Privacy in Health Disparity Research

Heng Xu, PhD, and Nan Zhang, PhD

Background: Most research on identifying and understanding health disparities focused on measurement strategies and analytics design over administrative and survey data. What has received less attention, however, is the complex interplay between privacy concerns and the needs of health disparity research.

Objectives: The main objective of the paper is to illustrate the fundamental challenges facing the reconciliation of privacy needs and health disparity research.

Research Design: We investigated whether privacy disparity exists over raw, unprocessed, public health data, and then examined the effect of a state-of-the-practice data anonymization process and a state-of-the-art privacy-preserving data sharing technique on both privacy disparity and the efficacy of health disparity research.

Results: We demonstrate that the essential elements for health disparity research, for example, the integration of demographics attributes such as race in public health datasets, could indeed lead to privacy disparity if not carefully treated.

Discussions and Policy Implications: The findings in this paper indicate that regulators must carefully study the privacy implications before mandating the inclusion of certain attributes like race and geolocation in a released dataset. For example, the Affordable Care Act requires all federally supported public health programs to collect data on race, ethnicity, sex, geographic location, etc., to the extent feasible. Although doing so could help identify health disparities, if the collected data are directly released without proper sanitization, the released data may introduce privacy disparities, causing further harm to underserved populations.

Key Words: health disparity, privacy, secondary data analysis

Although technological advances have improved the well-being of our society, socioeconomically disadvantaged, ethnic minorities, and other underserved populations continue to experience a disproportionate share of adverse health outcomes.1–3 To rectify this health disparity,4–6 it is critical for the research and medical communities to leverage Big Data technology for understanding the causes of health disparities and guiding interventions.7 For example, recent studies on addressing the limited availability of social determinants (eg, race, ethnicity)8 in administrative/survey data7,9 include the statistical inference of race/ethnicity from surname and location,10 demonstrating the power of data analytics on health disparity research.

What has received less attention is the complex interplay between health disparity research and the privacy implications of collecting health care data containing sensitive diagnosis and demographics information. Although precautions like removing personally identifiable attributes are often taken, privacy leakage happens even after such removals.11,12 This raises the question of how to study health disparities while protecting the privacy of individuals involved.

Privacy concerns over health care data have long been recognized: a meta-analysis13 identified privacy as a key concern brought by Big Data technologies in health care. Numerous solutions have been developed, informed consent is popular in practice, yet not as effective when the data are subject to secondary use,14 like health disparity research. Technically, there have been extensive work on privacy-preserving data publishing.15,16 Although directly using general-purpose privacy-protection tools appears straightforward, 2 special characteristics of health disparity research call for special treatment:

1) Privacy disparity: the specific vulnerability of underprivileged population to privacy intrusions. A critical consequence of privacy disclosure is collateral damage: an individual held liable not for his/her own behavior but those who are “similar” on social determinants.17 This harm is more pronounced for underprivileged groups, as 52% Americans in the lowest earning group (household income under $20,000) are “very concerned” about the collection of personal information, compared with just 37% in the highest group (over $100,000).18

2) Privacy protection masking inequality: privacy protection hides the presence of individuals, thereby increases the risk of masking inequalities: for example, to achieve δ-differential privacy,19 a state-of-the-art standard, we automatically attain kδ-differential privacy for any k individuals. That is, if we seek strong privacy protection (a small δ = 0.001) when there is a small underprivileged subpopulation (k = 100), then even the presence of this subpopulation becomes undetectable from privacy-preserved data, masking potential disparities.

SCOPE OF PAPER

This paper demonstrates the existence of privacy disparity in health care data, and examines the effectiveness of state-of-the-practice/art tools on protecting privacy and enabling health disparity research. It does not aim to solve all challenges facing privacy in health disparity research: the design of privacy-preserving tools can be extremely complex, demanding specific

From the American University, Washington, DC.
Supported in part by NSF 1343976, 1851637, and ARO W911NF-15-1-0020. The authors declare no conflict of interest.
Reprints: Heng Xu, PhD, Kogod School of Business, American University, 4400 Massachusetts Ave NW, Washington, DC 20016. E-mail: xu@american.edu.
Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.
ISSN: 0025-7079/19/S706–S172

S172 | www.lww-medicalcare.com

Medical Care • Volume 57, Number 6 Suppl 2, June 2019

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.
METHODS

Background

Dataset

We used the Pennsylvania Inpatient (PA) Dataset (486,924 records, 82 attributes) for Q1 2001.20 Personally identifiable information has been removed, but demographics like sex, age, race, ZIP code remain. Other fields include admission, treatment, and discharge information.

We chose the PA dataset because of its representativeness: the demographics attributes included in the dataset are widely prevalent in other datasets. Among 20 state-level inpatient databases maintained by Agency for Healthcare Research and Quality (AHRQ), all 20 include sex, 19 include age, 18 include race, and 17 include ZIP code.21 All 4 are included in AHRQ nationwide databases such as National Inpatient Sample (NIS)22 and Kids’ Inpatient Database (KID).23

It is also important to note that the results in this article do not depend on the specific composition of the PA dataset: none of the 4 demographics attributes (or even their combination) can guarantee the unique identification of an individual. Instead, each of these (and many other) attributes discloses partial identity information that causes privacy concerns. Thus, the results of this paper readily extend to datasets with different coverage of demographics and/or health-related attributes.

State-of-the-Practice Anonymization (Texas Procedure)

When releasing the Texas Inpatient Data, a dataset structurally similar to PA dataset, Texas Health Care Information Collection (THCIC) applied anonymization rules,24 also applicable to PA dataset:

- Suppress last 2 digits of ZIP with <30 patients.
- Suppress ZIP, facility-type indicators, and provider-ID if hospital has <50 discharges.
- Suppress ZIP and provider-ID if hospital has <5 discharges of a sex.
- Change Race to ‘Other’ and suppress ethnicity if hospital has <10 discharges of a race.

Although these rules were specific proposed by THCIC, their design principles are shared across anonymization attempts in public health practices: Department of Health and Human Services recommended the suppression of the last 2 digits of ZIP code as one of the measures for complying with HIPAA.25 New York State applies the same ZIP code suppression for HIV records starting in 2008.26

State-of-the-Art Differential Privacy Techniques

Depending on how privacy-preserved data are used, many different designs exist for differential privacy.16,27,28 We considered as an example \( \chi^2 \) independence test, and adopted MCIndep,19 a state-of-the-art differential privacy algorithm for this test.

Description of Methods

Privacy Disparity

To study privacy disparity, we examined whether releasing PA dataset levies different privacy risks against different races. An important issue here is how to measure privacy risk. Although numerous privacy measures have been proposed, no universal measure exists because the potential linkage between an administrative dataset and many extraneous (eg, social media) datasets makes it difficult to quantify privacy risk without understanding all linkage channels. Although modern techniques like differential privacy can be robust against extraneous linkage, they do not measure privacy disclosure of raw data. For the purpose of this article, we adopted 2 simple measures to study privacy disparity:

- **Percent uniqueness (PU):** percentage of records with a unique value combination on the attributes of interest.15 A higher PU naturally indicates higher privacy risk.
- **Information entropy (IE):** IE29 of the distribution of attributes of interest. Intuitively, IE measures the number of bits required to “describe” the attributes. The lower IE is, the less uncertainty there is for attribute values, the higher privacy risk becomes.

We measured PU and IE for 2 types of attributes: (1) medical-related ones like diagnosis codes; and (2) demographics ones like age and sex; and compared each across racial groups.

Our metrics are more generic than other existing ones, which are specific either to the technique used for privacy preservation,30 or to a predetermined baseline distribution for sensitive data.31 As a result, our metrics can be applied to both original and anonymized data, fitting the need of our research.

Examination of Anonymization

We then examined the effect of Texas procedure on PA dataset, specifically how it affects utility for health disparity research. To capture utility reduction, we considered 2 measures:

- **Number of suppressions:** count of individuals with attribute values suppressed.
- **Error incurred to analytical query answers:** we considered as a sample query the frequency distribution of principal diagnosis code (P_DIAG), and used the Kullback-Leibler (KL)-divergence32 to measure the distance between the raw distribution from PA data with that from the anonymized data. Intuitively, KL-divergence KL(\( P \rvert Q \)) captures the number of extra bits one needs to depict a distribution \( P \) after learning \( Q \). The smaller it is, the less impact anonymization has on data utility.

Examination of Differential Privacy

Our goal was to test whether, after achieving differential privacy, the dataset still supports statistics used in health disparity research. We applied MCIndep over PA dataset, and studied how it affects the outcome of an independence test between “Race” and “MedQual Atlas Midstay Nonresponder Indicator” (whether a patient is responding to treatment: Yes-No-Ineligible). To study how differential privacy could mask inequalities, we considered subsets of PA dataset, specifically: (1) all records with
race = Asian or white, (2) African American or white, and (3) Native American or white. Note that running independence test over original data, including all 3 subsets, rejects the null hypothesis with high significance ($P < 0.00001$). In setting the goal of $(\epsilon, \delta)$-differential privacy for MCIndep$_D$, we kept $\delta = 10^{-6}$ as, and varied $\epsilon$ from 1 to 0.01.

**RESULTS**

**Privacy Disparity**

Table 1 depicts the number of records for each race and for principal diagnosis, age, and ZIP code-PU, IE, and the number of unique values in each race. Privacy disparity is readily observable, as both PU and IE suggest disproportional privacy risks for Asians and Native Americans.

**Examination of Anonymization**

Table 2 shows the results of anonymization. Although the suppression percentage for ZIP is relatively uniform, the suppression of race disproportionately affects Asians ($11.08\%$). Although anonymization had little effect on whites (with KL-divergence $\sim10^{-6}$ bit), its effect on Native Americans is significant ($>2.5$ bits). Given the entropy of P_DIAG is only $\sim8$ bits, Texas procedure causes substantial utility loss for Asians and Native Americans.

**Examination of Differential Privacy**

Table 3 depicts the average $P$-value over 10 runs. Observe that, to achieve reasonable privacy ($\epsilon \leq 0.1$), one cannot reject the null hypothesis for the subset of Asians and whites. Even for the entire dataset, when $\epsilon = 0.01$, the average $P$-value (0.009) makes the rejection questionable. This demonstrates differential privacy masks disparities that would otherwise be obvious.

### TABLE 1. Privacy Disparity in the Pennsylvania Inpatient Dataset

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>African American</th>
<th>Asian or Pacific Islander</th>
<th>Native American or Eskimo</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>73.73%</td>
<td>10.91%</td>
<td>0.47%</td>
<td>0.29%</td>
<td>12.91%</td>
</tr>
<tr>
<td>Principal diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU</td>
<td>23.41%</td>
<td>32.90%</td>
<td>60.19%</td>
<td>60.85%</td>
<td>30.47%</td>
</tr>
<tr>
<td>IE</td>
<td>8.8659</td>
<td>8.8780</td>
<td>7.4454</td>
<td>5.5356</td>
<td>8.7936</td>
</tr>
<tr>
<td>UV</td>
<td>8.5101</td>
<td>3024</td>
<td>638</td>
<td>519</td>
<td>3128</td>
</tr>
<tr>
<td>Principal procedure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU</td>
<td>19.75%</td>
<td>33.12%</td>
<td>53.40%</td>
<td>70.18%</td>
<td>27.59%</td>
</tr>
<tr>
<td>IE</td>
<td>5.4798</td>
<td>5.1131</td>
<td>5.0427</td>
<td>5.7867</td>
<td>5.4755</td>
</tr>
<tr>
<td>UV</td>
<td>2132</td>
<td>1274</td>
<td>309</td>
<td>357</td>
<td>1341</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU</td>
<td>1.80%</td>
<td>0.93%</td>
<td>2.08%</td>
<td>10.11%</td>
<td>2.75%</td>
</tr>
<tr>
<td>IE</td>
<td>6.2480</td>
<td>6.2771</td>
<td>5.5528</td>
<td>5.9357</td>
<td>6.2276</td>
</tr>
<tr>
<td>UV</td>
<td>111</td>
<td>107</td>
<td>96</td>
<td>89</td>
<td>109</td>
</tr>
<tr>
<td>ZIP code</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU</td>
<td>32.28%</td>
<td>35.69%</td>
<td>33.68%</td>
<td>49.43%</td>
<td>35.10%</td>
</tr>
<tr>
<td>IE</td>
<td>8.5775</td>
<td>6.2363</td>
<td>7.1827</td>
<td>2.0587</td>
<td>7.7371</td>
</tr>
<tr>
<td>UV</td>
<td>6.4355</td>
<td>1020</td>
<td>380</td>
<td>87</td>
<td>2188</td>
</tr>
</tbody>
</table>

Bold represents values in a row which reflects the lowest risk for privacy disclosure, while italics represent the value with the highest risk in the row.

**DISCUSSIONS AND POLICY IMPLICATIONS**

We illustrated the challenge facing the balancing between privacy and health disparity research. Our results identified privacy disparity in public health data, and found anonymization and differential privacy techniques may disproportionately suppress information for minorities or mask health disparities. Although the specific results were generated from a state-level dataset, the issues unveiled extend to datasets larger (eg, nationwide databases) or smaller (eg, research datasets), even though their achievable tradeoffs between privacy and utility differ. For example, while a larger dataset naturally retains a higher statistical power after anonymization, the retained power still differs between various subpopulations, making signals detectable for some undetectable for other, minority, populations causing disparity concerns.

### TABLE 2. Effect After Applying the Texas Procedure

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>African American</th>
<th>Asian or Pacific Islander</th>
<th>Native American or Eskimo</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race suppression (count)</td>
<td>20</td>
<td>143</td>
<td>254</td>
<td>249</td>
<td>666</td>
</tr>
<tr>
<td>Race suppression (%)</td>
<td>0.006%</td>
<td>0.27%</td>
<td>11.08%</td>
<td>3.04%</td>
<td>0.14%</td>
</tr>
<tr>
<td>ZIP suppression (count)</td>
<td>14,134</td>
<td>577</td>
<td>46</td>
<td>265</td>
<td>17,537</td>
</tr>
<tr>
<td>ZIP suppression (%)</td>
<td>3.94%</td>
<td>1.09%</td>
<td>2.01%</td>
<td>3.24%</td>
<td>3.60%</td>
</tr>
<tr>
<td>Total suppression (count)</td>
<td>14,147</td>
<td>699</td>
<td>291</td>
<td>502</td>
<td>18,154</td>
</tr>
<tr>
<td>Total suppression (%)</td>
<td>3.94%</td>
<td>1.32%</td>
<td>12.69%</td>
<td>6.14%</td>
<td>3.73%</td>
</tr>
<tr>
<td>Entropy (P_DIAG)</td>
<td>7.4454</td>
<td>8.8780</td>
<td>8.8659</td>
<td>8.1622</td>
<td>8.9215</td>
</tr>
<tr>
<td>KL-divergence (P_DIAG) KL-Divergence (P_DIAG) KL-Divergence (QIP)</td>
<td>1.41 × 10^{-6}</td>
<td>0.1535</td>
<td>1.2972</td>
<td>2.5562</td>
<td>1.4989</td>
</tr>
</tbody>
</table>

Italics indicates which value in a row reflects the most (negative) effect on data utility. In the case of suppression count, this is corresponding to the highest percentage. In the case of information-theoretic measure, this is corresponding to the highest KL-divergence. $^*$ $P$ and Q are the distributions before and after anonymization, respectively (KL-divergence is asymmetric).

### TABLE 3. Average $P$-value for $\chi^2$ Independence Test After Differential Privacy

<table>
<thead>
<tr>
<th></th>
<th>Original Data</th>
<th>$\epsilon = 1$</th>
<th>$\epsilon = 0.1$</th>
<th>$\epsilon = 0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A, B, N, W)</td>
<td>1.69 × 10^{-24}</td>
<td>&lt;1 × 10^{-4}</td>
<td>&lt;1 × 10^{-4}</td>
<td>0.009</td>
</tr>
<tr>
<td>(B, W)</td>
<td>1.51 × 10^{-20}</td>
<td>&lt;1 × 10^{-4}</td>
<td>&lt;1 × 10^{-4}</td>
<td>&lt;1 × 10^{-4}</td>
</tr>
<tr>
<td>(A, W)</td>
<td>5.56 × 10^{-6}</td>
<td>1 × 10^{-4}</td>
<td>0.151</td>
<td>0.415</td>
</tr>
<tr>
<td>(N, W)</td>
<td>3.21 × 10^{-7}</td>
<td>&lt;1 × 10^{-4}</td>
<td>0.018</td>
<td>0.284</td>
</tr>
</tbody>
</table>

The variations in font represent whether the null hypothesis can be confidently rejected (bold) or cannot be rejected (italics). Bold-italics mean the rejection depends on the predetermined significance level (eg, rejected for 0.05, not rejected for 0.005).

A indicates Asian or Pacific Islander; B, African American; N, Native American or Eskimo; W, white.
Our findings have implications on both technical and policy research. Technically, instead of offering a one-size-fits-all privacy guarantee to all individuals for all purposes, the research community need to develop more flexible solutions to specifically enable health disparity research. One such possibility is to customize the privacy-preservation process to fit the utilization patterns of health disparity research, for example, in the case of differential privacy, by taking into account the importance of a statistic for disparity research when allocating the privacy budget.16

Policewise, regulators must carefully study the privacy implications before mandating the inclusion of attributes like race and geolocation in a released dataset. Although doing so could help identify health disparities, it may also have the danger of introducing privacy disparities, causing further harm to minorities. In contrast, mandating privacy could be problematic too, as the direct usage of off-the-shelf anonymization tools may mask disparities that are otherwise identifiable. Like in the technical case, it may be important to recognize the special needs of health disparity research and its inherent tension with privacy, and therefore, allow certain practical tradeoffs instead of mandating one over another. An example here is to allow certain patients to specifically consent to the usage of their data for health disparity research, while subjecting all other usages to more stringent privacy requirements. Although doing so introduces new challenges, for example, on ensuring the specially permitted data are not used for any other purpose, it provides a potentially effective solution to address the fundamental tension between enabling health disparity research and addressing privacy disparity.

Enabling health disparity research while avoiding privacy disparity is a complex challenge. It is our belief that addressing it requires extensive efforts in technical research and policy discussions, and our hope that this paper serves as a call for attention to the stakeholders and researchers.

REFERENCES