Privacy-Preserving Generation of Synthetic Pathology Reports for Information Extraction

Alejandra Lorenzo, Adrien Coulet, Claire Gardent

Privacy-Preserving Generation of Synthetic Pathology Reports for Information Extraction

Alejandra Lorenzo[1,2], Adrien Coulet[3], and Claire Gardent[1]

[1] CNRS, Université de Lorraine, Loria UMR7503, Nancy, France
Alejandra.Abdelnur@loria.fr, Claire.Gardent@loria.fr
[2] U2R, Unité de Radiothérapie République, Clermont-Ferrand, France
[3] Inria, Inserm, Université Paris Cité, HeKA UMR1346, Paris, France
Adrien.Coulet@inria.fr

Abstract. A long-standing goal in clinical NLP is extracting key variables from clinical text, but progress is hindered by domain shift, limited annotated data, and privacy constraints. In this paper, we propose a novel privacy-preserving method to generate synthetic report/data pairs that support information extraction by associating LLM-generated pathology reports with thirteen variables commonly found in real reports for breast cancer patients. We first generate synthetic tabular data for these variables and their possible values, comparing several tabular synthesizers and selecting PATE-CTGAN for its strong statistical fidelity and differential privacy guarantees. We then generate pathology reports using three LLMs to maximize linguistic diversity and conditioning generation on synthetic variable–value sets.
We show that a model fine-tuned on the resulting synthetic data substantially outperforms the same model without fine-tuning and improves performance over a 3-shot baseline using synthetic in-context examples, achieving 0.79 accuracy compared to 0.38 and 0.64, respectively. These results show that high-quality synthetic data can effectively compensate for limited annotated clinical data while enabling accurate and privacy-preserving clinical information extraction. The code is available at https://gitlab.inria.fr/aabdelnu/syntheticpreports.

Keywords: Clinical text · Information Extraction · Synthetic data · Differential privacy

1 Introduction

Cancer diagnosis and treatment rely on the accurate interpretation of pathology reports, which contain essential information about tumor location, grade, morphology, and classification. Extracting this information to populate structured diagnosis forms is critical for clinical decision-making, yet manual extraction is time-consuming and error-prone. Large language models (LLMs) offer a promising avenue for automating this process, but their effectiveness depends on access to substantial annotated data and the use of real clinical text is constrained by patient privacy regulations, leading to a scarcity of training data—an issue that

is particularly acute for languages with fewer resources than English [13]. Synthetic data generation provides a potential solution by enabling the creation of datasets that retain the statistical properties of real data while reducing privacy risks. Thus [12] generate table/text pairs by first, generating tabular data using a Bayesian network whose dependencies are defined by domain-experts and second, using this data as input to an LLM to generate clinical texts. Similarly, [15] automatically extracts table/text pairs from pdf files and uses these to train a table-to-text summarisation system.

In this work, we propose a novel privacy-preserving method to generate synthetic data for information extraction. This method consists in first creating realistic clinical tabular data and second, converting it into matching French reports using LLMs. We show that fine-tuning an LLM on these pairs significantly outperforms both zero- and few-shot baselines.

2 Generating Synthetic Tabular Data

Based on 3,880 breast cancer patient records, we derive synthetic data for thirteen clinically relevant variables (e.g., diagnosis, size and position of the tumor, Estrogen Receptor Status) in three main steps: discretization, comparison of four existing synthesizers and application of the selected synthesizer to the cancer data. Patient data were extracted from four French clinics belonging to the U2R group. Patient information and data pseudonymisation were performed in accordance with French regulation and the GDPR.

2.1 Discretization

First, we discretize variable values in order to aggregate fine-grained variables into clinically meaningful and broader categories. Categorical variables are consolidated into broader classes (e.g., *tumorectomie avec curage axillaire* → *tumorectomie*), while numerical variables are discretized into clinically meaningful ranges (e.g., tumor size (mm) → 0, 1–5, 6–10, 11–20, 21–50, 50+ mm). Discretization helps strengthen privacy protection as exact numerical values can make individuals more easily identifiable, especially for rare or extreme measurements. By grouping values into intervals, we reduce data granularity, which lowers the risk of linkage attacks based on unique numerical patterns, and we increase the size of equivalence classes (more patients share the same combination of variables) again reducing re-identification probability. Discretization was also shown to facilitate and stabilise synthetic data generation. In particular, differentially private tabular synthesizers, like MST or PATE-CTGAN, were shown to handle categorical distributions more effectively than continuous variables with wide or heavy-tailed ranges and [5] showed that optimizing the discretizer and number of bins improves utility.

2.2 Comparing synthesizers

We compare four existing synthesizers from two libraries, Synthetic Data Vault (SDV) [11] and SmartNoise [8]: a probabilistic model based on copulas (Gaussian Copulas from SDV), a generative adversarial network adapted to tabular data (CTGAN from SDV), and two variants that incorporate differential privacy guarantees (PATE-CTGAN, MST, from SmartNoise). The *Gaussian Copula Synthesizer* [11] learns marginal distributions, transforms them to a normal scale, estimates the covariance matrix, and samples synthetic rows from the resulting multivariate Gaussian before mapping them back to the original domain. *CTGAN* [16] is a conditional generative adversarial network designed specifically for tabular data, which extends the standard GAN framework with three innovations: a conditional generator to better model rare categorical values, Mode-Specific Normalization using Gaussian Mixture Models for continuous data, and conditional sampling to balance discrete categories. *PATE-CTGAN* [6] takes the state-of-the-art CTGAN tabular data synthesizer and applies PATE (Private Aggregation of Teacher Ensemble), to ensure Differential Privacy. Finally, *Maximum Spanning Tree (MST)* [10] is a simple probabilistic tabular data synthesizer that models dependencies between variables using a maximum spanning tree. While the Gaussian Copula preserves correlations and CTGAN captures complex non-linear relationships, both can still reproduce rare combinations, posing re-identification risks. Because of these risks, stronger guarantees are needed. Differential Privacy (DP) [4] provides a rigorous, mathematically provable standard for privacy protection. Under DP, the risk of any individual being identified is nearly the same whether or not their data is included in the dataset, making it resistant to membership inference and reconstruction attacks. This is particularly important when dealing with high-dimensional tabular data, where correlations and rare variable combinations can inadvertently expose individuals.

Evaluation Metrics. To determine which synthesizer best captures data distribution while preserving privacy, we evaluate the synthesized data along three main dimensions: fidelity, similarity and novelty.

Fidelity assesses the synthetic data's ability to reproduce the statistical distributions observed in the real data, both at the marginal level for individual variables and at the level of inter-variable relationships. We evaluate fidelity using *Accuracy* [14], the agreement between real and synthetic data, measured across univariate (single column), bivariate (column pairs), and trivariate (column triplets) distributions via averaged L1 distance.

Similarity [14] assesses whether the synthetic data preserves the overall structure of the real data within a learned representation space, without aiming for a one-to-one correspondence at the individual record level. First, the synthetic and training instances are represented as vectors in an embedding space, then the cosine distance between the centroid of each dataset is computed.

Finally, novelty focuses on privacy preservation. Here, the goal is to detect whether certain synthetic records are excessively close to real individuals. We use four metrics for novelty: *the proportion of identical matches between syn-*

thetic and training data, NNDR (Nearest Neighbor Distance Ratio), DCR share (Distance to Closest Record share) [14] and DCR Overfitting Score [3]. *NNDR* compares how close a synthetic record is to its nearest training *vs.* holdout record, where values near 1 mean balanced, below 1 suggest overfitting, and above 1 indicate high novelty. *DCR share* shows the proportion of synthetic records that are closer to some training record than to any holdout record. If this share is too large, it suggests overfitting / memorization. Ideally the synthetic data should behave similarly to holdout data in this respect. Lastly, *DCR Overfitting Score* evaluates whether synthetic samples are excessively close to the real (training) data compared to unseen holdout data. It computes a normalized ratio between 0 and 1, where lower scores (close to 0) indicates that a large fraction of synthetic samples are closer to real training data, suggesting possible overfitting or memorization, whereas higher scores (approaching 1) reflect greater novelty and privacy preservation.

Hyper-Parameters. We apply our synthesizers to 3,880 sets of variable-values pairs associated with anatomical pathology reports written by physicians for breast cancer patients. For the novelty metrics, this data was split 50/50 into training and holdout sets. For each synthesizer, ten independent synthetic datasets were generated and the resulting datasets evaluated, giving ten measurements per synthesizer. We then average the ten results to get the final score. The privacy budget ϵ was empirically evaluated across multiple values (from 2 to 5), and ϵ=3 was selected as it provides the best balance between data utility and privacy risk, consistent with findings in [1].

Table 1. Comparing the four Synthesisers.

Metric \ Synthesizer	**Gaussian Copula**	**CTGAN**	**PATE-GAN**	**MST**
Fidelity				
Overall Accuracy	86.59%	74.57%	82.56%	**92.22%**
Accuracy - Univariate	96.57%	86.68%	93.46%	**98.36%**
Accuracy - Bivariate	86.46%	74.17%	82.34%	**92.22%**
Accuracy - Trivariate	76.72%	62.86%	71.87%	**86.07%**
Similarity				
Cosine Similarity	0.984	0.886	0.973	**0.999**
Novelty				
↓ % Identical Matches	5.85%	**1.76%**	3.68%	17.62%
↑ NNDR	0.961	0.560	**1.080**	1.189
↓ DCR – Share	55.10%	56.47%	54.90%	**52.36%**
DCR Overfitting Score	**1.0**	**1.0**	**1.0**	**1.0**

Results and synthesizer selection. Following [7], who emphasize that the optimal synthetic data generator is not necessarily the one with the highest raw accuracy,

but rather the one that achieves the best trade-off between fidelity and privacy protection, we aim to strike a balance between realism and privacy: generating synthetic data that is sufficiently faithful to support the training of performant models, while remaining sufficiently distinct from the original data to ensure patient protection.

Table 1 reports the results for the four synthesizers. They show that CTGAN has the lowest and MST the highest fidelity. However, MST also produces the largest number of identical matches, indicating a higher risk of memorization and lower novelty. Gaussian Copula and PATE-CTGAN exhibit comparable fidelity, but PATE-CTGAN yields fewer identical matches. Balancing fidelity, and novelty, we therefore select PATE-CTGAN for synthetic medical data generation.

2.3 Generating Synthetic sets of Variable-Value Pairs

Using the PATE-CTGAN synthesizer, we generate 4,000 synthetic data instances (sets of 13 variable-value pairs) and select a representative subset of 1,000 instances to compare the outputs of three different LLMs while maintaining textual diversity and reducing generation and evaluation costs. To obtain this representative sample, we performed a stratified, frequency-weighted subsampling. First, we quantified data completeness by counting non-missing (non-"unknown") variables per instance and used this measure to define quartiles. Within each quartile, we examined the joint distribution of two clinically relevant variables namely, MORPHOLOGY and NUMBER OF LYMPH NODES REMOVED. For each quartile, we sampled 60 examples from the two most frequent variable combinations, 30 examples from the next two most frequent combinations, and 70 examples from all remaining combinations. This yielded 250 samples per quartile, totaling 1,000 records that preserve both category structure and completeness heterogeneity.

3 Generating Synthetic Pathology Reports

This second step uses LLMs to generate synthetic textual pathology reports conditioned on the previously generated variable-value pairs.

Data Preprocessing and Input Structuring The tabular data is first transformed into a format suitable for prompting an LLM. This preprocessing is particularly important for variable values previously discretized as numerical ranges. To ensure clinical plausibility while allowing narrative flexibility, these ranges are converted into descriptive expressions within the prompt. For instance, the tabular entry *lymph nodes examined: 6–12*, is transformed into the French version of *Number of lymph nodes examined: to be selected within the range of 6 to 12, depending on clinical consistency.* In this way, the LLM is informed of the possible values but is left free to choose within the given range.

Although the "diagnostic test" variable is mandatory in pathology reports, we observed cases in both real and synthetic data where it was missing. To address this issue, we apply the following expert rule: if the test is not specified but

the instance includes a tumor size value, the "diagnostic test" is set to "breast lumpectomy" (the most frequent diagnostic test with "tumor size" variable); otherwise, it is set to "biopsy" (the most frequent diagnostic test when no tumor size is indicated). We also observed that "hormonal receptors" values were systematically generated as either negative or positive whereas pathology reports also report a percentage of positivity when this variable is positive. To reproduce similar values, we provide the LLM with a random value drawn from a range of valid positivity percentages.

Models We select three LLMs for text generation based on their various expected strengths: Mixtral-8×7B for strong French generation, MedGemma-27B for medical-domain specialization, and Qwen3-30B-A3B as a large, instruction-tuned generalist model. For all three models, inference is performed using deterministic greedy decoding with temperature 0.

Evaluation As reference texts are unavailable, standard reference-based metrics such as BLEU or ROUGE cannot be used. Hence we adopt an LLM-as-Judge evaluation protocol using two LLMs (GPT-5.1 and Gemini-2.5-Pro) to assess both the faithfulness between texts and variable-value pairs and the global consistency of the document. For faithfulness, the LLM judge outputs, for each input variable-value pair, a JSON object with two fields: the field 'correct (yes/no)', indicates whether the text correctly mentions the input variable–value; the field 'reference', provides the supporting text span when applicable. For consistency, we prompt the LLMs to detect explicit contradictions for key clinical variables (e.g., tumor size, receptor status, grade, margins, lymph nodes) and assign each text a single consistency label out of the three possible ones {PASS, FAIL, or UNKNOWN}.

We assess the reliability of the LLM judges by comparing their predictions against human judgments. Two clinical experts were asked to manually annotate 50 synthetic pathology reports with their matching attribute-value pairs.The Kappa [2] value between the two human annotators reaches 0.90 for faithfulness (almost perfect) and 0.67 for consistency. For Humans–LLM agreement, we use Krippendorff's α [9], which is suitable for multiple annotators ($n = 3$ in our case). For both LLMs, the results show high consistency with human annotators for faithfulness α scores of 0.90 for GPT 5.1 and 0.82 for Gemini-2.5-Pro. Results are lower for consistency: 0.53 and 0.59, respectively. Based on these results, we used GPT 5.1 for faithfulness evaluation and Gemini-2.5-Pro for consistency.

We also assess the distribution gap between synthetic and real reports using lexical richness, repetition, entropy, compression, and structural-formatting metrics.

Results. Table 2 presents the faithfulness and consistency evaluations for the three LLMs. For each clinical variable, the table reports the per variable accuracy, defined as the proportion of variable-value pairs correctly mentioned in the generated report relative to the input data. The table also reports the overall accuracy, computed as the average across all variables, and the consistency

score, defined as the proportion of reports assigned a "PASS" value by the consistency evaluation. The overall accuracy across all variables is relatively high, with faithfulness ranging from 0.747 to 0.888 and consistency ranging from 0.792 to 0.929. The medical LLM MedGemma performs well in terms of both faithfulness (0.878) and consistency (0.929). For Qwen3, the lower accuracy on the *diagnostic_test* variable can be explained by the frequent use of generic report titles that do not clearly specify the procedure type and may be incompatible with biopsy reports.

Table 3 shows the distribution of document-level accuracy for each LLM. For each document, accuracy is defined as the proportion of clinical variables that exactly match the reference. Documents are then grouped into bins according to this proportion, and the table reports the percentage of documents falling into each bin. MedGemma and Mixtral achieve moderate to high document level accuracy for 92.9% of the documents.

Finally, the distribution gap analysis shows that real reports are longer, more lexically diverse, and more structurally segmented, whereas synthetic reports are shorter, more repetitive, more predictable, and less faithful to clinical formatting conventions. Details are provided in the code repository.

Table 2. Accuracy and Consistency of Synthetic Reports

Variable	**Qwen3**	**MedGemma**	**Mixtral**
diagnostic_test	0.125	1.000	0.994
number_lymph_nodes_removed	0.840	0.872	0.851
number_lymph_nodes_affected	0.519	0.871	0.728
extracapsular_tumor_spread	0.848	0.884	0.925
tumor_size	0.483	0.709	0.840
estrogen_receptor_status	0.961	0.974	0.957
progesterone_receptor_status	0.960	0.995	0.977
lymphovascular_invasion	0.405	0.712	0.605
her2_status	0.903	0.991	0.992
clear_margins	0.943	0.552	0.731
Ki67_percentage	0.923	0.996	0.983
morphology	0.992	0.948	0.974
sbr_grade	0.815	0.917	0.987
Overall	0.747	0.878	0.888
consistency	0.885	0.929	0.792

4 Information Extraction

To evaluate whether the synthetic training data improves IE in real world scenarios, we use it to fine-tune Mistral-7B-Instruct and compare the resulting model

Table 3. Document-level accuracy of Synthetic Reports (percentage of documents that achieved a given proportion of correctly generated clinical variables, N = 1000)

Accuracy Bucket - % of Correct Vars.	Qwen3 Med	Gemma	Mixtral
Perfect - 100%	0.3%	20.8%	27.2%
High - 85–99%	7.7%	30.8%	30.3%
Moderate - 70–84%	50.8%	41.3%	35.4%
Low - 50–69%	40.0%	7.1%	6.6%
Very Low - < 50%	1.2%	0.0%	0.5%

(Mistral-FT) with the same model in a 0- (Mistral-0S) and a 3-shot (Mistral-3S) setting. The fine-tuning data is composed of report/variable-value pairs, where each variable-value pair corresponds to those correctly generated within the report by any of the three LLMs. This yielded 33,277 training examples, i.e., about 2,500 reports for each of the thirteen variables. For each training instance, the input consists of a prompt including a pathology report, the variable to be extracted, its possible values or range and the required JSON output schema ({"value": ..., "reference": ...}. The output is the expected variable value. We fine-tuned using LoRA adaptation.

We evaluated the IE performance of the three models on a test set of 377 real breast cancer pathology reports manually annotated with gold standard labels for all target variables by experts.

Results Table 4 shows the results.

Adding synthetic examples as in-context demonstrations significantly improves performance over zero-shot prompting, showing their value even without fine-tuning. However, the fine-tuned model performs best overall, surpassing the 3-shot baseline, particularly on variables requiring numerical extraction or specific mention detection (e.g., Ki-67, tumor size, lymph nodes, extracapsular spread).

For some variables, the 3-shot baseline is competitive or slightly better. These correspond to attributes with a small, well defined set of possible values (e.g., positive/negative/unknown) and standardized phrasing in the reports, such as estrogen and progesterone receptor status, clear margins and diagnostic type. In such settings, a small number of in-context examples is often sufficient for the model to generalize.

Morphology is the hardest variable across settings, likely because the prompt uses an ICD-O classification name instead of a fixed label set, forcing the model to generate or extract free text instead of choosing from a limited set of labels.

5 Conclusion and Discussion

We propose a novel privacy-preserving approach for generating synthetic pathology reports from structured data. Applied to real-world breast cancer variables,

Table 4. IE Accuracy and F1 Scores for Fine-Tuned, 0- and 3-Shot Models.

Variable	Acc-0S	Acc-3S	Acc-FT	F1-0S	F1-3S	F1-FT
morphology	0.07	0.15	**0.44**	0.05	0.07	**0.47**
diagnostic_type	0.48	**0.73**	0.71	0.52	**0.76**	0.75
lymphovascular_invasion	0.55	0.63	**0.79**	0.69	0.68	**0.72**
sbr_grade	0.46	0.84	**0.90**	0.55	0.83	**0.90**
her2_status	0.47	0.71	**0.82**	0.55	0.79	**0.82**
Ki67_percentage	0.39	0.69	**0.89**	0.23	0.64	**0.87**
clear_margins	0.28	0.67	**0.77**	0.41	**0.53**	0.50
number_lymph_nodes_affected	0.24	0.66	**0.84**	0.24	0.36	**0.71**
number_lymph_nodes_removed	0.22	0.57	**0.81**	0.14	0.32	**0.69**
estrogen_receptor_status	0.71	**0.83**	0.80	0.79	**0.88**	0.83
progesterone_receptor_status	0.66	**0.88**	0.85	0.71	**0.89**	0.85
extracapsular_tumor_spread	0.15	0.37	**0.98**	0.14	0.13	**0.62**
tumor_size	0.24	0.53	**0.77**	0.15	0.27	**0.63**
Overall	0.38	0.64	**0.79**	0.41	0.58	**0.74**

the resulting synthetic report/data pairs remain faithful to real world distribution while avoiding the exposure of protected health information, thereby creating novel opportunities for the development of robust information extraction systems. We show that a model fine-tuned on this synthetic data achieves strong and consistent improvements in extraction performance across nearly all variables compared to the base model, while also improving overall performance over a 3-shot baseline that uses the generated synthetic pathology reports as in-context examples.

A limitation of this evaluation is that the extraction task targets the same variables used to generate the synthetic reports, so it does not assess generalization beyond the predefined schema. However, the approach can be extended to any new set of clinical variables. As it does not require annotated data, our approach can be applied to other types of clinical texts, provided the corresponding structured data are available. An important direction for future work would be to extend it to additional clinical domains and to other languages.

Acknowledgments. The authors thank SELARL U2R for providing the data and supporting this research. This work received government funding managed by the French National Research Agency under France 2030, reference number “ANR-23- IACL-0004.” (AI Chair Gardent: "Semantically Consistent LLM Based Text Generation"). It was also granted access to the HPC resources of IDRIS under the allocation AD011016561 made by GENCI.

Ethics Statement The study received approval from the institution’s internal bodies. The mandatory declaration was completed with the French regulatory authority in accordance with Reference Methodology MR-004 (ref. MR-004, no. 28088412).

Disclosure of Interests. Alejandra Lorenzo is affiliated with the organization that provided the data used in this study. The authors declare that this affiliation did not influence the design, analysis, or reporting of the results.

References

1. Appenzeller, A., Leitner, M., Philipp, P., Krempel, E., Beyerer, J.: Privacy and utility of private synthetic data for medical data analyses. Applied Sciences **12**(23), 12320 (2022)
2. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960). https://doi.org/10.1177/001316446002000104
3. DataCebo, Inc.: Synthetic Data Metrics (10 2023), `https://docs.sdv.dev/sdmetrics/`, version 0.12.0
4. Dwork, C.: Differential Privacy. In: International colloquium on automata, languages, and programming. pp. 1–12. Springer (2006)
5. Ganev, G., Annamalai, M.S.M.S., Mahiou, S., De Cristofaro, E.: The importance of being discrete: Measuring the impact of discretization in end-to-end differentially private synthetic data. arXiv preprint arXiv:2504.06923 (2025)
6. Jordon, J., Yoon, J., Van Der Schaar, M.: PATE-GAN: Generating synthetic data with differential privacy guarantees. In: International conference on learning representations (2018)
7. Kiran, A., Kumar, S.S.: A methodology and an empirical analysis to determine the most suitable synthetic data generator. IEEE Access **12**, 12209–12228 (2024)
8. Kopp, A.: Microsoft smartnoise differential privacy machine learning case studies. Microsoft Azure White Papers **14** (2021)
9. Krippendorff, K.: Computing krippendorff's alpha-reliability. Tech. rep., Annenberg School for Communication, University of Pennsylvania (2011)
10. McKenna, R., Miklau, G., Sheldon, D.: Winning the NIST contest: A scalable and general approach to differentially private synthetic data. arXiv preprint arXiv:2108.04978 (2021)
11. Patki, N., Wedge, R., Veeramachaneni, K.: The Synthetic Data Vault. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 399–410 (Oct 2016). https://doi.org/10.1109/DSAA.2016.49
12. Rabaey, P., Heytens, S., Demeester, T.: Simsum–simulated benchmark with structured and unstructured medical records. Journal of Biomedical Semantics **16**(1), 20 (2025)
13. Richter-Pechanski, P., Wiesenbach, P., Schwab, D.M., Kiriakou, C., Geis, N., Dieterich, C., Frank, A.: Clinical information extraction for lower-resource languages and domains with few-shot learning using pretrained language models and prompting. Natural Language Processing **31**(5), 1210–1233 (2025). https://doi.org/10.1017/nlp.2024.52
14. Sidorenko, A., Platzer, M., Scriminaci, M., Tiwald, P.: Benchmarking synthetic tabular data: A multi-dimensional evaluation framework. arXiv preprint arXiv:2504.01908 (2025)
15. Wu, H.Y., Zhang, J., Ive, J., Li, T., Gupta, V., Chen, B., Guo, Y.: Medical scientific table-to-text generation with synthetic data under data sparsity constraint. In: NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research (2023)
16. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. Advances in neural information processing systems **32** (2019)