Small Data Privacy Protection: An Exploration of the Utility of Anonymized Data of People with Rare Diseases

Haley MacLeod, MS, Jacob Abbott, MS, Sameer Patil, PhD
Indiana University, Bloomington, IN, USA

Abstract

Sociotechnical researchers have recently begun studying people with rare diseases. There is potential for impact if data can be anonymized and shared so additional research can take place. However, this data also presents a high risk of re-identification because of the rarity of the diseases. Using existing research on data protection techniques, we generate an anonymized version of a rare disease data set and explore the utility of this data in replicating existing rare disease research. We also explore the utility of this data for seven additional use cases generated by other researchers. We find the loss of utility varies depending on the use case, analysis method, and evaluation metrics.

Introduction

Researchers have dedicated themselves to learning about chronic illness populations to improve existing support technologies, design new interventions, or gain medical insight from personal health data. A more recent area of focus is rare diseases – conditions that, by definition, impact an extremely small number of people. In the US, a rare disease is defined as one that impacts less than 0.06% of the population[1]. Although each individual disease is rare, it is estimated that 10% of people worldwide have one of the approximately 7,000 different known rare diseases[1].

Compared to common chronic illnesses, rare diseases are not as well understood in the medical literature and do not have the same constrained set of symptoms to design for. There is potential for impact by reporting results from studies of rare disease communities and even greater impact if data can be made available to citizen scientists and researchers alike so that additional research can take place. The success of many citizen science initiatives relies on open data sharing, but sharing rare disease data poses challenges from a data protection and privacy perspective. Simply removing direct identifiers from quantitative data sets containing people with rare diseases is insufficient. Only a small number of attributes are needed to re-identify an individual from such a data set, particularly in conjunction with the name of the (rare) disease. For example, amyotrophic lateral sclerosis (ALS) is a disease that, although widely known because of the recent ALS ice bucket challenge[2], impacts less than 0.01% of Americans. Using only gender, age, and country of an individual, we can narrow down to close to approximately 1,000 people in the US. For those who live in a smaller country–say Denmark–we could narrow down to approximately 20 people.

We examined a reasonably small ($n = 341$) data set of self-reported behavioral data of people with rare diseases[3]. We generated an anonymized version of this data set with the idea that it could be released to the public to facilitate additional research. Next, we explored changes in the utility of the data as a result of anonymization. To do this, we replicate the original analysis[3] using the anonymized data. We also studied the change in utility for seven additional use cases generated by other researchers.

Data Protection & Anonymization

Making data publicly available can be extremely beneficial and lead to further research that might not be possible otherwise, especially in the case of rare diseases. However, there are numerous examples of re-identification of anonymized health data by linking quasi-identifiers to other public data sets. For example, Sweeney et al.[4] used quasi-identifiers (date of birth, zip code, and gender) to link a database of health insurance claims (with direct identifiers removed) to a voter registration list, resulting in identification of individuals listed in both databases.

To combat these kinds of privacy threats, Sweeney introduced k-anonymity[4]. A data set can be said to have k-anonymity protection if *"the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release"*[4]. The goal is to release a version of a data set *"with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful"*[4]. Typically, k-anonymity is achieved by *suppressing* (replacing values of the attribute with an asterisk) or *generalizing* (replacing the attribute with a broader category) variables.

Method

In our analysis of the trade-off between privacy and utility in sharing rare disease data, we rely on MacLeod et al.'s[3] data set. This data consists of information about self-reported online behavior and perception of health care providers by people with rare diseases and common chronic illnesses. The full data set consists of 341 responses and includes:

1. ***demographic information*** (age, gender, country of residence, marital status, employment status, and education),
2. ***disease information*** (disease name, number of years of experiencing symptoms, number of years since diagnosis, perceived severity of symptoms),
3. ***technology use information*** (frequency of Internet use, devices owned, use of health applications, health information seeking behaviors, health tracking behaviors, health information sharing behaviors, health support seeking behaviors), and
4. ***perception of health care providers*** (number of specialists, helpfulness of different sources (health professionals, friends/family, other patients, online sources) for different health tasks or questions).

From this data set, we consider age, gender, country of residence, martial status, employment status, education, and disease name to be quasi-identifiers, as these have been used for re-identification in previous research[4]. We defined generalization hierarchies for each quasi-identifier and used ARX[5] to generate a k-anonymized version of this original data set. We begin with $k = 2$ (i.e., an individual in the data set cannot be distinguished from at least one other individual in the data set) to assess the changes in utility when the most basic, minimal protection is applied.

The result of this 2-anonymization is a data set in which age, country, and martial status are removed entirely. Disease names are replaced with a designation of whether the person has a common chronic illness, a rare disease, or both. All other quasi-identifiers remain unchanged. We use this new 2-anoymized data set to replicate our previous analysis[3] that used a range of classification methods to predict whether someone has a rare disease or a common chronic illness based on the self-reported behavioral survey data.

This use case is not necessarily representative of the ways *other* researchers might use this data. We originally used the data to distinguish between people with rare diseases and those with common chronic illnesses. We made decisions about how the data should be discretized, removing disease names entirely, relying solely on the labels "one or more rare diseases", or "one or more common chronic illnesses" as the prediction target. We similarly made decisions about how best to discretize features such as age, years of experiencing symptoms, and years since diagnosis. All these decisions were made based on the specific use case and goals of the research.

To investigate alternative use cases, we asked five machine learning researchers to suggest a possible uses based on a description of the data. The researchers were given freedom to select features and subsample and discretize the data as they saw fit to complete their proposed task. Additionally, they were asked to repeat their pre-processing on the anonymized version of the data set and complete the task again. Note that the use case is the *same* for the original and anonymized data sets, however, the pre-processing schemes may vary. Because the anonymization process involves suppressing and generalizing variables, the researchers sometimes chose (or were forced to) use different pre-processing approaches based on domain expertise and intuition. Researchers also chose the evaluation metric(s) to use. For each use case, we compared the performance of the original data set against the anonymized version. The difference in utility relative to the performance of the original (sensitive) data is calculated as:

$$\frac{Protected - Original}{Original} * 100$$

Findings

Our experiments focus on two key questions: (**Q1**) How much utility is lost when we use an anonymized version of the data to predict disease? and (**Q2**) How much utility is lost when we use an anonymized version of the data for alternative use cases? We present each in turn.

Table 1: Results of replicating MacLeod et al.'s study using the original data (Orig.) and protected data (Prot.). We report on the relative difference (Dif.) between these two versions of the data for three evaluation metrics: true positive rate (TPR) and two weighted F measures (F_3 and F_5).

	TPR			F_3			F_5		
	Orig.	**Prot.**	**Dif.**	**Orig.**	**Prot.**	**Dif.**	**Orig.**	**Prot.**	**Dif.**
Naïve Bayes	0.507	0.524	3.35%	0.507	0.528	4.14%	0.507	0.525	3.55%
Logistic Regression	0.413	0.456	10.41%	0.400	0.448	12.00%	0.402	0.453	12.69%
5-Nearest Neighbors	0.288	0.262	-9.03%	0.299	0.279	-6.69%	0.293	0.268	-8.53%
Decision Trees	0.408	0.262	-35.78%	0.412	0.271	-34.22%	0.410	0.265	-35.37%
Random Oversampling	0.511	0.454	-11.15%	0.504	0.471	-6.55%	0.508	0.460	-9.45%
Random Undersampling	0.657	0.553	-15.83%	0.632	0.566	-10.44%	0.647	0.558	-13.76%
SMOTE	0.472	0.360	-23.73%	0.471	0.382	-18.90%	0.472	0.368	-22.03%
Soft-Margin Gradient Boosting	0.972	0.708	-27.16%	0.825	0.795	-3.64%	0.909	0.793	-12.76%

Replication (Predicting Disease Type)

To answer **Q1**, we replicated our previous analysis[3] to compare the performance of the anonymized version of this data against the performance of the original version of the data. The goal of the original work (and of our replication) was to predict whether someone had a rare disease or a common chronic illness. We use the same classifiers and evaluation metrics as the original study[3]. Table 1 shows the results of replicating the analysis with the original data set and with the anonymized version.

Soft-margin functional gradient boosting (Soft-FGB)[6] achieves the best results for all three metrics. There is a noticeable drop in performance using soft-FGB (27.16% drop in true positive rate between the original and anonymized versions of the data set). Yet, as Table 1 shows, the reduced true positive rate on the anonymized data (0.708) is still better than the performance of the other classifiers on the original data (0.262–0.657). Importantly, the decrease in performance is the smallest when we examine the F_3 score (3.64%), suggesting that the anonymization hurts recall more than precision. This is problematic as it was recall that we were aiming to maximize in our goal of being able to identify the rare disease examples (more so than the common chronic illness ones). Therefore, the results indicate that the utility of the anonymized data depends on choice of evaluation metric.

Researcher-Generated Use Cases

To answer **Q2**, we asked five researchers with expertise in machine learning to suggest use cases or tasks suitable for the data. Each researcher was responsible for decisions around pre-processing, classifier choice, and evaluation metrics. We compared the performance of the original (sensitive) data against the anonymized data for each use case suggested by the researchers.

Use Case: Disease Type. Two researchers aimed to predict disease type, but they took different approaches. The first researcher used a random forest (RF) approach with 5-fold cross validation, comparing the original data to the anonymized data. He further aimed to make this prediction with the demographic information removed, using the original data set and the anonymized data set. He used accuracy as their evaluation metric.

Table 2: Results (accuracy, AUC-PR) of predicting disease type with and without demographic information for the original and anonymized data sets.

	With Demo.			No Demo.		
	Orig.	**Anon.**	**Dif.**	**Orig.**	**Anon.**	**Dif.**
RF	0.828	0.738	-10.87%	0.865	0.729	-15.72%
LSVM	0.520	0.519	-0.19%	-	-	-

The second researcher used a linear support vector machine (LSVM) and nested-cross-validation (30 trials of four stratified-folds for outer cross-validation loop and 10 stratified-shuffle-splits for inner loop in the hyper-parameter search). As in MacLeod et al.'s[3] analysis, this researcher chose to remove disease name from the original data set stating, *"that felt like cheating"*. For evaluation metrics, this researcher chose Area Under the Precision and Recall curve (AUC-PR), averaged over the 30 trials.

Table 3: Results (accuracy) of predicting quality of life with and without disease information for the original and anonymized data sets.

	With Disease			No Disease		
	Orig.	**Anon.**	**Dif.**	**Orig.**	**Anon.**	**Dif.**
NB	0.359	0.359	0.00%	0.327	0.308	-5.81%
2-NN	0.308	0.306	-0.65%	0.315	0.333	5.71%
5-NN	0.303	0.324	6.93%	0.315	0.313	-0.63%
15-NN	0.326	0.306	-6.13%	0.299	0.327	9.36%
RF	0.332	0.292	-12.05%	-	-	-

Table 4: Results (accuracy) of predicting social network use for original and anonymized data.

	Orig.	**Anon.**	**Dif.**
RF	0.511	0.474	-7.24%

Table 5: Results (accuracy) of predicting most helpful sources of support for the original and anonymized data.

	Orig.	**Anon.**	**Dif.**
DT	0.521	0.445	-14.59%
NB	0.503	0.533	5.96%
LR	0.545	0.447	-17.98%

Table 6: Results (accuracy) of predicting number of specialists

	Orig.	**Anon.**	**Dif.**
RF	0.471	0.474	0.64%

Table 7: Results (accuracy) of predicting health information sharing

	With Demo.			No Demo.		
	Orig.	**Anon.**	**Dif.**	**Orig.**	**Anon.**	**Dif.**
NB	0.628	0.632	0.64%	0.637	0.639	0.31%
2-NN	0.620	0.620	0.00%	0.566	0.616	8.83%
5-NN	0.607	0.598	-1.48%	0.605	0.598	-1.16%
15-NN	0.602	0.639	6.15%	0.604	0.593	-1.82%

Table 8: Results (accuracy) of predicting disease name

	Orig.	**Anon.**	**Dif.**
BN	0.377	0.00	-100.00%
Bag	0.375	0.00	-100.00%
Boost	0.359	0.00	-100.00%

The difference in performance between the original and anonymized data for the LSVM approach is virtually nonexistent (Table 2). When we compare this finding against the results of the random forest approach and several of the approaches attempted in MacLeod et al.'s[3] work, this suggests that a linear support vector machine approach to this task is more robust to the loss of data; the utility of the data depends on the choice of classifier.

Use Case: Quality of Life. Two researchers aimed to be able to predict "How severely do your symptoms impact your life?" In particular, one researcher wanted to understand the importance of the disease type in making this prediction. This researcher first used the original and anonymized data sets with all variables included, then re-ran the analyses on the anonymized data set with additional variables removed (all demographic information including disease type removed in the first iteration and all demographic information except disease type removed in the second iteration). This researcher chose to use Naïve Bayes and k-nearest neighbours ($k = 2, 5, 15$). Based on these results, the first researcher concluded that disease type is an important attribute to predict quality of life when using an algorithm like Naïve Bayes, which is more susceptible to poor performance due to missing information. The other researcher aimed to make the prediction without demographic information (i.e., based purely on disease information, technology use, and perception of health care providers). This researcher used a random forest approach with 5-fold cross validation. Both researchers chose accuracy as the evaluation metric (Table 3). In this task, the change in performance is less noticeable – most of these differences could be easily attributed to expected statistical deviations. Therefore, using the anonymized data would be acceptable for this task.

Use Case: Social Network Use. One researcher proposed a use case where the goal was to use the devices owned by an individual to predict whether he or she used social networks as a source of health information (Table 4). He used a random forest approach with 5-fold cross validation and accuracy as the evaluation metric.

Use Case: Most Helpful Sources of Support. One researcher proposed using the data to predict the most helpful source of support. She used three different classifiers for this prediction (decision trees (DT), Naïve Bayes (NB), and logistic regression (LR)) and accuracy as her evaluation metric. As in the disease type prediction task (Table 5), this is a case where the change in performance varies by classifier; Naïve Bayes does not show the same drop in performance as decision trees and logistic regression.

Use Case: Number of Specialists. One researcher was interested in predicting how many specialists someone sees. He used a random forest approach with 5-fold cross validation and accuracy as the evaluation metric. This task, and to a lesser extent the social network use task (Table 6), may both be cases where using the anonymized data would be sufficient. However, this use case was only attempted with a single classifier. It's possible that if the researcher were to use different classifiers or metrics this would not continue to be the case.

Use Case: Sharing Information Online. One researcher suggested predicting whether someone shares information about medical treatments online. As in previous examples, the researcher compared the original data against the anonymized data. Further, she looked at the data with all demographic data removed. She found the technological features sufficient to predict the online sharing behavior of a person, rendering other demographic information unnecessary. Most of the differences are reasonably small and within expected statistical deviations (Table 7). Therefore, this is another case where the anonymized data would be sufficient for the task.

Use Case: Disease Name. One researcher attempted to classify the specific diseases of the individuals in the data set. For individuals with multiple chronic conditions, she made duplicate entries such that each case has exactly one disease. She included only those cases where there were at least two instances of the disease in the data set (removing conditions that occurred only once). She used Bayesian Networks (BN), Bagged Decision Trees (Bag), and Boosted Decision Trees (Boost) to make this classification (Table 8). Unsurprisingly, it is not possible to complete this task with the anonymized data, given that disease name is one of the suppressed variables in the anonymized data.

Discussion

Tasks like understanding the impact of disease type on quality of life (Table 3) and predicting online sharing behavior (Table 7) do not demonstrate much of a loss of utility. It is likely that researchers could complete these tasks using only the anonymized data instead of the sensitive data, thus preserving the privacy of people with rare diseases. Other tasks like predicting disease type (Table 2) or the most helpful sources of support (Table 5) are more challenging and depend on the classifier/evaluation metrics used and how influential the protected features are in making the classifications. In these cases, it is clear that having the original data set would be preferable for improving results but this introduces potential risk to the people in the data set. What is not clear is whether the protected data is sufficient to draw *some* conclusions. That is, if the only choices were using the protected data or not conducting the analysis at all, is the performance of the protected data still high enough that the analysis is worth doing? This is an interesting question for future work.

Finally, tasks like diagnosing one's disease (Table 8) are impossible to complete without the original data because the target of the prediction (disease name) is one of the variables that is entirely suppressed in the anonymized version (the only remaining information is whether the disease in question is common or rare). If researchers would like to undertake this kind of prediction task, they would need to enter into a special agreement with those responsible for the data. An alternative for future research may be to explore different taxonomies or ways of clustering diseases such that instead of two categories of disease (rare vs. common) there could be many.

Our goal with this work was to explore the possibility of being able to release a single protected version of a rare disease data set that meets some minimal privacy preserving criteria (in this case, 2-anonymity such that individuals in the data set cannot be distinguished from at least one other individual in the data). Our findings have shown that this 2-anonymity protected data set is not universally useful; higher values of k would only lead to greater generalization and suppression of data, thus reducing the performance even further. It may be possible to maintain utility at higher values of k in a larger data set, but rare disease data sets, by definition, tend to be small.

Designing tools for privacy-preserving data sharing that take into account community input will be critical to enabling additional research, while protecting the privacy of research subjects. In most cases currently, researchers will collect the data and do some cleaning or pre-processing, and then make it available in some repository for others to use. This might be appropriate in cases of non-human subjects data or other low privacy risk contexts, but it may not be appropriate in the case of behavioral health data (especially rare diseases). It may be possible to create anonymized versions that are better tailored to specific use cases, but we are hesitant to release multiple differently protected versions of the same data set in case of attacks. It is valuable to consider as many use cases as possible prior to anonymizing and releasing the data.

Tools for privacy preserving data sharing should consider community input on the uses of data. One option here is to release a description of the data or a list of the variables/features without releasing the actual data itself. Community members could submit use cases or questions they were interested in answering from the data. This was the case in Wagner et al.'s[7] Device Analyzer project, where the authors solicited input from researchers about *"questions that the research community would like to answer using our dataset"*[7].

It might not be possible to *completely* assess the usefulness or information gain from each specific feature without actually having access to the data, but if the researchers know as much as possible about the use cases they can better generate a single anonymized version that takes these tasks into consideration. In our research, there weree several use cases where the change in performance is negligible or tolerable; understanding the potential use cases allows us to assess the need for sharing the sensitive data via special agreement vs. just providing the protected version.

Conclusion

Although specific rare diseases impact only very small numbers of people, the number of people impacted by a rare disease worldwide is substantial. Despite these numbers, rare diseases are largely under researched. Appropriately anonymizing data such that it can be released to the community can enable additional research to take place without as much burden on researchers to collect new data from scratch.

This paper provides an initial exploration of the utility of anonymized rare disease data. Unsurprisingly, we find that the loss function depends on the use case, the choice of classifier, and the evaluation metric used. Thus, we outline a need for tools that protect privacy of participants while increasing the amount of research that can take place and decreasing the burden on researchers. These tools should allow community members to contribute information on the potential uses of the data before the anonymized version of the data is generated, such that the anonymized version can be as useful as possible. Usable tools for privacy-preserving data sharing and better methods of obtaining informed consent can allow additional research to take place on previously underserved populations while still preserving privacy and protecting against risks.

Acknowledgments

We thank Suyog Chandramouli, Rion Brattig Correia, Raksha Kumaraswamy, Katherine Metcalf, Phillip Odom, and Shuo Yang for their expertise. We also thank CLEAR Health Information and the Rare Disease Foundation for their support of our rare disease research.

References

[1] Shire Human Genetic Therapies. Rare Disease Impact Report: Insights from patients and the medical community. Shire Human Genetic Therapies; 2013.

[2] Wicks P. The ALS Ice Bucket Challenge – Can a splash of water reinvigorate a field? Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2014;15(7-8):479–480.

[3] MacLeod H, Yang S, Oakes K, Connelly K, Natarajan S. Identifying Rare Diseases from Behavioural Data: A Machine Learning Approach. In: Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016 IEEE First International Conference on. IEEE; 2016. p. 130–139.

[4] Sweeney L. k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(05):557–570.

[5] Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX - A comprehensive tool for anonymizing biomedical data. In: AMIA Annual Symposium Proceedings. vol. 2014. American Medical Informatics Association; 2014. p. 984–993.

[6] Yang S, Khot T, Kersting K, Kunapuli G, Hauser K, Natarajan S. Learning from Imbalanced Data in Relational Domains: A Soft Margin Approach. In: Proceedings of the 2014 IEEE International Conference on Data Mining. ICDM '14. IEEE Computer Society; 2014. p. 1085–1090.

[7] Wagner DT, Rice A, Beresford AR. Device Analyzer: Large-scale Mobile Data Collection. SIGMETRICS Perform Eval Rev. 2014 Apr;41(4):53–56.