Generative Adversarial Networks for Creating Synthetic Free-Text Medical Data: A Proposal for Collaborative Research and Re-use of Machine Learning Models

Suranga N. Kasthurirathne, PhD[1,2], Gregory Dexter[3], Shaun J. Grannis MD, MS[1,2]
[1]Regenstrief Institute, Indianapolis, IN, USA; [2]Indiana University School of Medicine, Indianapolis, IN, USA; [3]Purdue University, Indianapolis, IN, USA

Abstract

Restrictions in sharing Patient Health Identifiers (PHI) limit cross-organizational re-use of free-text medical data. We leverage Generative Adversarial Networks (GAN) to produce synthetic unstructured free-text medical data with low re-identification risk, and assess the suitability of these datasets to replicate machine learning models. We trained GAN models using unstructured free-text laboratory messages pertaining to salmonella, and identified the most accurate models for creating synthetic datasets that reflect the informational characteristics of the original dataset. Natural Language Generation metrics comparing the real and synthetic datasets demonstrated high similarity. Decision models generated using these datasets reported high performance metrics. There was no statistically significant difference in performance measures reported by models trained using real and synthetic datasets. Our results inform the use of GAN models to generate synthetic unstructured free-text data with limited re-identification risk, and use of this data to enable collaborative research and re-use of machine learning models.

Introduction

Rapid uptake of Health Information Systems (HIS) has enabled the accessibility and availability of structured and unstructured electronic health data. These data, together with the rapid evolution of Artificial Intelligence (AI) and various analytical and machine learning toolkits has led to the widespread development of machine learning solutions(1, 2) designed to address organizational-level challenges using organizational-level data. However, the current U.S. regulatory framework limits sharing of Patient Health Identifiers (PHI) outside the healthcare organization(3). Limited or burdensome data access hinders (a) sharing and re-using machine learning solutions across larger audiences, (b) promoting inter-organizational collaboration addressing various healthcare challenges, and (c) building generalized machine learning models targeting diverse populations.

There have been significant efforts to de-identify structured and unstructured patient data for research and dissemination purposes(4, 5). Traditional de-identification efforts focus on the perturbation of potentially identifiable patient demographic attributes such as names, addresses, identifiers, and contact information via randomization, suppression or generalization(6, 7). However, such efforts are not foolproof – patient records scrubbed of PHI may be susceptible to re-identification based on residual clinical information contained in symptoms, diagnosis, medications or lab results(8). This significantly impacts de-identification of structured data due to difficulty in identifying potentially sensitive information from free-text data. Researchers have proposed various approaches for creating synthetic data that mimics clinical patterns in medical records as a solution to re-identification risk based on clinical information(9). A synthetic patient dataset that has been scrubbed of any PHI elements using traditional de-identification methods would be significantly harder to re-identify than a real dataset that has only been scrubbed of PHI elements. However, previous synthetic data generation efforts have resulted in data that are not sufficiently realistic for machine learning(7).

Generative Adversarial Networks (GAN) are a class of deep learning algorithms that offer significant promise to improve synthetic data generation. GAN algorithms are implemented by a system of two neural networks(10). One neural network, the generator, attempts to create synthetic data, while the other neural network, the discriminator, seeks to distinguish between synthetic data and real data. As these networks are trained, the generator network successfully develops synthetic data that cannot be flagged by the discriminator. Initial GAN models were designed to mimic real-valued data(10). As such, they have been used to produce high quality categorical(11) and image datasets(12, 13). In the healthcare domain, GAN models have been used to generate numerical clinical data that is statistically similar to real data(7, 14).

Recent improvements to GAN algorithms enable them to generate synthetic free-text data(15). Researchers have applied these models to successfully generate text data such as molecules encoded as text sequences, musical melodies(16), reviews, dialogues(17), poetry and image captions(18). These innovations offer much potential to the medical field, where a large quantity of clinical information may be trapped within unstructured free-text(19, 20).

We evaluate the potential to leverage GAN models to produce synthetic unstructured free-text medical data that closely reflect characteristics of real data, and thus, may be used to develop machine learning models that approximate similar models created using original data. Next, we will assess the re-identification potential of these synthetic, informationally similar, unstructured datasets.

Materials and methods

Test data selection

We extracted all laboratory messages pertaining to cases of Salmonella reported to the Indiana Network for Patient Care (INPC)(21) during 2016-2017. The INPC is a statewide Health Information Exchange (HIE) that facilitates interoperability across 117 hospitals, 38 health systems, other free-standing laboratories, and physician practices across the state of Indiana. We parsed these messages, which were obtained in the form of Health Level Seven (HL7) version 2 messages, and extracted the free-text report data included in each message. Laboratory messages for salmonella were selected due to the semi-structured nature of the HL7 messages, which allowed us to separate PHI from the unstructured text, as well as the brevity of the free-text laboratory messages. Each message was manually reviewed, and labelled as positive or negative for Salmonella. We randomly selected 90% of each of the positive and negative salmonella messages, hereafter known as positive (train) and negative (train) datasets for training GAN models. The remainder of the datasets, hereafter known as the positive (holdout) and negative (holdout) datasets, were used to test the performance of GAN generated data.

Development of GAN models for synthetic data generation

We adopted SeqGAN, a GAN algorithm designed to generate textual data(22). SeqGAN models approach the sequence generation procedure as a sequential decision-making process. The generative model is treated as an agent of reinforcement learning; the state is the generated tokens while the action is the next token to be generated. The discriminator evaluates the sequence and feeds back the evaluation to guide the learning of the generative model(22). GAN models consist of a number of parameters that can be fine-tuned to optimize model performance. We explored model performance by training multiple GAN models using the positive (train) and negative (train) datasets, and varying several parameters (Appendix A). We adopted a Gaussian distribution as the default initial parameter for all generators. Performance of these models were compared using two document similarity based metrics; embedding similarity, which measures similarity between two documents using similarity between word embeddings, and NLL-test, which evaluates a model's capacity to fit real test data(15). Optimal models selected using this approach were used to generate positive (synthetic) and negative (synthetic) laboratory messages. To build compatible decision models, we generated n positive synthetic reports, where n equals the number of positive (train) messages, and m negative synthetic reports, where m equals the number of negative (train) messages. We trained SeqGAN models using Texygen, a benchmarking platform for GAN based text generation models(15).

Machine learning process

Features extracted from the positive (train) and negative (train) datasets, (jointly known as the real dataset), and positive (synthetic) and negative (synthetic) datasets (jointly known as the synthetic dataset) were used to train multiple classification models using the following approach.

Feature extraction

We mimicked the feature extraction process adopted in our previous work on predicting cancer cases using free-text data obtained from the INPC(23, 24). We developed a Perl script to parse the positive and negative training datasets, and identify all unique stemmed tokens present within these reports. Next, we used the Negex algorithm(25) to identify the context of use (positive or negative) for each stem. We counted the presence of each feature in positive and negated context, and used this data to prepare an input vector for each laboratory message. A similar approach was used to generate vectors of counts representing each message in the synthetic dataset.

Decision model building and evaluation

We applied the Gini impurity(26) metric to rank features in the real and synthetic datasets by order of importance. We used subsets of the top 5, 10, 15 and 20 features selected from the real and synthetic datasets to train a series of decision

models using the Random Forest classification algorithm(27). Random forest was selected due to its proven track record in health care decision-making applications(24, 28, 29). The real and synthetic decision models were tested using feature vectors derived from the positive and negative holdout datasets. We calculated sensitivity (True Positive Rate or Recall), specificity (True Negative Rate), F1-Measure and Area Under the ROC Curve (AUC) for each decision model. Paired t-tests were used to compare the performance of synthetic and real data-based decision models.

Evaluation of re-identification risk

Risk of presence disclosure, aka membership inference assesses an attackers ability to determine if any real patient records in their possession were used to train GAN models by comparing these records against the synthetic patient dataset(30, 31). Assessing risk of presence disclosure ensures the privacy of individuals whose data was used to train a decision model(32), as well as the interests of the healthcare entity where the individual received treatment(33). Thus, presence disclosure is a widely evaluated measure of re-identification risk(34, 35). We assessed risk of presence disclosure using the following experiment(7); we re-purposed vectors that represented the training and synthetic datasets using binary values representing the presence or absence of each feature in positive and negative context. We compared each synthetic record with all training messages using various hamming distance cutoff scores(36), a measure of the minimum number of substitutions required to change one string into the other. The identification of a synthetic record that matched with any training message using a hamming distance equal or smaller to the hamming score threshold would label it as a 'match' to the synthetic record under study. We computed the frequency of matches across each hamming score threshold, and used these metrics to evaluate re-identification risk.

Figure 1 presents our complete workflow, from data extraction to decision model evaluation.

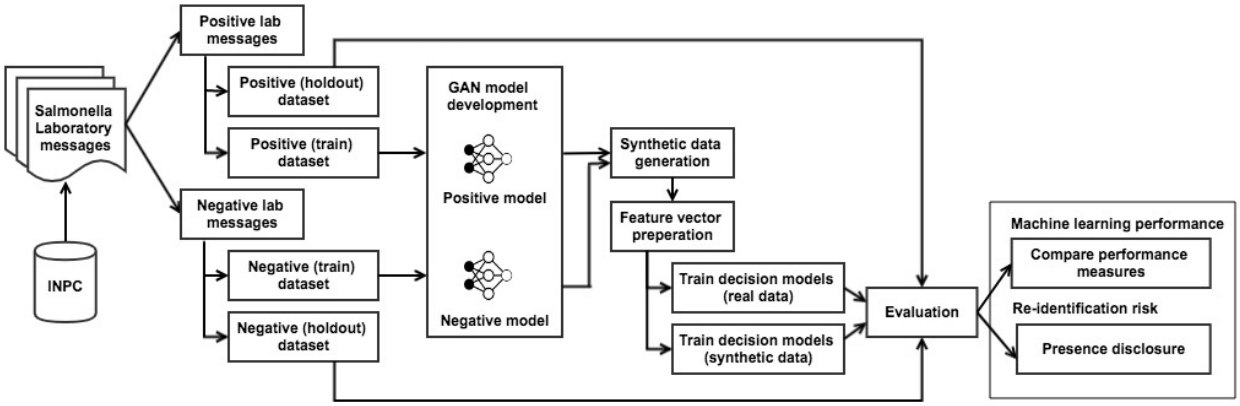


Figure 1. A workflow depicting our study approach from laboratory message extraction to decision model evaluation.

Results

We identified a total of 6,770 laboratory messages pertaining to salmonella. Manual review labelled 1,213 (17.91%) of these messages as positive, and 5,557 (82.08%) as negative. We identified optimal SeqGAN models for generating positive and negative laboratory messages using hyperparameters identified in appendix A. Using these models, we generated 1092 positive synthetic messages and 5001 negative synthetic messages to correspond with the 90% training messages for each dataset. Appendix B presents representative samples from the positive (train), negative (train), positive (synthetic) and negative (synthetic) lab report sets. These samples were manually reviewed, and any PHI elements masked. As seen in appendix B, the only PHI elements identified within these report sets were date, time and report identifier fields. We computed several Natural Language Generation (NLG) measures to evaluate similarity between real and synthetic datasets; a) Bilingual Evaluation Understudy (BLEU) scores(37) are widely used to compare similarity between real and synthetic datasets. We calculated BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores that evaluated the quality of synthetic datasets using 1-gram, 2-gram, 3-gram and 4-gram matches respectively. b) Google-BLEU (GLEU) scores, a measure that seeks to address limitations in BLEU score calculations and are better suited for sentence level comparisons(38). The GLEU score is a composite of all 1-grams, 2-grams, 3-grams and 4-gram matches (table 1).

NLG measure	Positive (train) vs. Positive (synthetic)	Negative (train) vs. Negative (synthetic)
BLEU-1	0.913	0.944
BLEU-2	0.675	0.742

BLEU-3	0.480	0.552
BLEU-4	0.331	0.409
Google-BLEU	0.249	0.328

Table 1. Comparison of real and synthetic datasets using various NLG measures.

These results are comparable to those produced by prior researchers(15), and indicate considerable similarity between real and synthetic datasets. Further, NLG measures comparing negative (train) and negative (synthetic) datasets were higher than the positive (train) and positive (synthetic) datasets. We hypothesize this is because negative reports are more similar due to uniform text documenting negative status. The positive (train) dataset comprised of 2,551 unique stemmed features. 1,827 (71.6%) of these features were present within the positive (synthetic) dataset. The negative (train) dataset comprised of 5,803 unique stemmed features. 4,093 (70.5%) of these stemmed features were present within the negative (synthetic) dataset. With stop words and dates removed, the overall training dataset of positive and negative reports consisted of 3810 unique stemmed features. 2651 (69.6%) of these were present within the overall synthetic dataset. Appendix C lists the top 20 features identified across the real and synthetic datasets using gini impurity scores. Appendix D presents the overlap between the top 5, 10, 15, 20, 50 and 100 features identified across the real and synthetic datasets. We note significant similarity between real and synthetic datasets with between 70% to 80% overlap across each of the feature subsets being compared. Figure 2. presents the sensitivity, specificity, F1-measure and Area under the ROC curve scores reported by decision models built using the top 5, 10, 15 and 20 real and synthetic features upon being tested using the holdout test datasets.

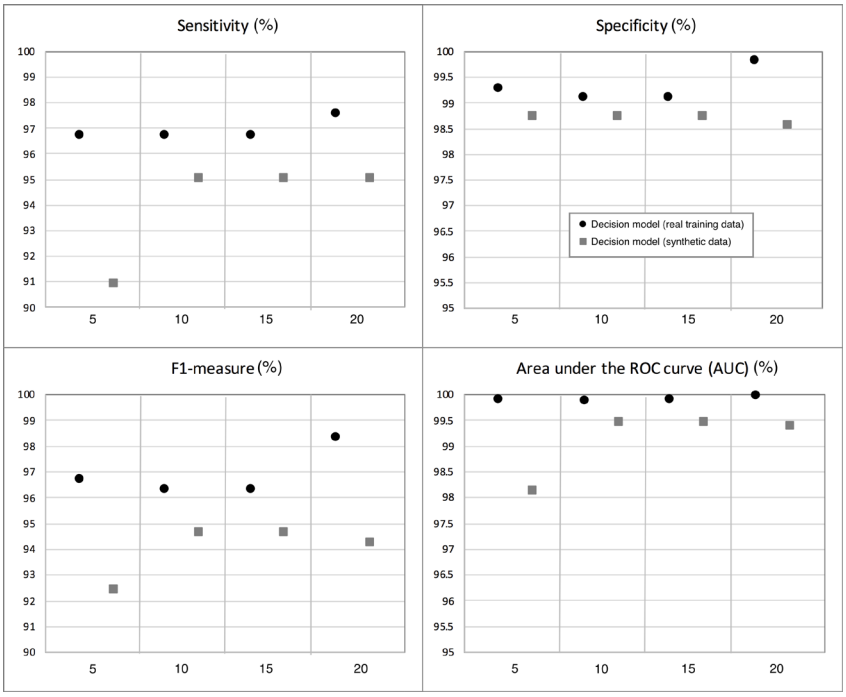


Figure 2. Sensitivity, specificity, F1-measure and Area under the ROC curve scores reported by decision models built using the top 5, 10, 15 and 20 real and synthetic features upon being tested using the holdout test datasets.

Due to the discriminatory power of the features, each model achieved high-performance measures despite being trained on a small number of features. Further, paired t-tests reported statistical significance levels (alpha) of > 0.65. As such, there is no significant difference between performance measures reported by real and synthetic decision models built using any of the feature subset sizes. Given high overlap between top 50 and 100 feature sets (appendix D), we hypothesize that decision models built using the top 50 and 100 real and synthetic features would also report statistically similar performance measures.

Evaluation of re-identification risk

Results of the presence disclosure test are presented in appendix E. We conclude that these results indicate acceptable levels of re-identification risk given that the number of positive matches identified by a hamming threshold of 10 was reasonably small.

Discussion

Our results further two challenges; the use of GAN models to generate synthetic free-text medical data with limited re-identification risk, and use of these datasets to develop machine learning models with statistically similar performance metrics to models developed using the original test data, thereby enabling cross-institutional collaboration and broader dissemination of machine learning models.

Comparison of unique features across test and synthetic datasets revealed that the synthetic dataset contained only 69.6% of the features in the test dataset. We attribute this to the mode collapse problem(39), which leads to reduced diversity of synthetic data(15, 17). NLG scores reported by our models were compatible to scores reported by other efforts to generate synthetic text data extracted from non-medical sources(15). However, we note that the synthetic data presented reduced syntactic/grammatical correctness (appendix B), a common pitfall in deep learning based text generation approaches(40). However, this was irrelevant for our purposes as we only sought to demonstrate that synthetic data could be used to replicate machine learning performance, and not as a tool for training or teaching of humans. Thus, no human evaluation of the synthetic reports was performed. The synthetic dataset also contained a quantity of recurring phrases such as hospital and laboratory test names. The recurrence of such phrases may have positively influenced NLG scores. Despite these limitations, there was 70-80% overlap between the top 5, 10, 15, 20, 50 and 100 features extracted from both datasets (appendix D). Performance measures generated by models trained using top 5, 10, 15 and 20 features extracted from the test and synthetic datasets were high, as well as statistically similar. These results present the possibility of using synthetic datasets to share machine learning solutions, and foster cross-institutional collaboration on various challenges.

Our findings help inform data de-identification efforts. As discussed previously, de-identification efforts involve (a) removal of PHI elements, and (b) addressing re-identification risk based on clinical information in patient records. Adoption of GAN models alone do not result in de-identified data. However, synthetic data generation reduces re-identification risk by creating new patient records with similar, but different content. It also removes any 1-to-1 mapping between test and synthetic reports. Our results using presence disclosure tests confirmed that synthetic datasets pose a small chance of re-identification based on clinical information. However, synthetic data produced by these efforts must undergo rigorous de-identification of PHI elements before they can be distributed for public use. Removal of PHI elements will not impact decision model performance as the top 100 features listed in appendix D did not include any PHI elements.

An alternate approach to evaluate re-identification risk is attribute disclosure, which evaluates an attackers ability to derive additional attributes (features) for a patient based on a subset of attributes they are aware of(41). We did not evaluate our datasets for attribute disclosure as a) unlike longitudinal patient datasets that consist of varied clinical diagnoses that are not necessarily related, the salmonella lab reports consisted of very specific features that are often highly correlated. Thus, it would be relatively easy to predict missing features in our dataset based on those present. Secondly, a considerable number of features presented very low prevalence across the laboratory reports. Thus, predicting 'absence' of these features was relatively easy. However, we argue that these factors pose low risk to the patient because unlike studies that deal with clinical diagnosis that if revealed, may impact the patient's privacy, our dataset focusses on salmonella alone. As an example, discovery of any of the top features listed in appendix C using an attribute disclosure attack would not lead to any harm beyond the awareness that the patient was tested for, and diagnosed as positive or negative for Salmonella. In contrast, attribute disclosure across a different dataset may lead to discovery of multiple clinical diagnosis, patient demographics or other treatment information. We propose the following hypothetical scenario to demonstrate how our approach could be applied in a real-life setting; An organization that possesses rich free-text data sources, but lacks adequate machine learning expertise can leverage our approach to create synthetic data. They de-identify and share the synthetic data with experts who use it to build machine learning models. Once optimal models have been identified, they can be implemented across the original dataset with compatible performance measures.

We identified a number of limitations in our study. Our test dataset consisted of structurally similar reports describing a very specific illness. This, together with the overall simplicity of our predictive outcomes (positive vs. negative for salmonella) may have contributed to our positive results. Datasets that are not structurally similar, nor restricted to a specific illness, or consist of more colloquial language may be harder mimic, and thus, produce less optimal results. Such datasets may require more robust decision models built using other free-text friendly GAN models such as Maximum-Likelihood augmented discrete Generative Adversarial Networks (MaliGAN)(42) or Long Text Generative Adversarial Networks (LeakGAN)(18), and more complex feature vectors consisting of n-grams. Further, our approach was restricted to mimicking synthetic free-text data. It is unclear if our models are able to learn or mimic the

significance of various numeric values such as age or other measurements present in free-text data. This limitation did not impact the performance of our current effort as no numerical values were selected as top features. However, it may impact models built using other datasets.

Future research avenues include use of GAN models to create truly de-identified synthetic free-text data that does not require additional de-identification, and expansion of our work across other more challenging healthcare datasets. Other researchers have demonstrated the ability to mimic numerical and categorical patient data using GAN models(7, 14). Integrating these efforts with ours would enable researchers to share comprehensive synthetic patient health records consisting of both structured and unstructured data for secondary research purposes. Furthermore, our study did not include any analysis of the readability or syntactic/grammatical accuracy of the synthetic reports. As such, these results are suitable for machine learning, and not for teaching or learning resources. Next steps include a) manual assessment of the readability and correctness of synthetic reports using human experts (Turing test), and b) investigation of other word and grammar-based measures that inform synthetic data assessment.

Conclusions

GAN models can be used to generate synthetic unstructured free-text medical data that can be used to replicate the performance of machine learning models with high, as well statistically similar results. Further, synthetic datasets poise limited risk of re-identification based on clinical features. As such, these synthetic datasets can be easily de-identified, and used to champion cross-organizational collaboration efforts.

Appendices

Appendix A. List of hyperparameters evaluated as part of the SeqGAN training process.

Parameter name	Description	Variations attempted
Pre-training epochs	The generator is trained for n epochs, followed by n epochs for the discriminator	Increments of 5 between 10 and 100
Adversarial epochs	Number of adversarial epochs	Increments of 5 between the values 5 and 50
Embedding dimensions	Dimensionality of embedding layer	32, 64, 128
Hidden dimensions	Number of neurons in hidden layer	32, 64, 128
sequence length	Length of each training sequence	Increments of 10 between the values 10 and 120

Appendix B. Representative samples of the train and synthetic datasets with HL7 tags removed.

Positive (train) messages

A) culture in progress. identifications performed by maldi tof mass spectrometry were developed and performance characteristics were determined by pcl alverno hammond in. salmonella species numerous. susceptibility not routinely performed. gastroenteritis due to non typhoidal salmonella spp. is generally self limiting in patients without underlying medical issues. for salmonella typhi isolates azithromycin is the drug of choice. identified by maldi tof mass spectrometry. sent to indiana state department of health.
B) additional organisms present as probable contaminants. salmonella species 100000 cfu/mlthis strain tested resistant to naladixic acid. treatment of extraintestinal salmonella infections may not be eradicated by fluoroquinolone treatment. therefore ciprofloxacin and levofloxacin are reported as resistant.

Negative (train) message

A) identifications performed by maldi tof mass spectrometry were developed and performance characteristics were determined by pcl alverno hammond in. no salmonella shigella aeromonas plesiomonas edwardsiella isolated. no predominant growth of klebsiella oxytoca present. one or more organisms were isolated and found to be normal flora through maldi tof mass spectrometry. campylobacter jejuni numerous. identified by maldi tof mass spectrometry. drugs of choice are ciprofloxicin erythromycin clindamycin tetracycline.
B) one or more organisms were isolated and found to be normal flora through definitive biochemical testing. no salmonella shigella plesiomonas edwardsiella or campylobacter isolated. no predominant growth of klebsiella

oxytoca present. aeromonas species moderate. susceptibility not routinely performed. aeromonas spp. are associated with gastrointestinal disease. symptoms are usually mild and self limiting. individuals with impaired immune systems or underlying malignancy are susceptible to more severe infection. antibiotics maybe indicated if symptoms are prolonged and in system icinfections. identified by maldi tof mass spectrometry.

Positive (synthetic) messages

A) client services present. moderate salmonella spp result progress called faxed to dr chang stroman at <time> on <date> by dr maritza office dr cahans office salmonella species sent to state lab salmonella group ser. jg salmonella spp
B) identifications performed by maldi tofmass spectrometry were developed and performance characteristics were determined by pcl alverno hammond in. salmonella species identified by maldi tof mass spectrometry. salmonella species sent to indiana indiana by felicia stipp [317 . e. coli o157 no campylobacter isolated sent to indiana state department of health reportable confirmed by indiana state dept. of further testing performed. salmonella performed performed on up please contact the laboratory if serotyping is required.

Negative (synthetic) messages

A) normal gi flora present no enteric pathogens isolated stool screened for salmonella shigella staphylococcusaureus campylobacter and sorbitol negative e. coli o157 this culture is a prior result no further 15d0662599 date called to difficile rn at dr. nichols office on <date>
B) test culture stool specimen type stool specimen received <date> <time> est final reports verified date/time <date> <time> final reports verified date/time <date> <time> no salmonella shigella species isolated no salmonella species no shigella species isolated no salmonella or shigella plesiomonas isolated. no shigella aeromonas plesiomonas edwardsiella or campylobacter isolated not routinely cultured is desired. no campylobacter specie scalled to indiana dept. of at <time>

Appendix C. List of top 20 features selected from the real and synthetic datasets using gini impurity scores.

Rank	Real (train) dataset	Synthetic dataset
1	Shigella	Salmonella
2	Salmonella	Speci
3	Speci	Shigella
4	Isol	Health
5	Campylobact	Campylobact
6	Health	Isol
7	Indiana	Sct
8	Group	State
9	Suscept	Group
10	Typhi	Confirm
11	Confirm	Chslb
12	Depart	Indiana
13	Ml	Suscept
14	Spp	Typhi
15	Call	Call
16	Cultur	Depart
17	Stool	Sent
18	Chslb	Test
19	Coli	Self
20	Enter	Cultur

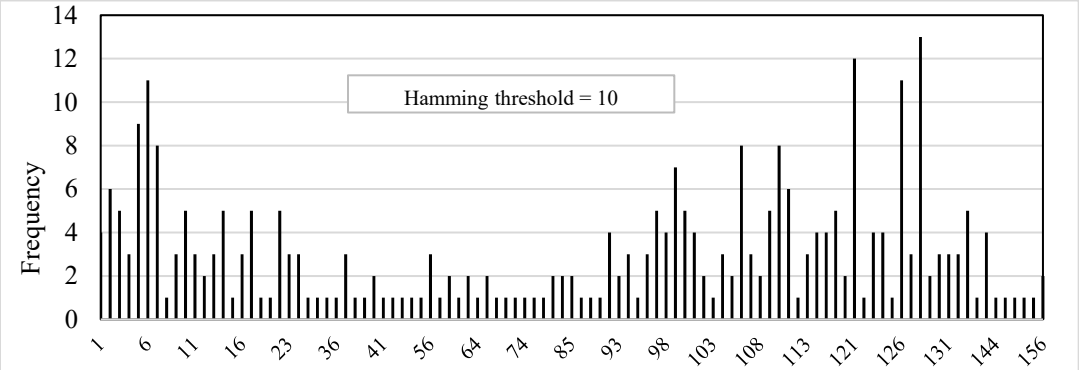
Appendix D. Intersection of top 5, 10, 15, 20, 50 and 100 features selected from the real and synthetic datasets using gini impurity scores.

Feature subset size	# features present in both datasets	List of features present in both datasets

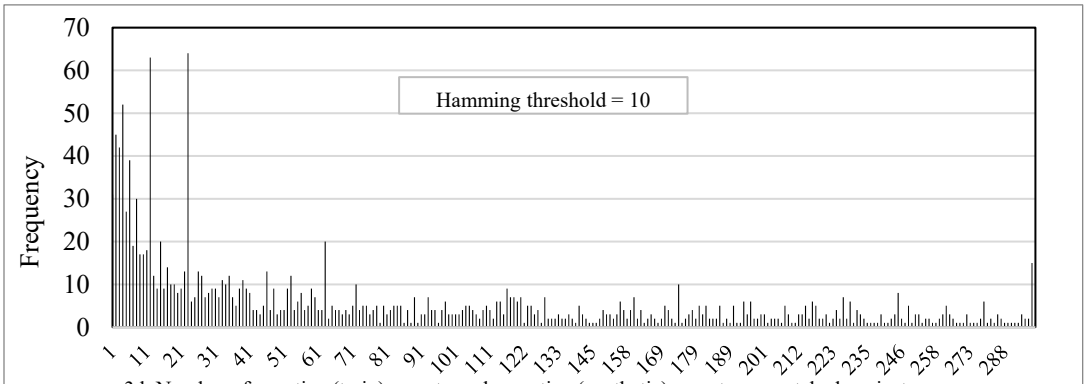
5	4 (80%)	salmonella, speci, shigella, campylobact
10	7 (70%)	speci, health, isol, group, salmonella, shigella, campylobact
15	12 (80%)	speci, indiana, campylobact, confirm, health, isol, group, suscept, typhi, salmonella, shigella, call
20	14 (70%)	chslb, speci, indiana, shigella, campylobact, confirm, health, isol, group, suscept, depart, salmonella, typhi, call
50	35 (70%)	sct, speci, non, due, isol, suscept, typhi, coli, report, chslb, call, indiana, cultur, confirm, diseas, issu, gener, azithromycin, sent, health, gastroenter, without, self, progress, depart, salmonella, stool, campylobact, enter, medic, test, final, group, shigella, tofmass
100	79 (79%)	perform, sct, present, speci, characterist, non, laboratori, due, isol, result, pathogen, infect, suscept, typhi, coli, identifi, report, chslb, serogroup, call, indiana, cultur, thi, lab, confirm, enzym, sourc, aerob, diseas, growth, issu, drug, gener, azithromycin, moder, maldi, specimen, sent, gastroenter, health, without, underli, serotyp, spectrometri, routin, self, toxin, numer, aeromona, follow, salmonella, progress, depart, stool, tofmass, ser, enter, campylobact, normal, develop, shiga, medic, patient, salsp, determin, test, mani, choic, mass, final, group, usual, see, board, date, shigella, access

Appendix E. Presence disclosure test

We computed frequency of a synthetic report matching with 1-to-n many real reports using a hamming score threshold of 10, which was selected based on its use in prior research(7). We *hypothesize*d that negative synthetic reports stood a much greater chance of matching with negative (train) reports because negative (train) reports tend to be similar to each other due to uniform text used to report a negative outcome. Thus, separate tests were performed against positive and negative datasets. We tabulated counts of how many positive reports were matched to each synthetic report. Next, we plotted the frequency of n synthetic reports matching with m training reports.



3.a Number of positive (train) reports each positive (synthetic) report was matched against



3.b Number of negative (train) reports each negative (synthetic) report was matched against

Figure 3. (3.a). Frequency of positive (synthetic) reports matched with positive (train) reports (hamming threshold <=10). (3.b). Frequency of negative (synthetic) reports matched with negative (train) reports irrespective of report status (hamming threshold <= 10).

We determined that re-identification risk is greater when,

a) A synthetic report is matched with a smaller number of real reports. Linking a synthetic report to a smaller number of real reports offer attackers a greater chance of pinpointing true matches via manual review. Re-identification risk falls as the number of real reports matched with a single synthetic report increases, as attackers must manually review each of these matches to pinpoint patients.
b) Synthetic reports are matched with real reports using smaller hamming cutoff thresholds. Smaller hamming distance thresholds indicate smaller differences between records, and thus, raises the likelihood that two matched reports are the same.

An evaluation of matches across a hamming distance of 10 presents that positive synthetic reports were matched to positive real reports (figures 3.a) at a lower rate than negative reports (3.b). As such, they poise significantly low chance of re-identification. We hypothesize that negative reports were matched with more certainty because they included boilerplate phrases reporting negative status. As anticipated, chances of matching a negative synthetic and real reports were larger than matching positive synthetic and real reports.

References

1. Callahan A, Shah NH. Machine Learning in Healthcare. Key Advances in Clinical Informatics: Elsevier; 2018. p. 279-91.
2. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. The American journal of gastroenterology. 2010;105(6):1224.
3. Hodge Jr JG, Gostin LO, Jacobson PD. Legal issues concerning electronic health information: privacy, quality, and liability. Jama. 1999;282(15):1466-71.
4. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC medical research methodology. 2010;10(1):70.
5. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017;24(3):596-606.
6. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. bmj. 2015;350:h1139.
7. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. arXiv preprint arXiv:170306490. 2017.
8. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PloS one. 2011;6(12):e28071.
9. McLachlan S, Dube K, Gallagher T, editors. Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. Healthcare Informatics (ICHI), 2016 IEEE International Conference on; 2016: IEEE.
10. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al., editors. Generative adversarial nets. Advances in neural information processing systems; 2014.
11. Camino R, Hammerschmidt C, State R. Generating Multi-Categorical Samples with Generative Adversarial Networks. arXiv preprint arXiv:180701202. 2018.
12. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H, editors. Synthetic data augmentation using GAN for improved liver lesion classification. Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on; 2018: IEEE.
13. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434. 2015.
14. Beaulieu-Jones BK, Wu ZS, Williams C, Greene CS. Privacy-preserving generative deep neural networks support clinical data sharing. BioRxiv. 2017:159756.
15. Zhu Y, Lu S, Zheng L, Guo J, Zhang W, Wang J, et al. Texygen: A Benchmarking Platform for Text Generation Models. arXiv preprint arXiv:180201886. 2018.
16. Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. arXiv preprint arXiv:170510843. 2017.

17. Xu J, Sun X, Ren X, Lin J, Wei B, Li W. DP-GAN: Diversity-Promoting Generative Adversarial Network for Generating Informative and Diversified Text. arXiv preprint arXiv:180201345. 2018.
18. Guo J, Lu S, Cai H, Zhang W, Yu Y, Wang J. Long text generation via adversarial training with leaked information. arXiv preprint arXiv:170908624. 2017.
19. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. Jama. 2014;311(24):2479-80.
20. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems. 2014;2(1):3.
21. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, et al. The Indiana network for patient care: a working local health information infrastructure. 2005;24(5):1214-20.
22. Yu L, Zhang W, Wang J, Yu Y, editors. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. AAAI; 2017.
23. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. 2016;60:145-52.
24. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. Journal of biomedical informatics. 2017;69:160-76.
25. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BGJJobi. A simple algorithm for identifying negated findings and diseases in discharge summaries. 2001;34(5):301-10.
26. Breiman L, Friedman J, Olshen R. Stone, cj (1984) classification and regression trees. Wadsworth, Belmont, California. 2009.
27. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
28. Kasthurirathne SN, Vest JR, Menachemi N, Halverson PK, Grannis SJ. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. Journal of the American Medical Informatics Association. 2017;25(1):47-53.
29. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. Journal of biomedical informatics. 2016;60:145-52.
30. Nergiz ME, Clifton CJIToK, Engineering D. δ-presence without complete world knowledge. 2010;22(6):868-83.
31. Shokri R, Stronati M, Song C, Shmatikov V, editors. Membership inference attacks against machine learning models. Security and Privacy (SP), 2017 IEEE Symposium on; 2017: IEEE.
32. Dwork C, McSherry F, Nissim K, Smith A, editors. Calibrating noise to sensitivity in private data analysis. Theory of cryptography conference; 2006: Springer.
33. Truex S, Liu L, Gursoy ME, Yu L, Wei W. Towards demystifying membership inference attacks. arXiv preprint arXiv:180709173. 2018.
34. Rahman MA, Rahman T, Laganiere R, Mohammed N, Wang Y. Membership Inference Attack against Differentially Private Deep Learning Model. Transactions on Data Privacy. 2018;11(1):61-79.
35. Backes M, Berrang P, Humbert M, Manoharan P, editors. Membership privacy in MicroRNA-based studies. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016: ACM.
36. Hamming RWJBStj. Error detecting and error correcting codes. 1950;29(2):147-60.
37. Papineni K, Roukos S, Ward T, Zhu W-J, editors. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics; 2002: Association for Computational Linguistics.
38. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:160908144. 2016.
39. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X, editors. Improved techniques for training gans. Advances in Neural Information Processing Systems; 2016.
40. Chen L, Dai S, Tao C, Zhang H, Gan Z, Shen D, et al., editors. Adversarial Text Generation via Feature-Mover's Distance. Advances in Neural Information Processing Systems; 2018.
41. Matwin S, Nin J, Sehatkar M, Szapiro T. A review of attribute disclosure control. Advanced Research in Data Privacy: Springer; 2015. p. 41-61.
42. Che T, Li Y, Zhang R, Hjelm RD, Li W, Song Y, et al. Maximum-likelihood augmented discrete generative adversarial networks. arXiv preprint arXiv:170207983. 2017.