Royal Academy of Engineering Research Fellowship.

Contents

- 1 Introduction** 1
 - 1.1 The Need for Sharing Electronic Medical Record Data..... 1
 - 1.2 The Threat of Patient Re-identification 2
 - 1.3 Preventing Patient Re-identification..... 4
 - 1.4 Aims and Organization of the Book..... 6
 - References 6
- 2 Overview of Patient Data Anonymization**..... 9
 - 2.1 Anonymizing Demographics 9
 - 2.1.1 Anonymization Principles 9
 - 2.1.2 Anonymization Algorithms 12
 - 2.2 Anonymizing Diagnosis Codes..... 14
 - 2.2.1 Anonymization Principles 15
 - 2.2.2 Generalization and Suppression Models..... 18
 - 2.2.3 Anonymization Algorithms 19
 - 2.3 Anonymizing Genomic Data 26
 - References 27
- 3 Re-identification of Clinical Data Through Diagnosis Information** 31
 - 3.1 Motivation 31
 - 3.2 Structure of the Datasets Used in the Attack 33
 - 3.3 Distinguishability Measure and Its Application to EMR Data..... 33
 - 3.4 Utility Measures..... 36
 - References 38
- 4 Preventing Re-identification While Supporting GWAS** 39
 - 4.1 Motivation 39
 - 4.2 Background..... 40
 - 4.2.1 Structure of the Data..... 41
 - 4.2.2 Privacy and Utility Policies 42
 - 4.2.3 Anonymization Strategy 43
 - 4.2.4 Information Loss Measure 44

4.3	Algorithms for Anonymizing Diagnosis Codes	45
4.3.1	Privacy Policy Extraction.....	45
4.3.2	Utility-Guided Anonymization of CLInical Profiles (UGACLIP)	47
4.3.3	Limitations of UGACLIP and the Clustering-Based Anonymizer (CBA)	49
	References	52
5	Case Study on Electronic Medical Records Data.....	55
5.1	Description of Datasets and Experimental Setup	55
5.2	The Impact of <i>Simple Suppression</i> on Preventing Data Linkage and on Data Utility	55
5.3	Utility of Anonymized Diagnosis Codes.....	57
5.3.1	Supporting GWAS Validation	58
5.3.2	Supporting Clinical Case Count Studies.....	60
	References	64
6	Conclusions and Open Research Challenges.....	65
6.1	Threats Beyond Patient Re-identification	66
6.2	Complex Data Sharing Scenarios	67
	References	68
	Index.....	71

List of Figures

Fig. 2.1	A classification of heuristic search strategies	13
Fig. 2.2	Summary of generalization models	14
Fig. 2.3	An example of: (a) original dataset, and (b), (c) two anonymized versions of it	15
Fig. 2.4	An example of: (a) original dataset containing public and sensitive items, (b) a $(0.5, 6, 2)$ -coherent version of it, and (c) a generalization hierarchy.....	17
Fig. 2.5	Subpartitions created during the execution of <i>Partition</i>	21
Fig. 2.6	An example of (a) complete two-anonymous dataset, created by <i>Partition</i> , and (b) 2^2 -anonymous dataset, created by <i>Apriori</i>	21
Fig. 2.7	An example of (a) $(0.5, 2, 2)$ -coherent dataset produced by <i>Greedy</i> , (b) SARs used by <i>SuppressControl</i> , (c) intermediate dataset produced by <i>SuppressControl</i> , and (d) 0.5 -uncertain dataset produced by <i>SuppressControl</i>	24
Fig. 3.1	An example of: (a) original dataset containing patient names and diagnosis codes, (b) a de-identified sample of this dataset, and (c) a generalized sample of this dataset	32
Fig. 3.2	Percentage of patient records in <i>VNEC</i> that are vulnerable to re-identification when data are released in their original form	34
Fig. 3.3	Characteristics of the <i>VNEC</i> dataset: (a) frequency of the ICD codes, and (b) number of ICD codes per record	35
Fig. 3.4	The result of replacing 401.0, 401.1, and 493.00, in the dataset of Fig. 3.1a, with $(401.0, 401.1, 493.00)$	36

Fig. 3.5	The five-digit code 250.02 and its ascendants at different levels of the ICD hierarchy	37
Fig. 4.1	Biomedical datasets (fictional) and policies employed by the proposed anonymization approach: (a) research data, (b) identified EMR data, (c) utility policy (d) privacy policy, and (e) a five-anonymization for the research data	41
Fig. 4.2	Example of a research dataset.....	51
Fig. 4.3	Anonymizing the data of Fig. 4.2 using CBA	51
Fig. 4.4	Anonymized dataset produced by applying CBA [5] to the dataset of Fig. 4.2	52
Fig. 5.1	Percentage of patient records in <i>VNEC</i> that are vulnerable to re-identification when (a) data are released in their original form, and (b) when ICD codes in <i>VNEC</i> , which are supported by at most $s\%$ of records in <i>VP</i> , are suppressed.....	56
Fig. 5.2	Data utility after applying simple suppression, as it is captured by (a) <i>SL</i> , and (b) <i>RSL</i>	58
Fig. 5.3	Re-identification risk (shown as a cumulative distribution function) of clinical profiles	58
Fig. 5.4	Utility constraint satisfaction at various levels of protection for: (a) <i>VNEC</i> , and (b) <i>VNEC_{KC}</i>	59
Fig. 5.5	Example of a query that requires counting the number of patients diagnosed with a certain set of ICD codes	61
Fig. 5.6	Relative Error (<i>RE</i>) vs. k for the single visit case and for (a) <i>VNEC</i> , and (b) <i>VNEC_{KC}</i>	62
Fig. 5.7	Relative Error (<i>RE</i>) vs. k for the all-visits case and for (a) <i>VNEC</i> , and (b) <i>VNEC_{KC}</i>	63
Fig. 5.8	Mean of Relative Error (<i>ARE</i>) vs. k for the single-visits case and for (a) <i>VNEC</i> , and (b) <i>VNEC_{KC}</i>	63

List of Tables

Table 1.1	An example of: (a) original dataset, and (b) a de-identified version of it	3
Table 2.1	(a) Original dataset, and (b), (c) two different four-anonymous versions of it.....	10
Table 2.2	Summary of privacy principles for guarding against sensitive information disclosure.....	10
Table 2.3	Summary of existing grouping strategies w.r.t. their objectives	13
Table 2.4	Summary of algorithms for preventing identity disclosure in transaction data publishing	20
Table 3.1	Description of the <i>VNEC</i> dataset	35
Table 5.1	Description of the datasets used	56
Table 5.2	(a) Percentage of retained disease information after applying simple suppression with varying s , and (b) disease information retained after applying simple suppression with varying s	57
Table 5.3	Satisfied utility constraints for UGACLIP and ACLIP when $k = 2$ and for (a) the single-visit case, and (b) the all-visits case	60
Table 5.4	Satisfied utility constraints for UGACLIP and CBA for the single-visit case when (a) $k = 5$, and (b) $k = 10$	61

List of Algorithms

Algorithm 1	Partition($\tilde{\mathcal{D}}, \mathcal{C}, \mathcal{H}, k$)	20
Algorithm 2	Apriori($\tilde{\mathcal{D}}, \mathcal{H}, k, m$)	22
Algorithm 3	Greedy($\mathcal{D}, h, k, p$)	24
Algorithm 4	SuppressControl($\mathcal{D}, \rho$)	26
Algorithm 5	PPE($\mathcal{D}, \mathcal{F}, k$)	46
Algorithm 6	UGACLIP($\mathcal{D}, \mathcal{P}, \mathcal{U}, k$)	48
Algorithm 7	CBA($\mathcal{D}, \mathcal{P}, \mathcal{U}, k$)	50

