Assessing Authenticity and Anonymity of Synthetic User-generated Content in the Medical Domain

Tomohiro Nishiyama, Lisa Raithel, Roland Roller, Pierre Zweigenbaum, Eiji Aramaki

Assessing Authenticity and Anonymity of Synthetic User-generated Content in the Medical Domain

Tomohiro Nishiyama[1*] **Lisa Raithel**[2,3,4*] **Roland Roller**[2]
Pierre Zweigenbaum[4] **Eiji Aramaki**[1]
[1]Nara Institute of Science and Technology
[2]German Research Center for Artificial Intelligence (DFKI)
[3]TU Berlin, BIFOLD [4]Université Paris-Saclay, CNRS, LISN
[1]{nishiyama.tomohiro.ns5,aramaki}@is.naist.jp [3]raithel@tu-berlin.de
[2]roland.roller@dfki.de [4]pz@lisn.fr

Abstract

Since medical text cannot be shared easily due to privacy concerns, synthetic data bears much potential for natural language processing applications. In the context of social media and user-generated messages about drug intake and adverse drug effects, this work presents different methods to examine the authenticity of synthetic text. We conclude that the generated tweets are untraceable and show enough authenticity from the medical point of view to be used as a replacement for a real Twitter corpus. However, original data might still be the preferred choice as they contain much more diversity.

1 Introduction

Medical text is difficult to share, even for research purposes, as it contains information about patients that might reveal an individual's identity. This makes natural language processing in that domain difficult. Moreover, there have been concerns about sharing even publicly available data from social media in recent years. This is partially due to legal reasons (e.g., X (Twitter)) but also due to privacy concerns. While data sensitivity can be at least addressed by de-identification (removal of personal health identifiers) and anonymization (irreversible removal of all information that possibly links back to an individual) (Meystre et al., 2010), privacy aspects constitute an additional barrier (see (Vakili et al., 2022; Volodina et al., 2023; Ben Cheikh Larbi et al., 2023)).

Synthetic data generation bears much potential and a way out of this misery, particularly with the rise of generative models. Various attempts within and outside the medical domain generate synthetic clinical data and show that large datasets can be easily generated and models trained on them can compete with models trained on real data (Ive et al., 2020; Libbi et al., 2021; Giuffrè and Shung, 2023). Apart from that, the use of generative data has advantages such as structural similarity, information relevance, and subjective assessment (Guillaudeux et al., 2023). Furthermore, synthetic dataset have been shown to be useful in, e.g., epidemiology research, medical education and training, and algorithm testing (Gonzales et al., 2023).

> ***Ex1:*** I've heard a lot about people going blind after inoculation, but the ears too. If you have pneumonia, you can recover instantly with steroids, but the eyes and ears...
>
> ***Ex2:*** I've been taking azathioprine for 2 days now and I feel like it's working really well. But the side effect is a rash all over my body..

Figure 1: Example of a source (top) and a pseudo (bottom) tweet translated into English. The original Japanese text can be found in the appendix.

For example, Choi et al. (2017) employ Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to generate (English) electronic health records (EHRs) while Abedi et al. (2022) synthesize tabular medical data such as laboratory values. Amin-Nejad et al. (2020) use GPT-2 (Radford et al., 2019) amongst other models to create datasets of discharge summaries in English. These data are then used as either pure training/fine-tuning material (Choi et al., 2017) or to augment existing resources (Amin-Nejad et al., 2020; Abedi et al., 2022), resulting in a better performance of the trained models when compared to the low-resource setup in which they are usually fine-tuned. Hiebel et al. (2023) report that linguistic phenomena are reproduced while privacy is preserved in their generated datasets of French clinical case reports. The usefulness of the authors' synthetic corpus is extrinsically investigated by fine-tuning models for a clinical named entity recognition task. The perfor-

* Equal contribution

mance of the models yields promising results.

Although the use of generated data in corpora is highly significant from the viewpoint that medical language resources are difficult to make public, few studies evaluate the anonymity or authenticity of the data, including medical aspects. Melamud and Shivade (2019) investigate the privacy-preserving characteristics and utility of synthetic EHRs by introducing a new measure based on Pointwise Differential Training Privacy (PDTP) (Long et al., 2017). Another work is provided by Mclachlan et al. (2018), who propose a framework to investigate the "realism" of synthetic EHRs. They compare the generated information with the rules, constraints, and concepts used in the original EHR data. Their approach, however, does not apply to unstructured user-generated texts.

In contrast to related work, this paper examines the authenticity of synthetic *user content* with respect to health-related topics in Japanese. More precisely, we examine whether artificially generated user tweets about potential adverse drug effects (ADE) are authentic and privacy-preserving and, therefore, might be a valuable alternative resource for future research.

2 Dataset

The baseline of this work is a synthetic corpus of Japanese tweets in the context of drug intake and symptoms (Wakamiya et al., 2023). The source data (original) was collected using 68 medication names as keywords from a Japanese drug-name dictionary[1] in the Twitter API. During preprocessing, URLs and user names were removed. Only tweets containing mentions of drugs and symptoms were kept in the data. T5 (Raffel et al., 2020), a transformer-based encoder-decoder model, was fine-tuned on these data and finally used to generate 10,000 tweets per medication name. The created texts were filtered and manually annotated with 22 different adverse drug reactions. More details about the synthetic data can be found in the appendix and in Wakamiya et al. (2023). In this paper, we are examining how authentic these synthetic data are. In the following, we distinguish between *source* tweets, i.e., the original data, and *pseudo*-tweets, i.e., the synthetics tweets.

[1] https://sociocom.naist.jp/hyakuyaku-dic/

3 Method

We analyze the data in different ways to measure the authenticity and validate the anonymity of the pseudo-tweets. First, we examine the source and the pseudo data on the word level and compare the vocabulary of both datasets. Next, we analyze if the distribution of our target events in the synthetic data is similar to that in the source data. Finally, we directly compare a subset of synthetic and source tweets manually as well as automatically with respect to *naturalness*, *comprehensibility*, *medical correctness* and *anonymity*.

3.1 Vocabulary

First, we compare the vocabulary of both corpora to analyze the diversity of the source and the pseudo data. Since there are considerably more source tweets (441,151) than pseudo-tweets (10,000), we sample 5 times 10,000 messages from the source tweets and compare each sample to the pseudo-tweets. To this end, we tokenize all tweets using spaCy[2] and report the number of tokens, types, and the mean lengths of source versus pseudo-tweets.

Additionally, we compare the similarity of original and pseudo tweets with the FAISS library[3]. All tweets (original and pseudo) are embedded using SentenceBERT (Reimers and Gurevych, 2019)[4] and compared using cosine similarity. Finally, we compute the type-token ratio (TTR) (Johnson, 1944) as a function of corpus size, and we check the frequency of part-of-speech tags.[5]

3.2 Analysis of ADEs

We compare the distribution of adverse drug effects in the pseudo data to their distribution/frequency in the real world. We compare the data to the Japanese Adverse Drug Event Report database (JADER) [6], which contains information about medications and ADEs. Since JADER reports every single ADE, the relative frequency of an ADE is calculated by dividing the number of reports for each adverse drug reaction pair by the total number of reports on the 22 ADEs for that drug. Using the frequency,

[2] https://spacy.io/api, version 3.7.2., model "ja_core_news_trf"

[3] https://github.com/facebookresearch/faiss

[4] https://huggingface.co/sonoisa/sentence-luke-japanese-base-lite

[5] More details on the corpus statistics can be found in Appendix B.

[6] https://www.pmda.go.jp/safety/info-services/drugs/adr-info/suspected-adr/0003.html (in Japanese)

we calculate Pearson's and Spearman's correlation coefficients for each drug individually and for all drugs globally. We also categorize ADEs into a more frequent (MFG) and a less frequent group (LFG) based on this frequency, for each drug individually and for all drugs globally, and compare MFG and LFG using a t-test. As the source data is not annotated, we draw only a comparison between pseudo data and world knowledge (JADER), but not to the source data. In addition, we examine whether we can find ADEs in the pseudo data that are unknown according to JADER.

3.3 Direct Comparison

Next, we directly compare the content of the source and pseudo-tweets. For this, we randomly select 100 source tweets and 100 pseudo-tweets. We conduct a manual and an automatic (GPT-4 (OpenAI, 2023)) analysis, giving the following questions to human annotators and GPT-4. Both of the human annotators are native Japanese speakers and medically trained.

Q1: "Do you think a human wrote this message?" (*naturalness*)
Q2: "Do you understand what the person wants to say with this message?" (*comprehensibility*)
Q3: "Is this message medically correct?" (*medical correctness*)
Q4: "Does the message contain any identifying information?" (*anonymity*)

Each question could only be answered with "yes" or "no". The human annotators were encouraged to answer quickly, i.e., without overthinking their response. Based on the responses, we calculated the inter-annotator agreement using Cohen's κ (Cohen, 1960).

4 Results

In the following section, we briefly present the results of our analyses. More details, particularly tables and figures, can be found in the appendix.

4.1 Vocabulary

For the pseudo-tweets, we count 441,022 tokens in total and on average 646,773 $\pm$ 1,568 tokens for the sampled source tweets. When comparing the number of types in the vocabulary, we find 6,499 different types in the pseudo data, whereas the source tweets exhibit 21,079 $\pm$ 121 types per random sample batch. Further, the mean length of pseudo-tweets is 44 (median is 44), while the source tweets have a mean length of 64 (median is 68). The results of the other statistics are summarized in Appendix B.

4.2 Analysis of ADEs

The comparison between the overall drug-ADE pairs is presented in Figure 2. The left figure indicates that according to the frequency in JADER, frequent ADE pairs in the pseudo data also occur more frequently than the less frequent ADE pairs. The right figure analyzes the single drug-ADE pairs in more detail. A deeper analysis, however, shows that we cannot find a correlation between the frequency of drug-ADE pairs in our pseudo data and their occurrence in the real world, as reported in JADER. The figure shows, for instance, that various drug-ADE pairs occur with a much higher frequency in the pseudo data than JADER. Conversely, we can observe some frequent drug-ADE pairs hardly occur in the pseudo data.

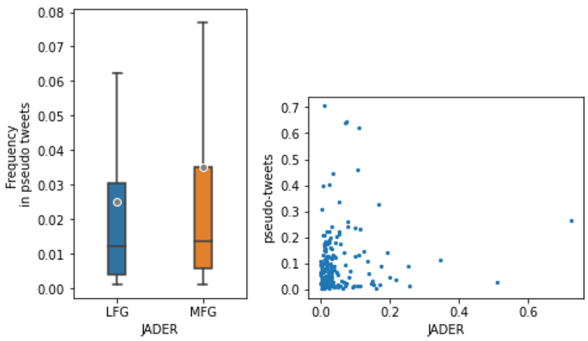

Figure 2: The frequency of ADEs from JADER and the pseudo data. (left): Comparison between MFG and LFG in the pseudo data. (right): Scatter plot between JADER and the pseudo data.

Figure 2 (right) does not show a strong association between JADER and the pseudo data, but the t-test result between MFG and LFG from all drug-ADE pairs indicated some association. When looking at the drugs individually, we found that only two of the drugs, *amiodarone* and *azathioprine*, were correlated with Pearson's and Spearman's correlation coefficient, respectively. Although the results of the t-tests in each drug showed no statistically significant differences, each of the means of MFG was greater than each of the means of LFG.

Next, we analyzed the drug-ADE pairs of the pseudo data. We found six pairs that were not listed in JADER, namely *azathioprine-constipation*, *amiodarone-insomnia*, *infliximab-insomnia*,

Ex3: Colchicine has been used for a long time and I have either constipation or diarrhea... I was told that if I get any side effects I can reduce it... but so far I've only had side effects.

Figure 3: Original tweet mentioning colchicine and constipation. (Translated from Japanese into English)

colchicine-asthma, *colchicine-constipation*, and *colchicine-hemorrhagic cystitis*. Of those six pairs, however, three could be found in the drug leaflets of the corresponding medications. For the remaining three, we cannot judge if this is correct from a medical perspective. Further analysis revealed that the pair *colchicine-constipation* at least occurred in the source data as shown in Figure 3, while the combinations *azathioprine-constipation* and *colchicine-asthma* did not.

4.3 Direct Comparison - Human

The human analysis shows a considerable disagreement between the two annotators on what can be considered a message written by a human (Q1). Normally, a higher Cohen's kappa closer to one is desirable, but in this result the Cohen's kappa closer to zero is desirable. The closer to zero, the better, because it means that the two human annotators are choosing more randomly which tweets are written by humans and which are generated by the model. Moreover, the results show that both annotators consider a slightly higher number of pseudo-tweets human-like than those from the source data.

Regarding the tweets' comprehensibility (Q2), most can be understood by both annotators. Again, there seems to be no major difference between pseudo and source tweets. Interestingly, both annotators agreed not to understand only eight pseudo and 13 source tweets.

Although our two annotators are medical experts, the results show a considerable disagreement (Cohen's kappa of 0.290) regarding which messages can be considered medically correct (Q3). However, there is a slight tendency towards source tweets being considered by both as medically correct (37 original versus 29 pseudo-tweets). The same applies to the joint agreement for medically incorrect tweets (23 versus 25).

Finally, regarding the anonymity of the data (Q4), the agreement of both annotators is very strong. Only up to four tweets (overlap of one tweet) were considered to contain identifying information. Notably, none of those four tweets were from the pseudo data. More details can be found in the appendix.

Ex4: Hanako, good evening... I couldn't tell him about my mental health... instead he gave me some Calonal because of a pressure headache...

Figure 4: Example of a part of a source tweet that contained a person's name, manually replaced here with 'Hanako' for publication.

4.4 Direct Comparison - Model

In contrast to the human analysis, GPT-4 only responded to the above-described questions for 198 tweets. Of those tweets, the model considered all messages human-generated and nearly all understandable (196/198). Moreover, 144 tweets were regarded as medically correct, of which a slightly larger portion came from the pseudo-tweets. The number of tweets considered medically correct by GPT-4, but as incorrect by both annotators, was 30. On the other hand, the number of messages considered medically incorrect by GPT but correct by both annotators was six. Finally, no message was considered by GPT-4 to be not anonymous.

5 Discussion

5.1 Vocabulary

Vocabulary inspection reveals a lower diversity of the pseudo-tweets compared to the source text messages, i.e., the source data generally contains more types and longer messages. The similarity comparison shows that given the pseudo-tweets and 4,000 source tweets, 1% of the pseudo-tweets are very similar to the original tweets, but not equal[7]. The generation process added content or reformulated the messages, leading to pseudo-tweets covering the same topics as the original tweets. The distribution of POS tags is similar in both datasets. Therefore, with respect to vocabulary, the pseudo data seems to be diverse, but not as diverse and creative as the source data. This aligns with research on the diversity of generated content (Chung et al., 2023) and might lead to an easy-to-learn dataset from which a machine-learning model cannot be generalized to other data.

5.2 Analysis of ADEs

Based on the investigated distribution of drug-ADE pairs, we conclude that the data is medically au-

[7]Except for one tweet, see details in Appendix B.

thentic to a certain degree. Further investigation by medical experts would be needed to arrive at a final conclusion.

5.3 Direct Comparison – Human Annotators

Naturalness A large number of pseudo-tweets were considered to be written by humans, whereas many source tweets were considered to be not written by humans. Moreover, the inter-annotator agreement on this task was very low (Cohen's kappa of 0.089). Therefore, we conclude that it is difficult to detect tweets written by humans and that our pseudo-tweets are sufficiently human-alike.

Comprehensibility Many tweets, even those written by humans, were not understood, and in fact, a larger percentage of the source tweets written by humans did not make sense to the annotators. This suggests that our pseudo-tweets are at least as comprehensible as the source tweets.

Medical correctness The annotations show that both annotators considered more source tweets medically correct. On the other hand, the annotators also show a strong disagreement with many tweets. Therefore, it is difficult to conclude that source data might be medically more accurate than synthetic data. Conversely, we can see a similar distribution of messages labeled as medically incorrect by both annotators (source=23; pseudo=25). In other words, this means that one out of four messages is medically incorrect. Although our subset was randomly sampled, this shows a concerning tendency and raises concerns about health-related information from social media.

Anonymity Most tweets did not include identifying information, as critical information and messages were filtered out beforehand. Interestingly, the only messages considered problematic regarding anonymity were still from the source data, not the pseudo data, as shown in Figure 4. However, we cannot guarantee that pseudo-tweets per se do not include identifying information, but we believe that removing critical information before training a generative model helps.

5.4 Direct Comparison – Model

While the question about comprehension might be too abstract for GPT-4, it fails to identify messages with identifying information. Moreover, regarding medical correctness, the model identified multiple tweets as correct, which, on the other hand, were labeled by both humans as incorrect and vice versa. Finally, regarding the differentiation between human-generated and synthetic tweets, GPT-4 and humans come to a similar conclusion: they are difficult to differentiate. However, GPT-4 is too optimistic and assigns all messages to human-alike.

6 Conclusion

In this work, we analyzed synthetically-generated tweets in the context of drug intake and adverse drug reactions. The data was compared to (real) user-generated messages regarding authenticity, privacy preservation, and medical correctness. The results show that the synthetic data has characteristics similar to the source data. From a linguistic point of view, the data shows less variation, but it contains a similar number of data with questionable medical correctness (as the original), and has a similar authenticity. In addition to that, pseudo data could serve as a "safety net" as it might be less likely to provide identifiable information. Finally, we believe that the findings are generally valid for different languages; however, larger and more complex models than T5 might increase the authenticity and correctness level but might easily reproduce sensitive information it has seen during training.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback on our paper. Our work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425, and by JPMJCR20G9, ANR-20-IADJ-0005-01, and DFG-442445488 under the trilateral ANR-DFG-JST AI Research project KEEPHA. Furthermore, we gratefully acknowledge funding from the German Federal Ministry of Education and Research under the grant BIFOLD24B.

Limitations and Ethical Considerations

JADER has a reporting bias because some of the reports are voluntary, which may have affected the results. The study targets generated Japanese social media messages. However, most analyses should also apply to other languages, especially those in the original corpus (Wakamiya et al., 2023). We recognize that the above data analysis is domain-specific, but similar tests could also be conducted for other areas.

Regarding ethical considerations, the following three methods were implemented to avoid privacy issues in the original Twitter data: Deleting the usernames in training data for the model, deleting the exact duplicates in the generated text from the source, and, with manual work of annotators, checking all of the synthetic data and making sure no identifying information remains. Models using original data were trained locally.

We further acknowledge that questions used to judge tweets with GPT-4 and the corresponding responses are (1) not reproducible as soon as an updated version of GPT-4 is released, and (2) might result in different responses when the questions are slightly modified or set up differently.

Finally, to assess the authenticity and diversity of the data, many more linguistic measures could be applied. This paper only presents a few as a complement to the medically inspired investigations.

References

Masoud Abedi, Lars Hempel, Sina Sadeghi, and Toralf Kirsten. 2022. Gan-based approaches for generating structured data in the medical domain. *Applied Sciences*, 12(14).

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

Iyadh Ben Cheikh Larbi, Aljoscha Burchardt, and Roland Roller. 2023. Clinical Text Anonymization, its Influence on Downstream NLP Tasks and the Risk of Re-Identification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 105–111, Dubrovnik, Croatia. Association for Computational Linguistics.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mauro Giuffrè and Dennis L. Shung. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6(1):186.

Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith. 2023. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny, and Pierre-Antoine Gourraud. 2023. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *npj Digital Medicine*, 6(1):37.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for Natural Language Processing. *npj Digital Medicine*, 3(1):69.

W. Johnson. 1944. *Studies in Language Behavior*. Psychological Monographs. American Psychological Association.

Claudia Alessandra Libbi, Jan Trienes, Dolf Trieschnigg, and Christin Seifert. 2021. Generating synthetic training data for supervised de-identification of electronic health records. *Future Internet*, 13(5).

Yunhui Long, Vincent Bindschaedler, and Carl A. Gunter. 2017. Towards measuring membership privacy. *CoRR*, abs/1712.09136.

S. Mclachlan, K. Dube, T. Gallagher, B. Daley, and J. Walonoski. 2018. The ATEN Framework for Creating the Realistic Synthetic Electronic Health Record. *Technologies (BIOSTEC 2018), 11th International Joint Conference on Biomedical Engineering Systems*.

Oren Melamud and Chaitanya Shivade. 2019. Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1–16.

OpenAI. 2023. GPT-4 Technical Report. Publisher: arXiv Version Number: 3.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Elena Volodina, Simon Dobnik, Therese Lindströ m Tiedemann, and Xuan-Son Vu. 2023. Grandma karl is 27 years old – research agenda for pseudonymization of research data. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 229–233.

Shoko Wakamiya, Lis Kanashiro Pereira, Lisa Raithel, Hui-Syuan Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17. National Institute of Informatics (NII)*.

A Appendix

A.1 Corpus Generation

The synthetic data creation consists of two steps, data generation and pre-processing. First, Japanese tweets were collected from Twitter (X), using 68 drug queries extracted from a Japanese drug-name dictionary[8] and the public Twitter API[9]. The text generation model was built from the collected tweets to produce Japanese pseudo tweets. URLs and user names in the original tweets were replaced with masks. Using a Japanese medical named entity recognizer, MedNER-CR-JA[10], tweets without any symptom expression were filtered out. T5 was fine-tuned on the remaining tweets to generate synthetic tweets mentioning a subset of 17 drugs.

During post-processing, the following tweets were filtered out; (i) pseudo-messages that do not mention any drug or symptom, (ii) pseudo-messages that are identical to any of the original tweets, and (iii) duplicates.

Finally, all tweets mentioning any of the 17 drugs were annotated manually. After counting the number of annotations describing positive ADE mentions, the 24 most frequent ones were chosen. In two cases, two similar ADEs were merged into one. Then, 22 ADEs were obtained as labels. More details can be found in Wakamiya et al. (2023).

A.2 Tables and Figures about analysis of ADEs and human comparison

Tables 1 and 2 and Figure 6 present the detailed results of the direct comparison of source and pseudo data, analyzed by human annotators and GPT-4. Figure 5 presents the detailed distribution of the drugs and their ADE in the pseudo data compared to JADER. Table 3 and Figures 10 and 11 show a detailed overview of the data's drug-ADE correlation. Finally, Figure 8 presents all example tweets from above in the original language (Japanese).

B Details on Corpus Statistics

The following will give more details on the corpus statistics we used to compare the original and pseudo-tweets. This is not exhaustive; there are many more interesting analyses that can be applied to the data.

[8] https://sociocom.naist.jp/hyakuyaku-dic/

[9] https://developer.twitter.com/en/support/twitter-api

[10] https://huggingface.co/sociocom/MedNER-CR-JA

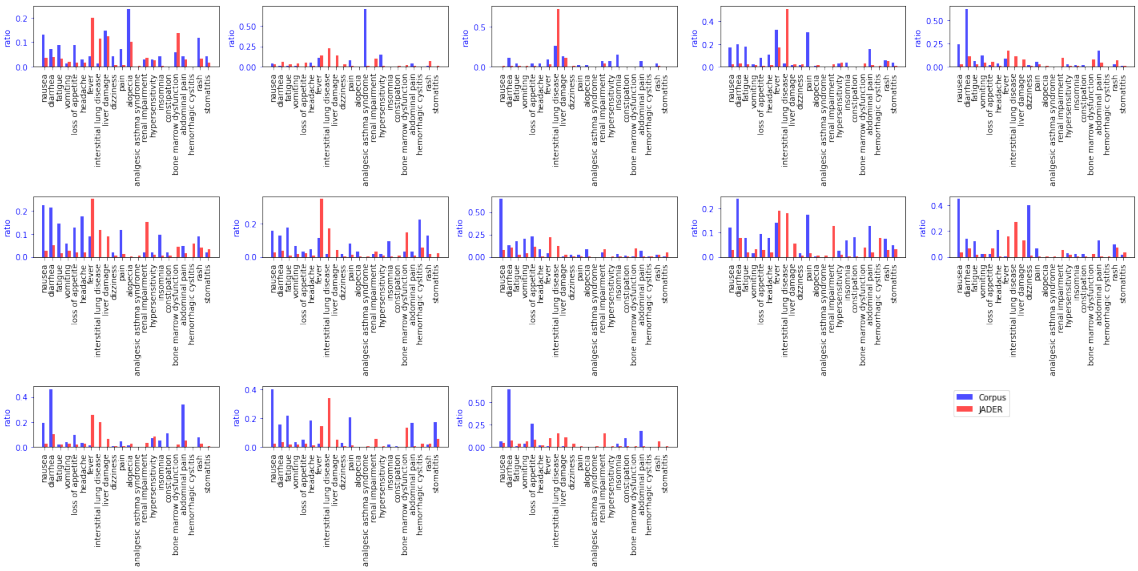

Figure 5: Distributions between JADER and the corpus

Q1		A1	
		yes	no
A2	yes	71 (31 / 40)	15 (7 / 8)
	no	83 (41 / 42)	31 (21 / 10)
Cohen's kappa		0.089	
Q2		A1	
		yes	no
A2	yes	126 (63 / 63)	38 (15 / 23)
	no	15 (9 / 6)	21 (13 / 8)
Cohen's kappa		0.281	
Q3		A1	
		yes	no
A2	yes	66 (37 / 29)	24 (7 / 17)
	no	48 (23 / 25)	48 (23 / 25)
Cohen's kappa		0.290	
Q4		A1	
		yes	no
A2	yes	1 (1 / 0)	1 (1 / 0)
	no	2 (2 / 0)	196 (96 / 100)
Cohen's kappa		0.393	

Table 1: Results from human judgment by annotator1 (A1) and annotator2 (A2).*The numbers are counts of original + pseudo (original / pseudo)

	GPT-4 Answer		Cohen's kappa	
	yes	no	A1	A2
Q1	198 (98 / 100)	0 (0 / 0)	0.000	0.000
Q2	196 (97 / 99)	2 (1 / 1)	0.088	0.014
Q3	144 (67 / 77)	54 (31 / 23)	0.237	0.298
Q4	0 (0 / 0)	198 (98 / 100)	0.000	0.000

Table 2: Results from model judgment by GPT-4

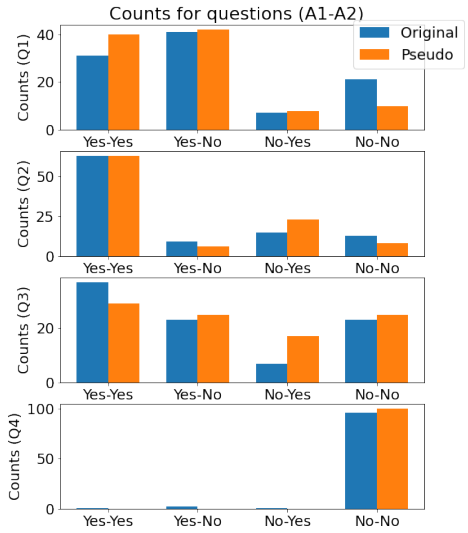

Figure 6: Results from human judgment by A1 and A2 (Barplots of Table 1) .

B.1 Statistics

Type-Token Ratio The type-token ratio (TTR) (Johnson, 1944) counts the number of types and divides the result by the number of tokens as a measure of diversity in a corpus. However, this ratio strongly depends on the corpus size, and therefore, it is often shown as a function of the corpus size.

Part-of-Speech Tags We further calculate the relative frequencies of the occurring POS tags in the data using spaCy for tagging.

Similarity We index the pseudo-tweets with the FAISS library and compare them using cosine similarity with a sample from the original data. This

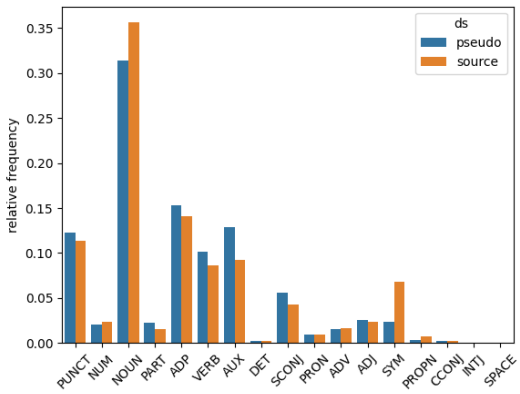

Figure 7: The relative frequency of the POS tags for the pseudo- and source tweets.

Ex1: 接種後失明は良く聞くけど耳もなんだ肺炎だとステロイドで即回復だけど目や耳って

Ex2: アザチオプリンを飲み始めて2日目だけどめっちゃ効いてる気がする。でも副作用で全身の発疹が凄い..

Ex3: コルヒチンは昔から使われてるし、便秘か下痢のどっちかかな…副作用出たら減らしてもいいから、次の病院まで飲み続けてって言われたんだけど、副作用しか今のとこないんだけど

Ex4: はなこちゃん，こんばんは... 精神的なことは伝えられずに終わりました～，そのかわり気圧頭痛が酷くてカロナール出して…

Figure 8: Japanese version of examples Ex1–Ex4. Ex1: source tweet. Ex2: pseudo tweet. Ex3: tweet mentioning colchicine and constipation. Ex4 where a person's name remained in the tweet (manually replaced here with 'はなこ' for publication).

sample contained only 4,000 original tweets since the computation was time-consuming.

B.2 Results and Discussion

Type-Token Ratio In Figure 9, we show the TTR for both datasets (the first 20,000 tokens), plotted against the corpus size. The source tweets clearly show a higher type-token ratio which decreases slower than the ratio of the pseudo-tweets.

Part-of-Speech Tags We show the relative frequencies of the occurring POS tags in Figure 7. The pseudo-tweets get tagged with 15 different POS tags, while the original data gets 16 POS tags. Nouns (NOUN), adpositions (ADP), auxiliaries

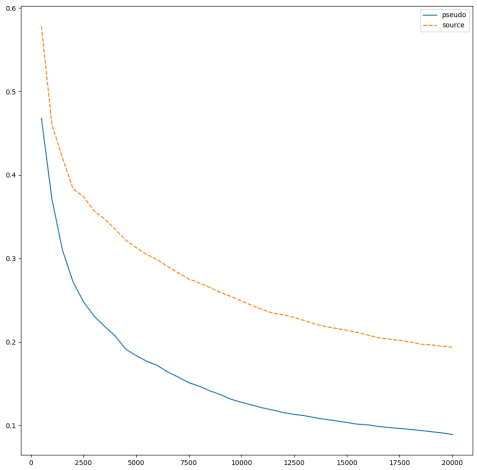

Figure 9: The type-token ratio as a function of corpus size for the source (orange) and pseudo data (blue) for an excerpt of the data.

	Pearson		Spearman		KS
drug	CC	p-value	CC	p-value	p-value
azathioprine	0.316	0.153	0.508	0.016	0.049
aspirin	-0.154	0.493	-0.014	0.951	0.007
amiodarone	0.762	0.000	0.391	0.072	0.020
infliximab	0.040	0.859	0.176	0.435	0.109
colchicine	0.314	0.154	0.205	0.359	0.632
cyclosporine	-0.096	0.670	-0.072	0.750	0.394
cyclophosphamide	0.110	0.627	0.205	0.361	0.109
cisplatin	0.184	0.411	0.269	0.226	0.872
tacrolimus	0.025	0.911	-0.137	0.542	0.394
minocycline	-0.212	0.343	-0.064	0.776	0.218
mesalazine	0.104	0.646	0.102	0.652	0.632
methotrexate	-0.215	0.336	-0.086	0.705	0.109
metformin	0.107	0.635	0.210	0.349	0.007
all drugs	0.088	0.140	0.126	0.034	-

Table 3: Pearson's and Spearman's correlation coefficient and p-values of the tests in each drug

(AUX) and punctuation markers (PUNCT)[11] are the most common POS tags for both corpora.

Similarity From the 4,000 samples we compared to the pseudo-tweets, we retrieved 86 hits that showed a cosine similarity higher than 0.9 and one that was exactly the same. However, from the 86 hits, only 39 were unique, i.e., one pseudo-tweet can have several very similar, but not exact nearest neighbors. The single pseudo-tweet that was identical to the source tweet did not contain any identifiable information and was basically a sequence of hashtags. However, this shows that even though the generation process included a diversity penalty, synthetic data might still be repetitive or near-repetitive.

[11] https://universaldependencies.org/u/pos/

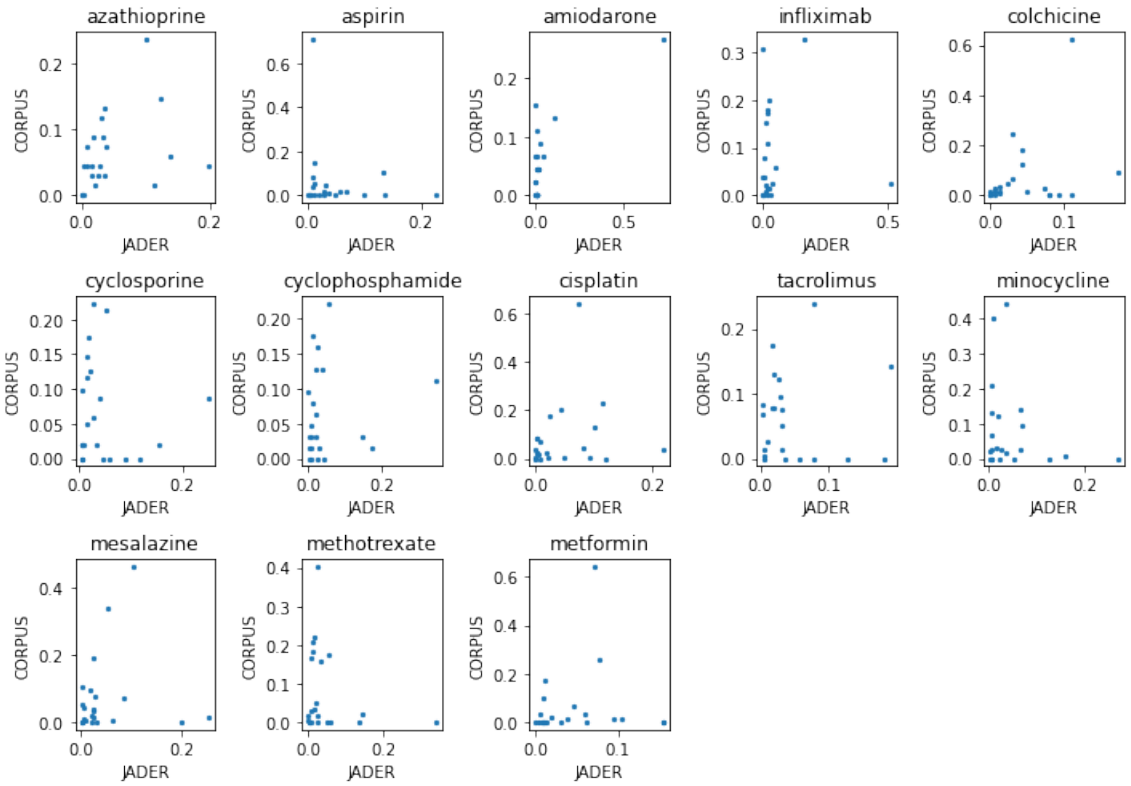

Figure 10: The frequency between JADER and the corpus in each drug

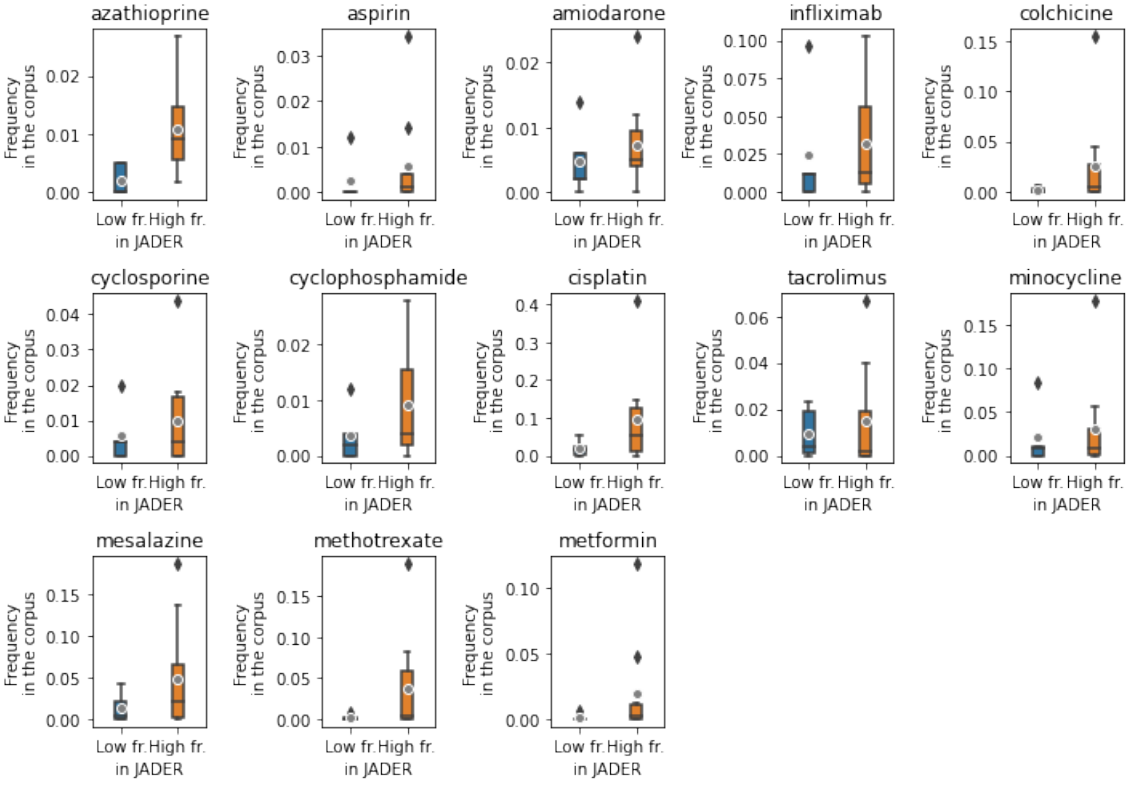

Figure 11: Comparison between MFG and LFG in each drug