Enabling Genomic-Phenomic Association Discovery without Sacrificing Anonymity

Raymond D. Heatherly[1]*, Grigorios Loukides[2], Joshua C. Denny[3], Jonathan L. Haines[4], Dan M. Roden[5], Bradley A. Malin[6]

1 Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, Tennessee, United States of America, **2** School of Computer Science and Informatics, Cardiff University, Cardiff, Wales, United Kingdom, **3** Department of Biomedical Informatics/Department of Medicine, School of Medicine, Vanderbilt University, Nashville, Tennessee, United States of America, **4** Department of Molecular Physiology and Biophysics/Center for Human Genetics Research, School of Medicine, Vanderbilt University, Nashville, Tennessee, United States of America, **5** Department of Medicine/Department of Pharmacology, School of Medicine, Vanderbilt University, Nashville, Tennessee, United States of America, **6** Department of Biomedical Informatics, School of Medicine/Department of Electrical Engineering and Computer Science, School of Engineering, Vanderbilt University, Nashville, Tennessee, United States of America

Abstract

Health information technologies facilitate the collection of massive quantities of patient-level data. A growing body of research demonstrates that such information can support novel, large-scale biomedical investigations at a fraction of the cost of traditional prospective studies. While healthcare organizations are being encouraged to share these data in a de-identified form, there is hesitation over concerns that it will allow corresponding patients to be re-identified. Currently proposed technologies to anonymize clinical data may make unrealistic assumptions with respect to the capabilities of a recipient to ascertain a patients identity. We show that more pragmatic assumptions enable the design of anonymization algorithms that permit the dissemination of detailed clinical profiles with provable guarantees of protection. We demonstrate this strategy with a dataset of over one million medical records and show that 192 genotype-phenotype associations can be discovered with fidelity equivalent to non-anonymized clinical data.

Citation: Heatherly RD, Loukides G, Denny JC, Haines JL, Roden DM, et al. (2013) Enabling Genomic-Phenomic Association Discovery without Sacrificing Anonymity. PLoS ONE 8(2): e53875. doi:10.1371/journal.pone.0053875

Editor: Noam Shomron, Tel Aviv University, Israel

Received November 15, 2012; **Accepted** December 3, 2012; **Published** February 6, 2013

Funding: This research was sponsored, in part, by grants U01HG006385 and U01HG006378 from the National Human Genome Research Institute and grants R01LM009989 and R01LM010685 from the National Library of Medicine. Dr. Loukides' research is supported by a Royal Academy of Engineering Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: r.heatherly@vanderbilt.edu

Introduction

Routine clinical care generates detailed, longitudinal information about a patient's health, medications, allergies, and treatment response. Recording and preserving these data, typically through an electronic medical record (EMR), can enable greater efficiency and effectiveness in the actions of care providers [1–3]. In the hopes of realizing the full potential of health information technology, the past several years has witnessed dramatic growth in the quantity and quality of clinical data [4], which, in turn, has become an invaluable resource for a wide range of secondary (i.e., not direct care) endeavors [5,6], including public health [7,8], quality assessment [9], and medical research [10,11]. With regard to the latter, EMRs are increasingly linked to biorepositories to enable large cost-effective association studies between genomes and an expanding range of phenotypes [12–15], such as atrioventricular conduction [16], white [17] and red [18] blood cell traits, hypothyroidism [19], and, more recently, the study of pharmacogenetic traits, including clopidogrel-response [20] and warfarin dose [21]. To facilitate transparency and enable reuse, collections of genotypes and DNA sequences tied to clinical knowledge are shared beyond the originating healthcare institutions, such as through the Database of Genotypes and Phenotypes (dbGaP) at the National Institutes of Health [22].

The majority of datasets currently shared via dbGaP, and similar environments, enable validation of known findings [23], but they lack the phenotypic detail necessary to support novel scientific investigations, thus slowing or preventing innovative biomedical research. A major obstacle to dissemination of clinically-rich datasets is the concern that disclosure of detailed records can cause privacy breaches, particularly in the form of patient re-identification [24,25]. Indeed, a growing number of studies illustrate how simple patient-specific data, such as demographics [26–29], hospital visit patterns [30], or insurance billing codes [31] – which correspond to International Classification of Diseases - 9th Revision (ICD-9) and are a core element of clinical phenotype specifications [13] – can be exploited for identification purposes. An additional concern is the contention that DNA sequence information is inherently identifiable [32], although patient-specific sequence databases to create such vulnerabilities are not (yet) generally available [33].

Concerns over re-identification can be mitigated through pragmatic governance models that integrate ethical, legal, and technical controls [34–37]. From a technical perspective, various approaches for the anonymization of patient-specific data have been proposed [38,39], but they are limited in their scope by considering unrealistically strong attackers. Of particular importance for the dissemination of clinical data, Loukides et al.

introduced an anonymization method for billing codes [40], but assumed the recipient of the data knows that a specific patient is a member of the cohort. While such a threat is plausible, it is not always likely and, in many situations, it is prudent for healthcare institutions to assume more realistic adversaries: for example, a recipient may only know that an individual was a patient at the hospital and not that they were a member of a specific research cohort [41,42]. We hypothesize that using larger populations for anonymization will yield more accurate biomedical knowledge discovery.

To investigate this hypothesis, we developed methods to anonymize datasets that contain a large amount of clinical data that account for varying degrees of a recipient's knowledge. To assess our models, we conducted an evaluation with three datasets derived from the EMR system of the Vanderbilt University Medical Center (VUMC), covering over one million patient records. Our findings illustrate that making more pragmatic assumptions on the capabilities of the recipient enables the dissemination of significantly greater quantities of patient-specific data in comparison to prior approaches. We find this method enables the dissemination of privacy-protected clinical data that support the discovery of phenome-wide associations equal to those previously published using non-protected information [43].

Results

We evaluated the influence of anonymization on two distinct types of knowledge discovery criteria. First, we summarize the quantity of clinical information retained in the anonymized datasets in comparison to the original resource. This provides a general sense of the quantity of clinical knowledge that can be disseminated. Second, we conducted Phenome-wide Association Studies (PheWAS) to characterize the extent to which phenotype-genotype associations are retained. In this scenario, all of our assessments are performed on the DEMO dataset.

Retention of General Clinical Information

Table 1 summarizes the quantity of clinical information retained in the anonymized datasets. We represent the changes through the use of two measures: Diagnosis Coverage (DC) and Code Coverage (CC). Diagnosis Count is a general measure of how many unique diagnoses are contained in the anonymized data, while Code Count is a measure of how many unique ICD-9 codes appear in the anonymized data. First, we examine the changes to the datasets resulting from anonymization process. It can be seen that *SD-Anon* yields the best retention of clinical information (99.99% DC and 99.98% CC) and *BioVU-Anon* has a slightly higher DC than $DEMO_D$ (99.99% and 99.57%, respectively). However, $DEMO_D$ has a higher CC than *BioVU-Anon* (80.78% and 77.02%, respectively). It is worth noting that this finding is influenced by the difference in CC and DC in the initial subsets (i.e., *DEMO* and *BioVU*). If the counts are considered in relation to the original SD, then *BioVU-Anon* has a DC of 19.71% and a CC of 67.65%, while $DEMO_D$ has a DC of 2.02% and a CC of 46.80%, showing that BioVU-Anon retains more information than $DEMO_D$ overall.

Next, we compare the information retained in the three Demonstration groups. Again, we see that $DEMO_S$, the SD anonymization, performs better than either of the other anonymizations (99.99% DC and 99.93% CC). Additionally, we find that $DEMO_D$ performs better than $DEMO_B$ in DC (99.57% and 91.50%, respectively), whereas the reverse is true for CC (80.78% and 96.84%, respectively).

Table 1. Summary statistics and information retention for the datasets in this study.

Original Dataset ***Anonymized Version***	**Code Count**		**Diagnosis Count**		**Population Size**
Synthetic					
Derivative(SD)*	**15,115**	**–**	**13,432,263**	**–**	**1,366,786**
SD-Anon	*15,112*	*99.98%*	*13,431,347*	*99.99%*	*1,366,552*
BioVU*	**13,275**	**–**	**2,647,056**	**–**	**104,904**
BioVU-Anon	*10,225*	*(77.02%)*	*2,646,872*	*(99.99%)*	*104,790*
Demonstration					
Group (DEMO)*	**8,734**	**–**	**272,080**	**–**	**5,994**
$DEMO_S$	*8,747*	*(99.93%)*	*272,043*	*(99.99%)*	*5,994*
$DEMO_B$	*8,476*	*(99.93%)*	*248,925*	*(91.50%)*	*5,595*
$DEMO_D$	*7,071*	*(80.78%)*	*270,867*	*(99.57%)*	*5,971*

Code Count and Diagnosis Count are the number of unique ICD-9 (or generalized set of ICD-9 codes) and total number of diagnoses for all records in the anonymized dataset, respectively. In this table, *corresponds to the original (i.e., non-anonymized) datasets.

doi:10.1371/journal.pone.0053875.t001

In combination, these findings partially confirm our earlier hypothesis. $DEMO_S$, which is derived from the largest population, results in the best retention of general clinical information among the DEMO anonymizations. However, neither $DEMO_B$ nor $DEMO_D$ clearly outperforms the other, which we discuss below.

In Tables 2 and 3, we show the full results we measured following the anonymization. In Table 2, we report the expanded DC information. In the left part of the chart, we report our findings without generalization - that is, if codes which occur in fewer than *k* records are removed from the set, rather than aggregated. For this, we report three measures. First, the count of diagnoses in the data set. Second, the as a percentage of the count to the total number of diagnoses in the SD (SD%). This measure allows us to determine how much information this anonymization retains of the entire population data set. Finally, we report the ratio of the count to the total number of diagnoses in its similar, non-anonymized data set. For example, for *BioVU-Anon*, the Local % measure compares to *BioVU*. As we hypothesized, we see in that even without generalization, *SD-Anon* still represents the highest retention of data of the SD (99.97%). We further see that BioVU-Anon and $DEMO_D$ each have lower percentages of the SD, as is expected. However, we also see that they each contain less

Table 2. Full Diagnosis Count information retained.

Dataset	**Diagnosis Count**			**Diagnosis Count**		
	without Generalization			**with Generalization**		
	Count	**SD %**	**Local %**	**Count**	**SD %**	**Local %**
SD-Anon	13428542	99.97%	99.97%	13431347	99.99%	99.99%
BioVU-Anon	2639298	19.65%	99.71%	2643872	19.68%	99.88%
$DEMO_D$	269868	2.01%	99.20%	270867	2.02%	99.57%
$DEMO_S$	271970	2.02%	99.97%	272043	2.03%	99.99%
$DEMO_B$	248467	1.85%	91.33%	248925	1.85%	91.50%

doi:10.1371/journal.pone.0053875.t002

Table 3. Full Code Count information retained.

Dataset	Code Count			Code Count		
	without Generalization			with Generalization		
	Count	SD %	Local %	Count	SD %	Local %
SD-Anon	13525	89.48%	89.48%	15112	99.98%	99.98%
BioVU-Anon	9785	64.74%	73.71%	10225	67.65%	77.02%
DEMOD	6952	45.99%	79.42%	7071	46.78%	80.78%
DEMOS	8681	57.43%	99.18%	8747	57.87%	99.93%
DEMOB	8179	54.11%	93.44%	8476	56.08%	96.84%

doi:10.1371/journal.pone.0053875.t003

information with respect to their original dataset as well (99.71% and 99.20%, respectively).

If we turn our attention to a comparison of the three Demonstration datasets, we see that $DEMO_S$ is still, as expected, the best performing. We note that our earlier observation about the divergence from our expectation degrading $DEMO_D$ and $DEMO_B$ still holds true here.

Next, in Table 3, we show the extended CC information retained. Similarly to Table 3, we report the original count and use the same additional measures to determine the information retention.

Here, we note several interesting findings. First, while the DC without and with generalization (99.97%, 99.99%) were similar, this is not the case with CC. Without generalization, *SD-Anon* retains only 89.48% of codes, while with generalization this retention is 99.98%. Considering these two statistics together, we see that by generalizing, we are able to release information about 10.5% of codes - approximately 1,500 codes – that, while generalized in the release, would be completely absent from the data were we to simply suppress them. We see similar, though less dramatic, results from the other two data sets.

If we again turn our attention to a comparison of the three DEMO sets, we see that, again, $DEMO_S$ clearly has more information retention than the other two DEMO sets. Most strikingly, $DEMO_S$, using the Local % measure, retains almost 30% more codes in the anonymized set (approximately 4,500 codes) over $DEMO_D$ even without generalization. If we also consider generalized codes, $DEMO_S$ keeps approximately 19% more codes over $DEMO_D$.

Again, we note that the difference in retention between $DEMO_D$ and $DEMO_B$ that was visible in the DC measure does not appear in the Code Count measure. Instead, we see that, as hypothesized, the larger the original set of data that is anonymized, the more information we retain in the anonymized data set (as measured by number of codes available for evaluation). Even though $DEMO_B$ performs better using this measure, the clearly superior data set is $DEMO_S$.

Retention of Genotype-Phenotype Associations

A Phenome-Wide Association Study (PheWAS) [19,44,45] assesses which clinical phenotypes from across a collection of concepts (in this case, a set of related billing codes grouped according to semantic similarity) are associated with a specific genomic region of interest. Patients are marked as either cases or controls according to the presence and absence of certain billing codes. In its simplest form the analysis determines the genotype distributions and calculates a χ^2 statistic, with an associated p-value and an odds-ratio. Conditions with $p \leq 0.05$ corrected for multiple comparisons are considered significant. We note that the point of the analysis in this paper is to determine the level of information loss resulting from anonymization of this dataset; specifically, we acknowledge that each of the conditions labeled as significant here only indicate a potential significance in a PheWAS discovery analysis, and are not necessarily conclusive. To conduct the present analysis, we focus on the anonymized demonstration cohorts and the six single nucleotide polymorphisms (SNPs) analyzed in the original PheWAS study of Denny et al. [43]. Table 4 compares the associations discovered in the anonymized and original datasets.

We first look at Type I errors, which we refer to as lost associations. These correspond to conditions determined to be significant in the original PheWAS study on DEMO, but were found to have $p > 0.05$ in an anonymized dataset. It can be seen that only $DEMO_S$ yields no lost associations across all SNPs. By contrast, $DEMO_D$ has only a single SNP (*rs1333049*) where there were no lost associations. Every other SNP has at least one lost association. $DEMO_B$ sustained lost associations in each SNP. Notably, $DEMO_B$ sustained a larger number of lost associations than $DEMO_D$ for every SNP.

Next, we turn our attention to Type II errors, which we refer to as false associations. These correspond to conditions with $p > 0.05$ in the original study, but $p \leq 0.05$ in the anonymized dataset. It can be observed that $DEMO_S$ is the only anonymization which in all cases has no additional significant associations reported. Similarly, $DEMO_D$ has one SNP (*rs6457620*) which has no new associations, though this is not the same SNP that sustained no lost associations. Again, $DEMO_B$ yielded new associations for each SNP.

PheWAS Case Studies

Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, and Figure 6 illustrate how *p*-values change for the phenotype associations across all SNPs in the form of QQ-plots. A perfect similarity would be represented by the line $y = x$, such that points along this line indicate the value in the original and anonymized analysis are equivalent. By contrast, points that deviate from this line indicate a change in the p value, such that the distance to the line indicates the magnitude of change. It can be seen that the $DEMO_S$ results lie consistently along the basis line, indicating that the values calculated in the original and anonymized PheWAS were approximately equivalent. For $DEMO_B$ and $DEMO_D$, however, there are more differences in the *p*-values between those derived from the original and anonymized datasets. Below, we highlight some specific changes in each PheWAS study conducted.

The plots in Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, and Figure 12 depict the change in $-log_{10}(p)$-values. It can be seen that $DEMO_B$ resulted in value changes of various frequencies, but that the $DEMO_S$ remained consistent across the association studies, with nearly 700 conditions (i.e., ICD-9 groupings) retaining their original *p*-value. We can also see that $DEMO_D$, while having fewer and smaller changes, those changes were significant enough to alter the effects of the significant conditions.

rs1333049. Figure 1 indicates there were at least three associations expected between approximately 0.5 and 1. Yet, in the anonymized dataset, the *p*-values for these associations were all close to 0, indicating a high likelihood of association. An example of a condition in this affected region is 440.00 *atherosclerosis*. There were also lost associations, such as condition 486 pneumonia, unspecified organism which, in the original PheWAS, was considered a significant association, but $DEMO_B$ has a changed *p*-value such that the condition is now non-significant.

Table 4. Results of anonymization on PheWAS Analysis for six SNPs.

SNP	Phenotype Associations at $p \leq 0.05$ in PheWAS						
	Original number of associations	Lost Associations			False Associations		
		(Type I Error)			(Type II Error)		
		$DEMO_D$	$DEMO_B$	$DEMO_S$	$DEMO_D$	$DEMO_B$	$DEMO_S$
rs1333049	30	0	9	0	1	13	0
rs2200733	27	2	8	0	1	7	0
rs2476601	33	1	4	0	4	6	0
rs3135388	39	4	12	0	2	9	0
rs6457620	35	3	7	0	0	12	0
rs17234657	28	3	9	0	1	6	0
Total	192	13	49	0	9	53	0

$DEMO_D$, $DEMO_B$, and $DEMO_S$ are the Demonstration group when anonymized, extracted from the BioVU anonymization, and extracted from the SD anonymization, respectively. Original is the number of significant associations (p 0.05) found in the PheWAS when conducted on pre-anonymized data. Identical is the number of associations which were the same between studies. Lost is the number of associations that were lost in the anonymized study. False is the number of associations that were determined as significant in the new study but were not in the original.
doi:10.1371/journal.pone.0053875.t004

***rs*2200733.** Figure 3 indicates there was an association which was anticipated to have a *p*-value of approximately 1.6. Yet, in the anonymized dataset, the *p*-value for this association was close to 4, indicating a high likelihood of association.

rs2476601. Figure 2 indicates there was an association which was anticipated to have a *p*-value of approximately 0.5. Yet, in the anonymized dataset, the *p*-value for condition 762 was close to 2, indicating a high likelihood of association.

rs3135388. Figure 4 indicates that condition 134 would originally have a lower likelihood of association. However, the anonymization, its *p*-value has sufficiently changed for it to be considered significant. Similarly, in $DEMO_B$, condition 223 originally had a value of approximately 1.6. In the anonymization, however, this value decreased to approximately 0.3, making it far less likely to be labeled significant.

rs6457620. Figure 5 indicates an expected value of at least 0.75 for a number of conditions, including 761, 976, 828, and 127, that, when anonymized, appear to be at or near 0.

rs17234657. Figure 6 indicates an expected value of at least 1.5 for two conditions, 976 (1.5) and 140 (3.6), when anonymized, appear to be at or near 0, completely removing them from significance.

Summary of Findings

In terms of general information retention, $DEMO_S$ always outperformed $DEMO_D$ and $DEMO_B$. This result suggests that the larger the initial set from which the subset is drawn, the more likely it is that deidentification can retain associations that are being sought. In terms of PheWAS, $DEMO_S$ exactly matched the evaluation performed on DEMO the non-anonymized Demonstration cohort. We have also shown that, when compared to the

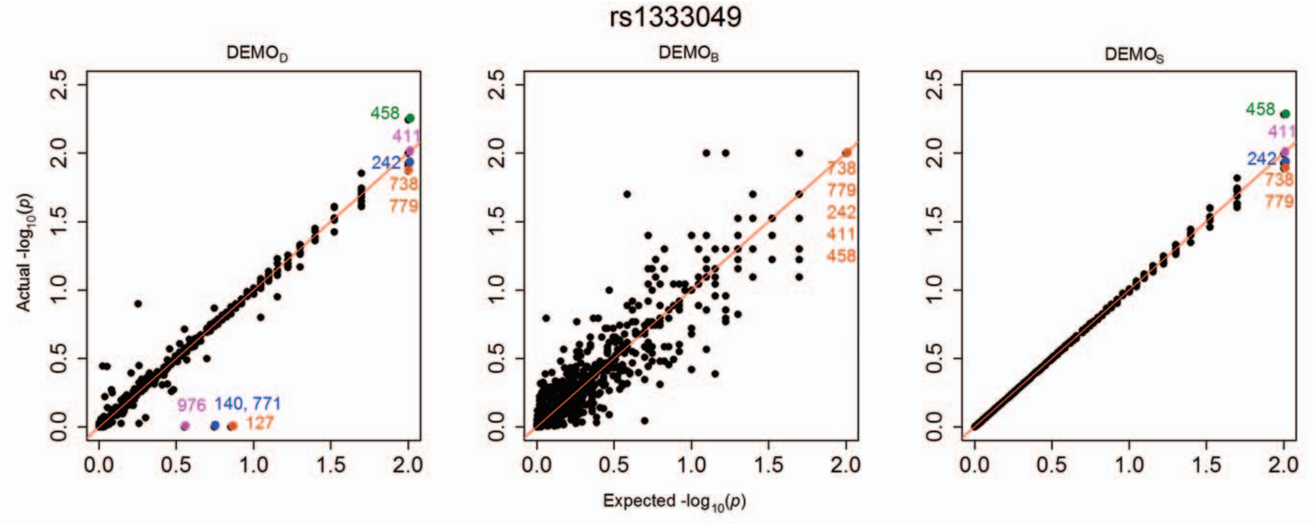


Figure 1. Changes in p-values for associations between clinical conditions and SNP rs1333049 presented as a QQ-plot for left) $DEMO_D$, middle) $DEMO_B$, and right) DEMOS. Descriptions of the annotated conditions in the plots are provided in Table 5.
doi:10.1371/journal.pone.0053875.g001

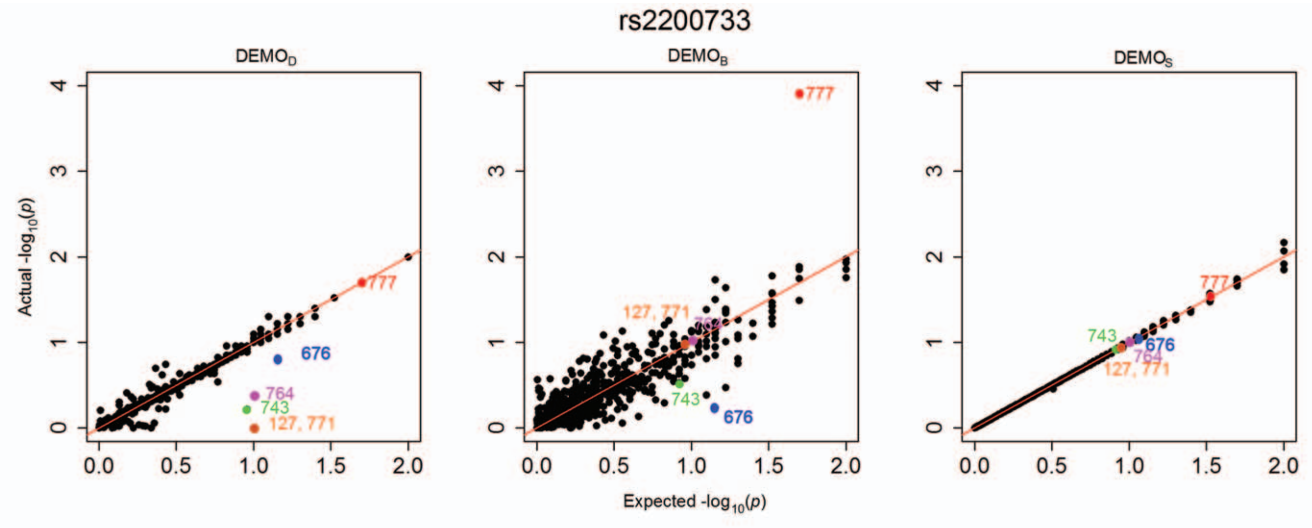


Figure 2. Changes in p-values for associations between clinical conditions and SNP rs2476601 presented as a QQ-plot for left) $DEMO_D$, middle) $DEMO_B$, and right) DEMOS. Descriptions of the annotated conditions in the plots are provided in Table 5.
doi:10.1371/journal.pone.0053875.g002

total amount of information in the original data (i.e., the SD), $DEMO_B$ outperforms $DEMO_D$, but that when anonymized data is compared to its non-anonymized data, the opposite is true in some of our measures. Additionally, $DEMO_B$ has a much greater number of lost and false associations than $DEMO_D$ (or $DEMO_S$).

The relationship between $DEMO_B$ and $DEMO_D$ is not quite as paradoxical as it may appear. In BioVU, when generalizing codes, some records have repeated incidents of related, low-frequency conditions. For example, consider record 49532 in Figure 13a. Notice that in one visit, both codes 401.00– *malignant hypertension* - and 401.01– *benign hypertension* - are present. However, in Figure 13b, the codes have been replaced with ⟨401.00,401.01⟩ (read as: "400.00 and/or 400.01"). Considering just these codes, the DC in the original dataset equals two. However, once these codes are transformed into ⟨400.00,400.01⟩, the result is a single generalized code, which halves the number of diagnoses in the anonymized dataset, yielding a DC of one. While the expectation was that BioVU would yield better results due to its significant increase in size, DEMOs population was selected to satisfy several specific phenotypes. As a result, records in DEMO were much more similar than records within BioVU. Consequently, less generalization was necessary to obtain $DEMO_D$ than $DEMO_B$.

This does not indicate that the anonymization strategy is ineffective. Instead, we have shown that it is important for the data holder to anticipate how the post-anonymization data will be used. If the data is intended to assist in hypothesis validation for a very specific cohort, then use of $DEMO_D$ may be sufficient. If, however, the data is intended to support hypothesis generation, then the use of $DEMO_B$ may be preferable. Regardless of the end use, however, $DEMO_S$ provides the most benefit to either task. As an additional merit to the use of $DEMO_S$, any further cohort that is drawn from SD-Anon is subject to the same protection, which

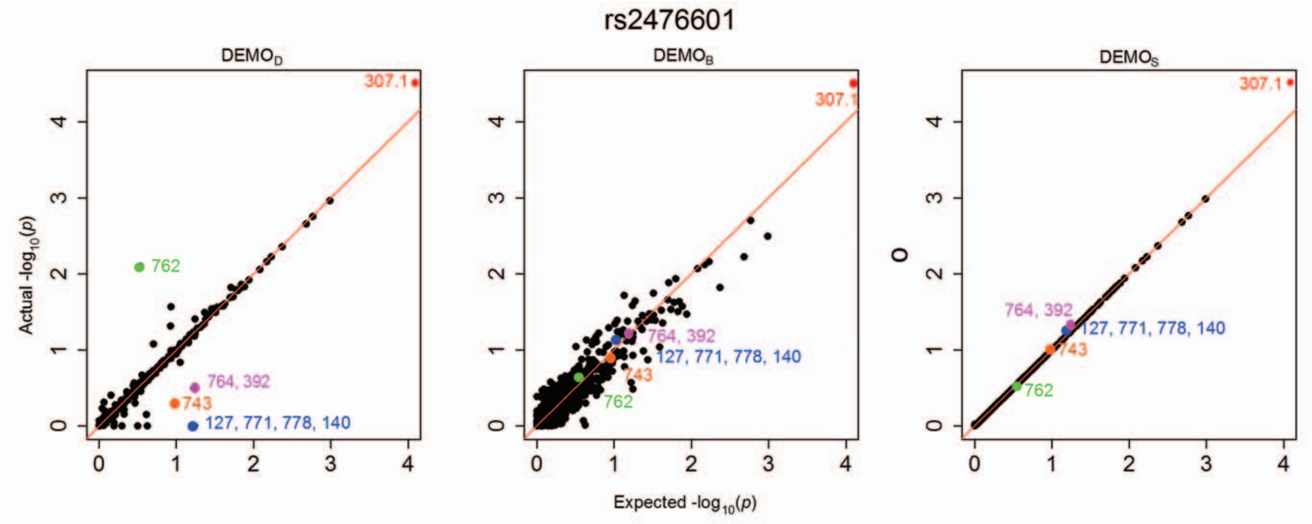


Figure 3. Changes in p-values for associations between clinical conditions and SNP rs2200733 presented as a QQ-plot for left) $DEMO_D$, middle) $DEMO_B$, and right) DEMOS. Descriptions of the annotated conditions in the plots are provided in Table 5.
doi:10.1371/journal.pone.0053875.g003

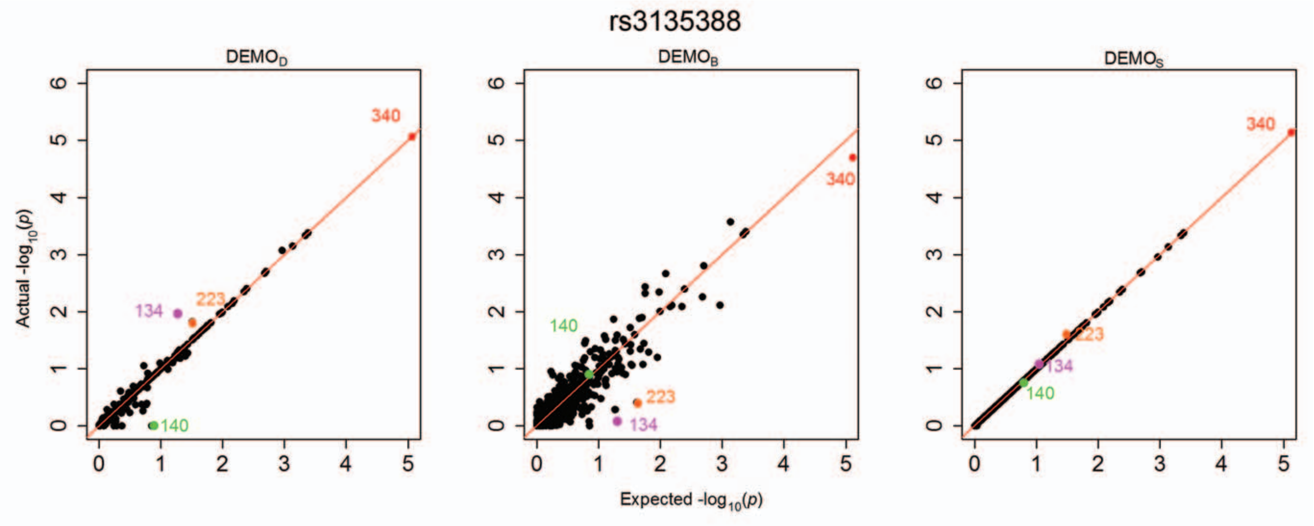


Figure 4. Changes in p-values for associations between clinical conditions and SNP rs3135388 presented as a QQ-plot for left) $DEMO_D$, middle) $DEMO_B$, and right) $DEMO_S$. Descriptions of the annotated conditions in the plots are provided in Table 5.

doi:10.1371/journal.pone.0053875.g004

means that if a user is grouped into two different cohorts, the exact same information will be revealed about this user to both groups. Separate anonymizations for data selections may not hold this property. Further research is necessary to determine what privacy claims, if any, may hold over repeated anonymizations of separate cohorts.

Discussion

The anonymization method proposed in this paper is a significant improvement over prior approaches. It enables healthcare institutions to account for adversaries of varying strengths. Moreover, our analysis illustrates that when an adversary is aware that a patient was a member of the hospitals general population (as opposed to as a specific cohort), the utility of the anonymized cohort is virtually equivalent to the pre-anonymized results. These results suggest that when reasonable adversarial models are applied in the context of large medical facilities, phenome-wide annotation of clinical populations could be anonymized, allowing public sharing of such data, without sacrificing research findings. Adoption of such a principled approach could enable much greater utility of extant research data sets such as currently stored within dbGaP.

This finding indicates that rather than selecting the smallest possible subset of data that may need to be released, there is significant value in anonymizing the entire body of data at an institution. Release of even subsets of these data provide far more data to subsequent researchers, while still maintaining a high standard of privacy for the patients reflected in these data. This implies that institutions may be able to publicly release large, dense datasets for various research purposes with provable privacy guarantees.

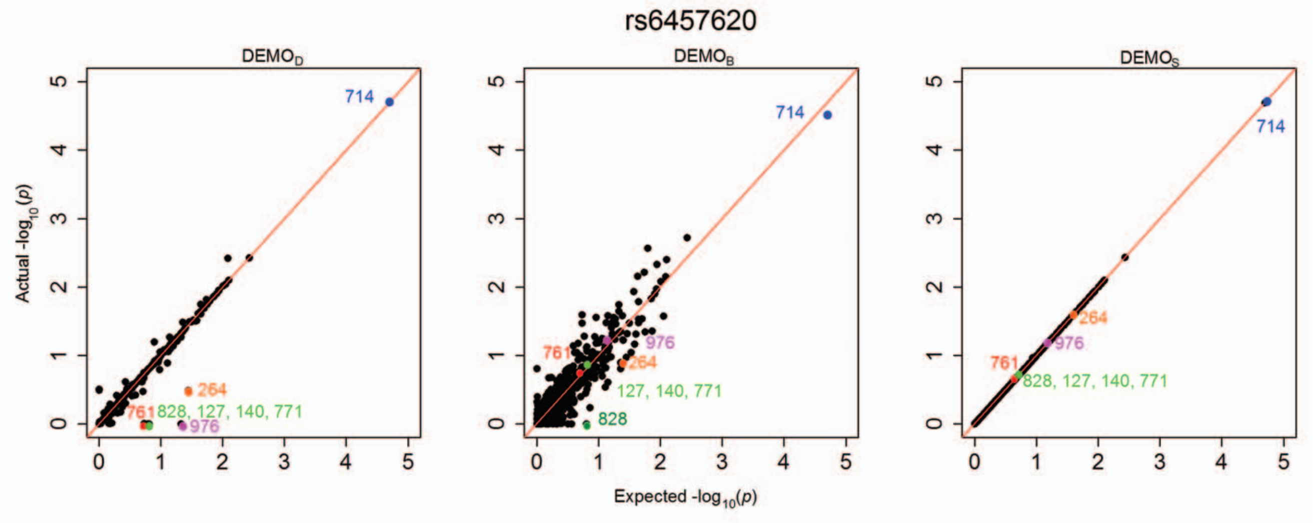


Figure 5. Changes in p-values for associations between clinical conditions and SNP rs6457620 presented as a QQ-plot for left) $DEMO_D$, middle) $DEMO_B$, and right) $DEMO_S$. Descriptions of the annotated conditions in the plots are provided in Table 5.

doi:10.1371/journal.pone.0053875.g005

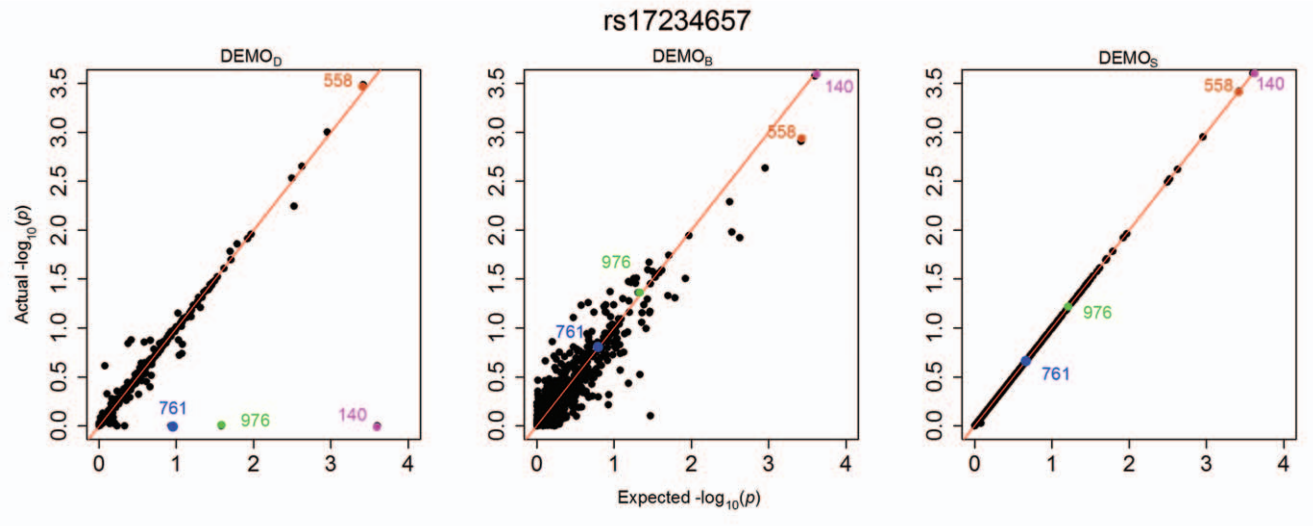


Figure 6. Changes in p-values for associations between clinical conditions and SNP rs17234657 presented as a QQ-plot for left) $DEMO_D$, middle) $DEMO_B$, and right) $DEMO_S$. Descriptions of the annotated conditions in the plots are provided in Table 5.
doi:10.1371/journal.pone.0053875.g006

Our study does include limitations, which can serve as guidelines for future research. First, from a technical perspective, the clinical code generalization strategy employed by the anonymization algorithm does not guarantee minimizing the amount of information loss incurred by the anonymization. For example, our algorithm chooses only one potential generalization from among many options. We chose this method of generalization because it is known to be computationally intractable for such

Table 5. Descriptions for the condition codes presented in the QQ-plots.

Code	Condition
127	Other intestinal helminthiases
134	Other infestation
140	Malignant neoplasm of lip
223	Benign neoplasm of kidney and other urinary organs
242	Thyrotoxicosis with or without goiter
264	Vitamin A deficiency
307.1	Eating disorders
392	Rheumatic chorea
411	Ischemic heart disease
458	Hypotension
558	Other and unspecified noninfectious gastroenteritis and colitis
676	Other disorders of the breast associated with childbirth and disorders of lactation
714	Rheumatoid arthritis and other inflammatory polyarthropathies
738	Other acquired musculoskeletal deformity
743	Congenital anomalies of eye
761	Fetus or newborn affected by maternal complications of pregnancy
762	Fetus or newborn affected by complications of placenta, cord, and membranes
764	Slow fetal growth and fetal malnutrition
771	Infections specific to the perinatal period
777	Perinatal disorders of digestive system
778	Conditions involving the integument and temperature regulation of fetus and newborn
779	Other and ill-defined conditions originating in the perinatal period
828	Multiple fractures involving both lower limbs, lower with upper limb, and lower limb(s) with rib(s) and sternum
976	Poisoning by agents primarily affecting skin and mucous membrane, ophthalmological, otorhinolaryngological, and dental drugs

doi:10.1371/journal.pone.0053875.t005

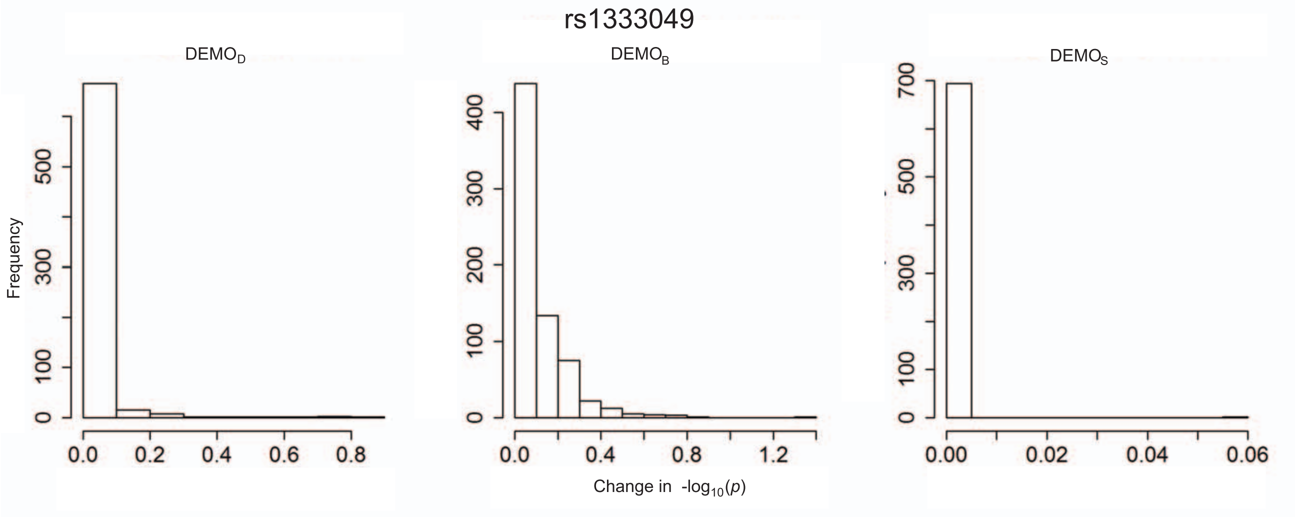


Figure 7. Distribution of p-value changes for associations between clinical conditions and SNP rs1333049 for left) $DEMO_D$, middle) $DEMO_B$, and right) DEMOS.
doi:10.1371/journal.pone.0053875.g007

data to be produced in a manner that minimizes information loss [46,47]. Nonetheless, we suspect that additional heuristics may be devised which can provide improvements or alternatives to our results.

Second, from an implementation perspective, it is important to note that certain healthcare institutions may be more likely to be attacked than others. As a consequence, we recommend that a healthcare institution assess the anticipated capabilities of their data recipients before adopting an anonymization strategy such as the one presented in this manuscript. For instance, healthcare institutions may choose a weaker adversarial model if they anticipate that the data recipient is a credentialed scientific investigator as opposed to an unknown individual in the general public [33]. Similarly, healthcare institutions manage vastly different volumes of data. While we have shown here that the utility and privacy impact on data of this magnitude are beneficial, further work is needed to determine what volume of data is necessary to obtain similar findings.

Methods

Study Overview

A summary of the datasets analyzed in this study are reported in Table 1, while their relationships are visually depicted in Figure 14.

The first dataset corresponds to a HIPAA de-identified (see Methods) version of all VUMC patient records, called the Synthetic Derivative (SD) [48], which contains 1,366,786 records. The second dataset corresponds to a subset of this resource for which the VUMC collected de-identified DNA samples, called BioVU ($n = 104{,}904$). The third dataset, referred to as DEMO ($n = 5{,}944$), is a subset of BioVU records that were previously analyzed to demonstrate the feasibility of phenome-wide association studies (PheWAS), using specific genotypes, via information

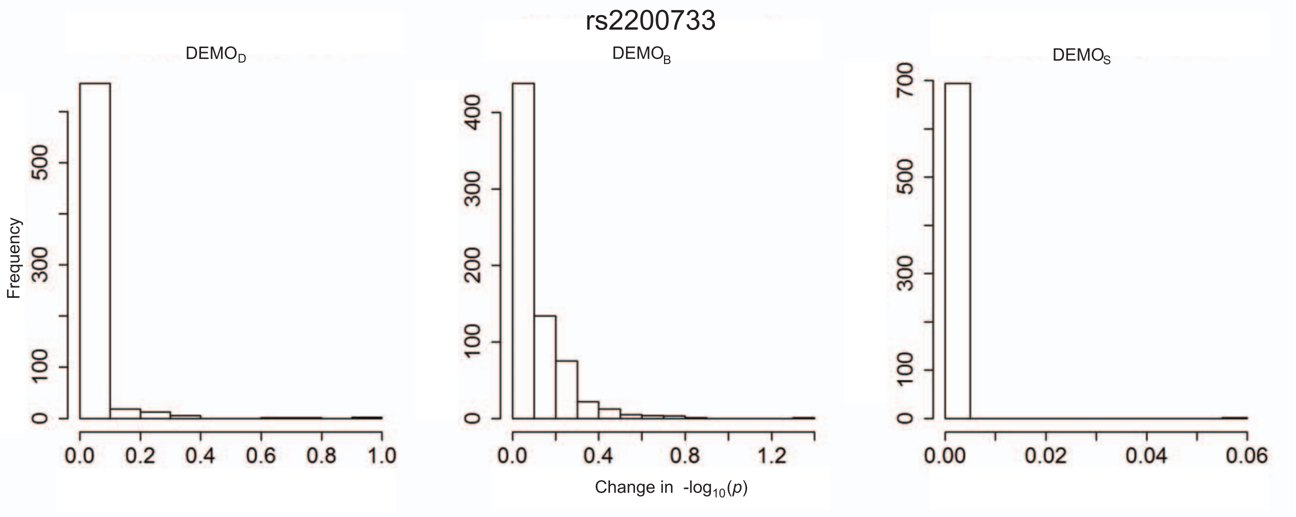


Figure 8. Distribution of p-value changes for associations between clinical conditions and SNP rs1333049 for left) $DEMO_D$, middle) $DEMO_B$, and right) DEMOS.
doi:10.1371/journal.pone.0053875.g008

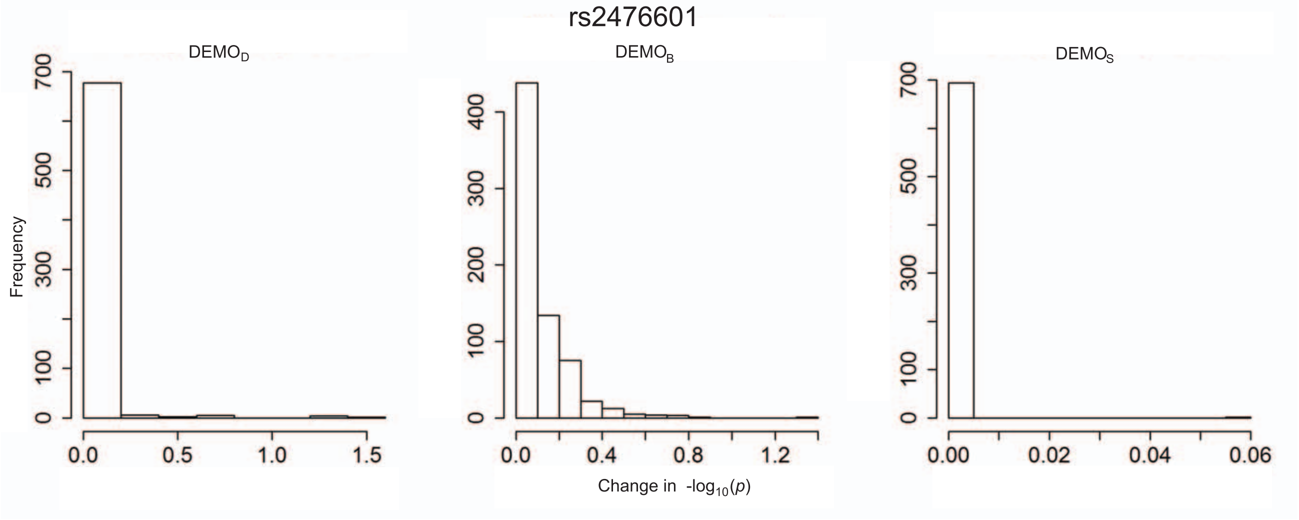


Figure 9. Distribution of p-value changes for associations between clinical conditions and SNP rs2476601 for left) $DEMO_D$, middle) $DEMO_B$, and right) $DEMO_S$.
doi:10.1371/journal.pone.0053875.g009

in existing EMRs [43]. In this dataset, each patient record is divided into a series of visits made to VUMC-affiliated healthcare providers. Each visit is characterized by the clinical activities that transpired, including diagnoses made, medications prescribed, and laboratory test results. An example of the structure of such records is depicted in Figure 13a. For this study, we anonymize the ICD-9 billing codes in the records, but we remark that our method is sufficiently general to apply to any standardized vocabulary of clinical events.

To model how cohorts are disseminated for validation and reuse, we developed a novel anonymization strategy that enables a subset of the SD to be shared for research purposes. In short, this strategy yields a patient record composed of diagnoses across all their visits. The information is anonymized, such that for any set of disclosed ICD-9 codes obtained at any one visit, there are at least k records in the anonymized resource with this combination of codes across all visits. For illustration, Figure 13b depicts a fictional example of anonymized records, with k set to 2. In our evaluation, we set k to 5, which is a level of protection commonly applied in practice [41].

There are several ways in which data can be anonymized to account for the knowledge of the recipient. Figure 14 depicts the various strategies. We begin with all data contained within the SD, from which we select two subsets. The first subset is BioVU and the second is DEMO. Each of the three datasets is then anonymized to create SD-Anon, BioVU-Anon, and $DEMO_D$, respectively. To examine the effect that each of the anonymizations have on subsequent analysis, we then extract the records which are in DEMO from SD-Anon and BioVU-Anon, creating $DEMO_S$ and $DEMO_B$, respectively. Note, $DEMO_S$, $DEMO_B$, and $DEMO_D$ each contain the same records, but the specific

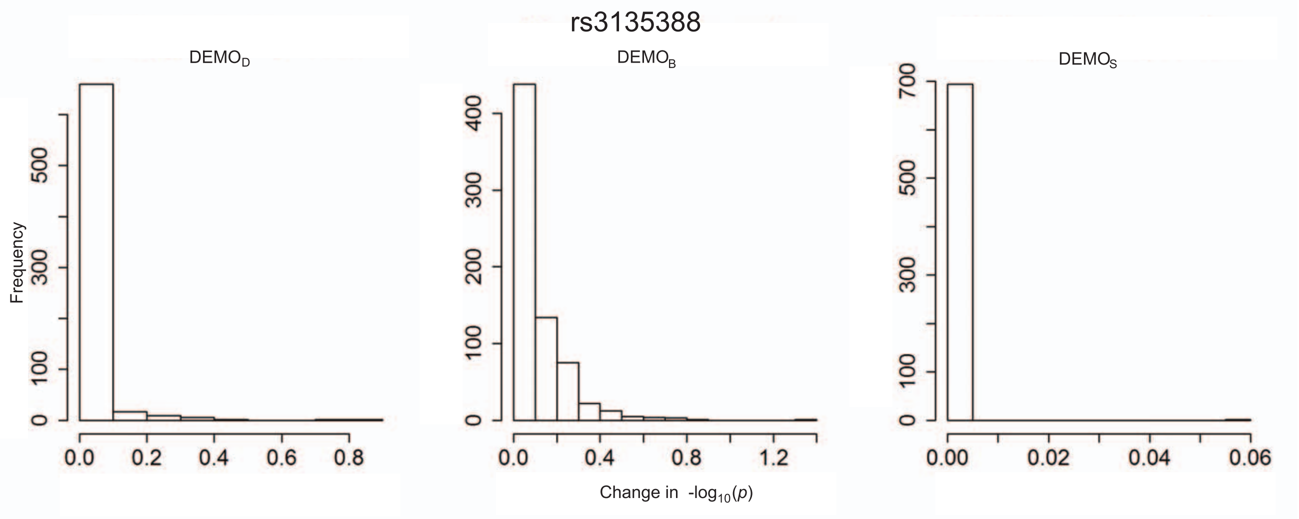


Figure 10. Distribution of p-value changes for associations between clinical conditions and SNP rs3135388 for left) $DEMO_D$, middle) $DEMO_B$, and right) $DEMO_S$.
doi:10.1371/journal.pone.0053875.g010

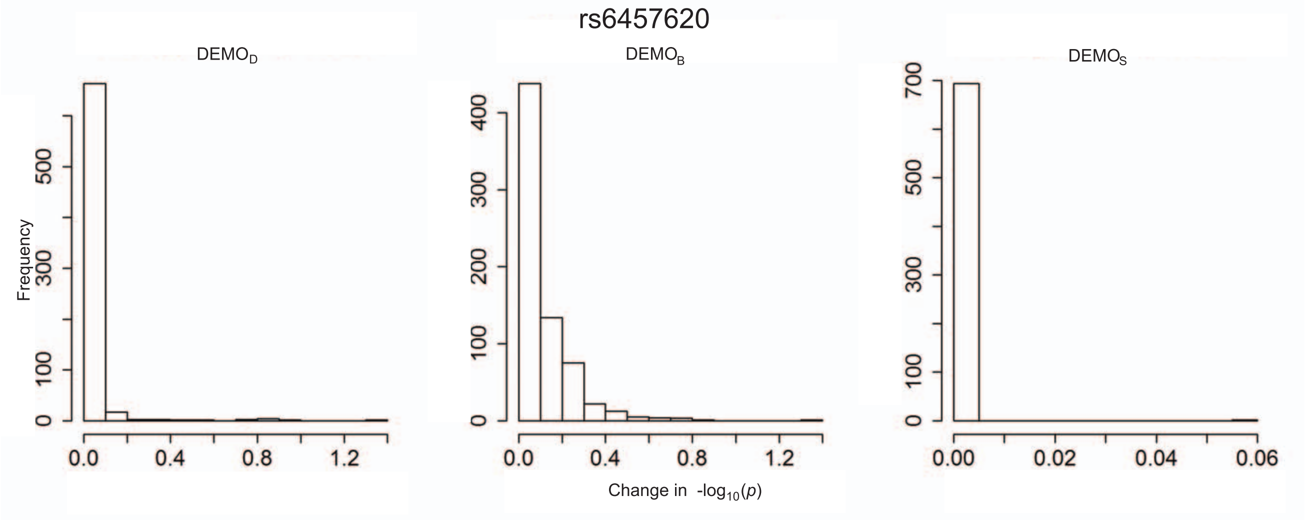


Figure 11. Distribution of p-value changes for associations between clinical conditions and SNP rs6457620 for left) $DEMO_D$, middle) $DEMO_B$, and right) $DEMO_S$.
doi:10.1371/journal.pone.0053875.g011

clinical codes within those records are different due to the anonymization process.

Privacy Models and Methods

There are a variety of computational models that have been proposed for protecting biomedical data. Most recently, randomization strategies, notably those based on differential privacy [49,50], have been suggested. These approaches perturb records through a controlled, but random, process (e.g., addition of codes not originally diagnosed). Such a framework provides strong proofs of privacy, but may be insufficient to support new studies at varying levels of granularity. Moreover, if care is not taken in its design, this strategy could lead to strange data representations (e.g., juvenile patients diagnosed with Alzheimers disease), and in the co-occurrence with a chance rare genetic event (e.g., a rare functional mutation in an exon), could lead to an erroneous association. Thus, we focused on data protection models that remain true to the underlying data. To do so, we adopted a variation of the k-anonymization principle [51], which states that any combination of potential identifiers in the resultant dataset must match at least k records. This principle has been applied to various types of patient-level data, such as demographics [41], as well as clinical codes [40]. To achieve privacy in our setting, we enforced a constraint which states that, for each visit of a patient, there are at least k patients who have the same set of diagnosis codes from some visit in the resulting dataset. This model allows us to represent an adversary with a moderate, but manageable, level of knowledge regarding patient information released by the institution.

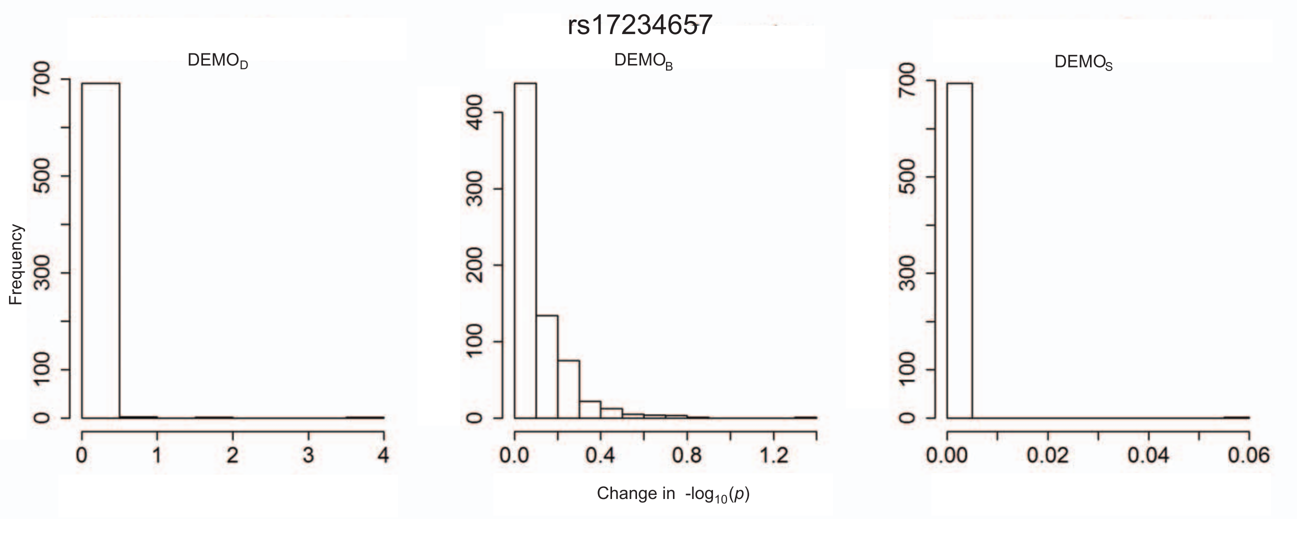


Figure 12. Distribution of p-value changes for associations between clinical conditions and SNP rs17234657 for left) $DEMO_D$, middle) $DEMO_B$, and right) $DEMO_S$.
doi:10.1371/journal.pone.0053875.g012

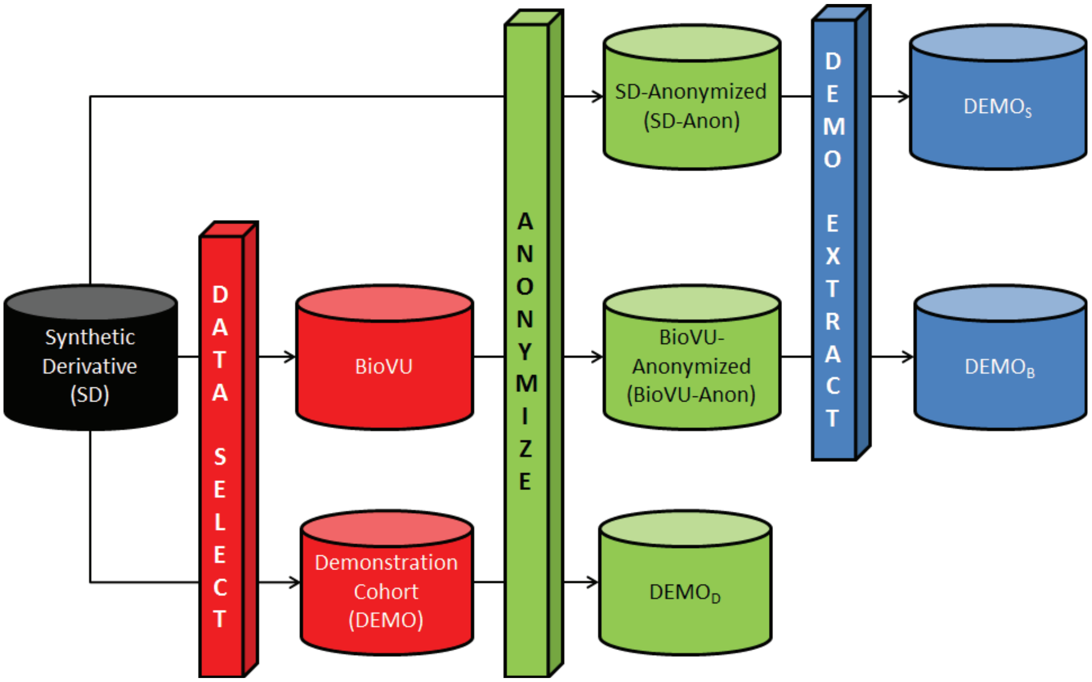


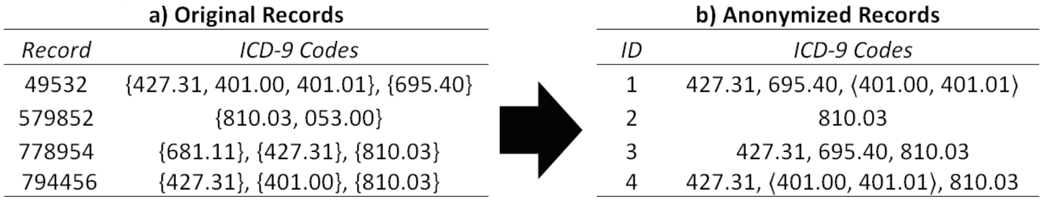
Figure 13. A fictional example of patient-specific records of diagnosis codes in the a) original resource and b) corresponding 2-anonymized result. The braces ("{ ldots }") demarcate the set of diagnoses received in a visit to a healthcare provider, while the brackets ("⟨ . . . ⟩") denote codes that have been generalized in accordance with the anonymization discussed herein.
doi:10.1371/journal.pone.0053875.g013

Health Data De-identification According to Federal Regulation and Residual Risks

Our goal is to enable biomedical analysis with patient-level records while thwarting re-identification attempts. Returning to Figure 13, the SD is de-identified according to the Privacy Rule of the Health Information Portability and Accountability Act of 1996 (HIPAA). This was accomplished by removing eighteen specific features of the data, including direct identifiers (e.g., patient names and residential address), quasi-identifiers (e.g., dates of birth, death, and healthcare provider visits), and specific identification numbers or codes (e.g., medical device identification numbers). Despite the removal of such information, many records may be uniquely distinguishable based on the combination of their diagnosis codes. [31] For instance, imagine that an attacker knows a patient, say "Alice" (49532), was assigned billing codes 427.31 *atrial fibrillation* and 401.00 *hypertension* in a hospital visit. Then, according to the depiction to the left of Figure 13, Alice will be uniquely identified in the original dataset. This means that the attacker learns Alice was additionally diagnosed with code 695.40 *systemic lupus* (as well as any other codes or DNA sequences in the released dataset). However, in the anonymized version of the table to right of Figure 13, an attacker would be unable to determine whether this patient is Record 1 or Record 4.

Clinical Concept Anonymization Process

To satisfy anonymization requirements, we invoke a system of code generalization. The generalization replaces a specific ICD-9 code with a group of codes which are semantically similar. For example, to successfully anonymize a dataset, we may need to generalize the code 810.01 *closed fracture of sternal end of clavicle* to the code "810.00 and/or 810.01" *closed fracture of clavicle, sternal and/or unspecified*. However, this generalization introduces the need for guidelines on what codes may acceptably be generalized together, which are called utility constraints. For instance, generalizing



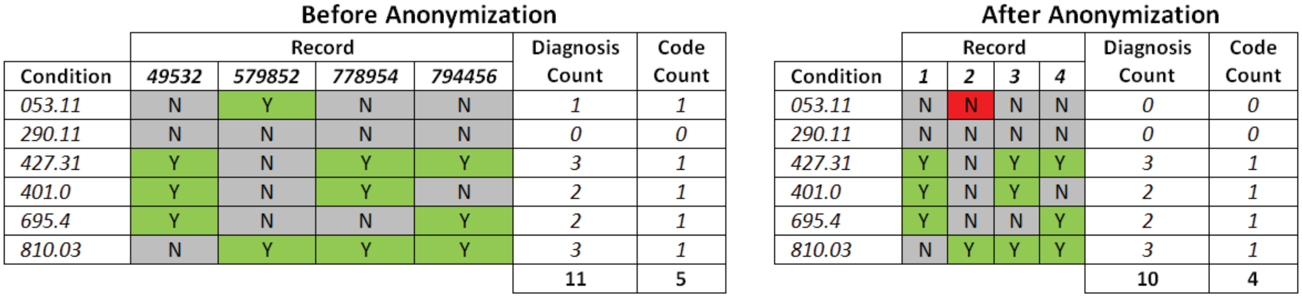
a) Original Records

Record	ICD-9 Codes
49532	{427.31, 401.00, 401.01}, {695.40}
579852	{810.03, 053.00}
778954	{681.11}, {427.31}, {810.03}
794456	{427.31}, {401.00}, {810.03}

b) Anonymized Records

ID	ICD-9 Codes
1	427.31, 695.40, ⟨401.00, 401.01⟩
2	810.03
3	427.31, 695.40, 810.03
4	427.31, ⟨401.00, 401.01⟩, 810.03

Figure 14. Datasets used for comparison of anonymization strategies. The DATA SELECT process is an extraction of some records of the SD into a smaller, specific dataset, such as BioVU or a demonstration cohort. The ANONYMIZE process is the anonymization algorithm described in this manuscript. The DEMO EXTRACT process selects the remaining records associated with the Demonstration cohort from a larger, anonymized dataset. The resultant datasets are as follows: anonymized version of the Synthetic Derivative (SD-Anon); anonymized version of BioVU (BioVU-Anon); SD-Anon, from which the demonstration group is extracted ($DEMO_S$); BioVU-Anon, from which the demonstration group is extracted ($DEMO_B$); and the anonymized version of the demonstration cohort ($DEMO_D$). $DEMO_S$, $DEMO_B$, and $DEMO_D$ each represent different anonymizations of the Demonstration group.
doi:10.1371/journal.pone.0053875.g014



Before Anonymization

	Record				Diagnosis	Code
Condition	49532	579852	778954	794456	Count	Count
053.11	N	Y	N	N	1	1
290.11	N	N	N	N	0	0
427.31	Y	N	Y	Y	3	1
401.0	Y	N	Y	N	2	1
695.4	Y	N	N	Y	2	1
810.03	N	Y	Y	Y	3	1
					11	5

After Anonymization

	Record				Diagnosis	Code
Condition	1	2	3	4	Count	Count
053.11	N	N	N	N	0	0
290.11	N	N	N	N	0	0
427.31	Y	N	Y	Y	3	1
401.0	Y	N	Y	N	2	1
695.4	Y	N	N	Y	2	1
810.03	N	Y	Y	Y	3	1
					10	4

Figure 15. An illustration of the computation for Code Count and Diagnosis Count. The table to the left illustrates if a record contained a condition (green "Y") or did not (grey "N") in the original dataset before anonymization. The table to the right illustrates if a record lost a diagnosis (red "N"). Notice that the second record lost one diagnosis, which resulted in both the Code Count and Diagnosis Count to be lowered by a score of 1.
doi:10.1371/journal.pone.0053875.g015

810.01 to "940 and/or 810.01" *burns* or *closed sternal fracture of the clavicle* would have introduced an entirely new condition "burns" instead of simply obscuring the specific region of the clavicle which had been broken. To prevent such occurrences, we use a hierarchy that defines what generalizations are allowed. For this work, we use the hierarchy described in [43], which is a complete mapping of all ICD-9 codes developed for clinical phenotype-genotype association studies.

For the anonymization, our approach generalizes ICD-9 codes with frequency (i.e., the number of records that contain the code in one or more visits) below a threshold *k* within the dataset together. First, we place each code in a bin corresponding to its frequency. For instance, all codes assigned to only one patient are stored in the first bin. For each bin with value less than k, the process generalizes the codes within that bin as permitted by the utility constraints. Next, the support is calculated for the new, generalized code, which is moved into the appropriate bin. On each subsequent iteration, we group adjacent bins together (i.e., bins one and two are grouped together, bins three and four are grouped together) until all bins representing frequency less than *k* have been grouped together. Note that by merging adjacent bins (such as one and two), the result does not necessarily get moved into bin three. Instead, the frequency is recalculated for all patients who would have the new, generalized code. After this point, any codes remaining with frequency less than *k* are suppressed.

At this point, the dataset satisfies the *k*-anonymization requirement and can be shared. However, assuming that the institution holding the data does not wish to release the entire anonymized dataset, this is the point at which subsets may be drawn from the data. For example, suppose that external researchers were interested in patients who had ischemic heart disease. Records of specific interest could be extracted from the anonymized data and then released to these researchers.

Computation of Diagnosis Count and Code Count

In Figure 15, we show an example computation of Diagnosis Count (DC) and Code Count (CC). In the left part of the figure, we show a non-anonymized example data set containing six conditions, 053.11, 290.11, 427.31, 401.0, 695.4, and 810.03, and four records - A, B, C, and D. We represent the datum that Record B was diagnosed with condition 053.11 in some visit as a "Y" (also shown as a green cell highlight). The absence of this diagnosis is represented as an "N" (also shown as a grey cell highlight), as shown in the cell represented by condition 053.11 and Record A.

As shown, Diagnosis Count is simply the number of times that a particular diagnosis is assigned across all patients in the set. Since Record B is the only one that contains the diagnosis 053.11, its Diagnosis Count is 1. Alternatively, Code Count is a count of the number of codes that have positive diagnoses in that data set. As shown in Figure 15, code 290.11 has no records with that diagnosis. As such, its Code Count is 0. Since each other code has at least one record with a positive diagnosis, each other count is 1, giving the data set a Code Count of 5.

On the right side of Figure 15, we show a possible change in our measures following anonymization. In this instance, the anonymization has suppressed Record 2's diagnosis of 053.11 (now represented as "N" in a red cell highlight). Because this decreased the number of diagnoses in the data set, the Diagnosis Count decreased by 1. However, since this was also the only diagnosis of condition 053.11, that code is no longer represented in the data; thus, the Code Count also decreased by 1.

Acknowledgments

The authors wish to thank Melissa Basford, Sunny Wang, and Jay Cowan for assistance in gathering and organizing the data used in this study.

Author Contributions

Conceived and designed the experiments: RDH BAM GL JCD. Performed the experiments: JDC RDH. Analyzed the data: RDH BAM GL JCD. Contributed reagents/materials/analysis tools: JCD DMR RDH. Wrote the paper: RDH GL BAM DMR JCD JH.

References

1. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, et al. (2005) Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Affairs 24: 1103–1117.
2. Kawamoto K, Houlihan C, Balas E, Lobach D (2005) Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj 330: 765.
3. Buntin M, Burke M, Hoaglin M, Blumenthal D (2011) The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health Affairs 30: 464–471.
4. Blumenthal D (2009) Stimulating the adoption of health information technology. New England Journal of Medicine 360: 1477–1479.
5. Safran C, Bloomrosen M, Hammond W, Labkoff S, Markel-Fox S, et al. (2007) Toward a national framework for the secondary use of health data: an american medical informatics association white paper. Journal of the American Medical Informatics Association 14: 1–9.
6. Piwowar H, Becich M, Bilofsky H, Crowley R (2008) Towards a data sharing culture: recommendations for leadership from academic health centers. PLoS medicine 5: e183.

7. Lazarus R, Klompas M, Campion F, McNabb S, Hou X, et al. (2009) Electronic support for public health: validated case finding and reporting for notifiable diseases using electronic medical data. Journal of the American Medical Informatics Association 16: 18–24.
8. Elkin P, Froehling D, Wahner-Roedler D, Brown S, Bailey K (2012) Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Annals of Internal Medicine 156: 11–18.
9. Weiner M, Embi P (2009) Toward reuse of clinical data for research and quality improvement: The end of the beginning? Annals of internal medicine 151: 359–360.
10. Jensen P, Jensen L, Brunak S (2012) Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics.
11. Fadly A, Rance B, Lucas N, Mead C, Chatellier G, et al. (2011) Integrating clinical research with the healthcare enterprise: from the re-use project to the ehr4cr platform. Journal of Biomedical Informatics.
12. Green ED, Guyer MS and National Human Genome Research Institute (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470: 204–213.
13. Kho A, Pacheco J, Peissig P, Rasmussen L, Newton K, et al. (2011) Electronic medical records for genetic research: results of the emerge consortium. Sci Transl Med 3: 79re1.
14. Kohane I (2011) Using electronic health records to drive discovery in disease genomics. Nature Reviews Genetics 12: 417–428.
15. McCarty C, Chisholm R, Chute C, Kullo I, Jarvik G, et al. (2011) The emerge network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics 4: 13.
16. Denny J, Ritchie M, Crawford D, Schildcrout J, Ramirez A, et al. (2010) Identification of genomic predictors of atrioventricular conductionclinical perspective using electronic medical records as a tool for genome science. Circulation 122: 2016–2021.
17. Crosslin D, McDavid A, Weston N, Nelson S, Zheng X, et al. (2011) Genetic variants associated with the white blood cell count in 13,923 subjects in the emerge network. Human genetics : 1–14.
18. Kullo I, Ding K, Jouni H, Smith C, Chute C (2010) A genome-wide association study of red blood cell traits using the electronic medical record. PLoS One 5: e13011.
19. Denny J, Crawford D, Ritchie M, Bielinski S, Basford M, et al. (2011) Variants near¡ i¿ foxe1¡/i¿ are associated with hypothyroidism and other thyroid conditions: Using electronic medical records for genome-and phenome-wide studies. The American Journal of Human Genetics 89: 529–542.
20. Delaney J, Ramirez A, Bowton E, Pulley J, Basford M, et al. (2011) Predicting clopidogrel response using dna samples linked to an electronic health record. Clinical Pharmacology & Therapeutics.
21. Ramirez A, Shi Y, Schildcrout J, Delaney J, Xu H, et al. (2012) Predicting warfarin dosage in european-americans and african-americans using dna samples linked to an electronic health record. Pharmacogenomics : 1–12.
22. Mailman M, Feolo M, Jin Y, Kimura M, Tryka K, et al. (2007) The ncbi dbgap database of genotypes and phenotypes. Nature genetics 39: 1181–1186.
23. Walker L, Starks H, West K, Fullerton S (2011) dbgap data access requests: A call for greater transparency. Science Translational Medicine 3: 113cm34–113cm34.
24. Stone M, Redsell S, Ling J, Hay A (2005) Sharing patient data: competing demands of privacy, trust and research in primary care. The British Journal of General Practice 55: 783.
25. Kalra D, Gertz R, Singleton P, Inskip H (2006) Confidentiality and consent in medical research: Confidentiality of personal health information used for research. BMJ: British Medical Journal 333: 196.
26. Benitez K, Malin B (2010) Evaluating re-identification risks with respect to the hipaa privacy rule. Journal of the American Medical Informatics Association 17: 169–177.
27. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M (2006) Evaluating common de-identification heuristics for personal health information. Journal of Medical Internet Research 8.
28. El Emam K, Dankar F, Vaillancourt R, Roffey T, Lysyk M (2009) Evaluating the risk of re-identification of patients from hospital prescription records. The Canadian Journal of Hospital Pharmacy 62: 307.
29. Li F, Zou X, Liu P, Chen J (2011) New threats to health data privacy. BMC Bioinformatics 12: S7.
30. Malin B, Sweeney L (2004) How (not) to protect genomic data privacy in a distributed net-work: using trail re-identification to evaluate and design anonymity protection systems. Journal of Biomedical Informatics 37: 179–192.
31. Loukides G, Denny J, Malin B (2010) The disclosure of diagnosis codes can breach research participants' privacy. Journal of the American Medical Informatics Association 17: 322–327.
32. of Health UD, Services H (2011) Advanced notice of proposed rulemaking: human subjects and reducing burden, delay, and ambiguity for investigators. Federal Register 76: 44512–44531.
33. Malin B, Loukides G, Benitez K, Clayton E (2011) Identifiability in biobanks: models, measures, and mitigation strategies. Human genetics : 1–10.
34. Anderson N, Edwards K (2010) Building a chain of trust: using policy and practice to enhance trustworthy clinical data discovery and sharing. In: Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies. ACM, 15–20.
35. Malin B, Karp D, Scheuermann R (2010) Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. Journal of investigative medicine: the official publication of the American Federation for Clinical Research 58: 11.
36. Roche P, Annas G (2001) Protecting genetic privacy. Nature Reviews Genetics 2: 392–396.
37. Wylie J, Mineau G (2003) Biomedical databases: protecting privacy and promoting research. TRENDS in Biotechnology 21: 113–116.
38. El Emam K, Dankar F, Issa R, Jonker E, Amyot D, et al. (2009) A globally optimal k-anonymity method for the de-identification of health data. Journal of the American Medical Informatics Association 16: 670–682.
39. Ohno-Machado L, Bafna V, Boxwala A, Chapman B, Chapman W, et al. (2012) idash: integrating data for analysis, anonymization, and sharing. Journal of the American Medical Informatics Association 19: 196–201.
40. Loukides G, Gkoulalas-Divanis A, Malin B (2010) Anonymization of electronic medical records for validating genome-wide association studies. Proceedings of the National Academy of Sciences 107: 7898.
41. El Emam K, Dankar F (2008) Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association 15: 627–637.
42. Gionis A, Mazza A, Tassa T (2008) k-anonymization revisited. In: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. IEEE, 744–753.
43. Denny J, Ritchie M, Basford M, Pulley J, Bastarache L, et al. (2010) Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics 26: 1205–1210.
44. Avery C, He Q, North K, Ambite J, Boerwinkle E, et al. (2011) A phenomics-based strategy identifies loci on apoc1, brap, and plcg1 associated with metabolic syndrome phenotype domains. PLoS genetics 7: e1002322.
45. Pendergrass S, Brown-Gentry K, Dudek S, Torstenson E, Ambite J, et al. (2011) The use of phenome-wide association studies (phewas) for exploration of novel genotype-phenotype relation-ships and pleiotropy discovery. Genetic epidemiology.
46. Loukides G, Gkoulalas-Divanis A, Malin B (2011) Coat: Constraint-based anonymization of trans-actions. Knowledge and Information Systems 28: 251–282.
47. Terrovitis M, Mamoulis N, Kalnis P (2008) Privacy-preserving anonymization of set-valued data. Proceedings of the VLDB Endowment 1: 115–125.
48. Roden D, Pulley J, Basford M, Bernard G, Clayton E, et al. (2008) Development of a large-scale de-identified dna biobank to enable personalized medicine. Clinical Pharmacology & Therapeutics 84: 362–369.
49. Vinterbo S, Sarwate A, Boxwala A (2012) Protecting count queries in study design. Journal of the American Medical Informatics Association 19: 750–757.
50. Wasserman L, Zhou S (2010) A statistical framework for differential privacy. Journal of the American Statistical Association 105: 375–389.
51. Sweeney L (2002) k-anonymity: A model for protecting privacy. International Journal of Uncertainty Fuzziness and Knowledge Based Systems 10: 557–570.