

Construction and Evaluation of a Public LiverTox Benchmark for Herbal- and Drug-Induced Liver Injury with Duplication-Controlled Synthetic Data Generation

Tae-Yoon KIM¹ and Jung-Hyun KIM^{2,*}

¹ Department of Korean Traditional Medicine Practice, Jaseng Medical Foundation, Seoul, Republic of Korea.

² Department of Acupuncture and Moxibustion, Kyung Hee University Hospital at Gangdong, 892 Dongnam-ro, Gangdong-gu, Seoul, 05278, Republic of Korea.

E-mails: pacifier9@naver.com; dan_mi725@naver.com

* Author to whom correspondence should be addressed;

Received: 11 April 2026 / Accepted: 24 September 2026 / Published online: 30 September 2026

Abstract

Purpose: We constructed a parser-audited public benchmark of herbal-induced liver injury (HILI) and non-herbal drug-induced liver injury (DILI) from LiverTox portable document format (PDF) chapters and assessed duplication-controlled synthesis. **Methods:** A PDF-based pipeline converted LiverTox narratives and tables into structured patient-episode records, followed by manual audit and correction. Under an agent-level split, three generators were compared: bootstrap with jitter, an unconstrained Bayesian Gaussian mixture model (BGMM), and a quantile-bounded distance-to-closest-record (DCR) rejection BGMM (QBound+DCR BGMM). Evaluation used parser audit metrics, feature-wise Wasserstein distance, correlation preservation, near-duplicate rate, and train-on-synthetic, test-on-real (TSTR) classification. **Results:** The final benchmark included 64 cases from 31 agents (20 HILI; 44 DILI). The QBound+DCR BGMM model reduced the pooled near-duplicate rate to $35.6\% \pm 6.3\%$, compared with $60.4\% \pm 4.9\%$ for BGMM and $100.0\% \pm 0.0\%$ for bootstrap with jitter, and eliminated class-conditional near-duplicate generation. However, fidelity worsened. The TSTR AUROC was 0.233 ± 0.119 , compared with 0.183 ± 0.083 for BGMM, 0.539 ± 0.280 for bootstrap, and 0.222 for the train-on-real reference. **Conclusion:** The benchmark supports reproducible evaluation and shows that duplication control improves novelty but introduces a fidelity-utility trade-off.

Keywords: Liver Diseases, Drug-Induced; Phytotherapy; Machine Learning; Data Mining.

Introduction

Drug-induced liver injury (DILI) remains one of the most challenging problems in clinical hepatology because it is idiosyncratic, phenotypically heterogeneous, and fundamentally a diagnosis of exclusion. Similar difficulties extend to herbal-induced liver injury (HILI), where exposure histories are often complex, product composition may be uncertain, and formal causality assessment is essential for separating plausible injury from coincidental association. Contemporary hepatology guidance therefore emphasizes the structured evaluation of exposure chronology, competing etiologies, biochemical phenotype, severity, and causality rather than reliance on isolated laboratory abnormalities alone [1-3]. The HILI evidence base is particularly fragmented compared with conventional DILI, and published reports often differ in diagnostic rigor, causality adjudication, and completeness of laboratory detail [4, 5].

At the same time, toxicovigilance research increasingly depends on the computational reuse of heterogeneous clinical evidence. LiverTox is an openly accessible resource of the National Institute of Diabetes and Digestive and Kidney Diseases hosted by the NCBI Bookshelf; it provides curated, agent-specific chapters on drug-, herbal-, and

dietary supplement-induced liver injury, including clinical narratives, likelihood assessments, laboratory patterns, and case material relevant to DILI and HILI [6]. This structure makes LiverTox clinically rich but computationally difficult to reuse: it is a chapter-based narrative resource, not a ready-to-analyze patient-level benchmark. Case material is distributed across prose, heterogeneous tables, and variable reporting styles, which complicates reproducible extraction of structured records suitable for machine learning, benchmarking, or synthetic data research.

Synthetic tabular data generation has been proposed as a method to support analysis when real datasets are small, access-restricted, or privacy-sensitive. Representative methods, such as conditional generative adversarial networks for tabular data, have shown that realistic synthesis of heterogeneous structured data is feasible, while healthcare-specific studies have further argued that synthetic patient data may support method development, benchmarking, and limited forms of data sharing [7-9]. However, the synthetic data literature also makes it clear that apparent realism is insufficient. Synthetic datasets must be evaluated jointly for distributional fidelity, downstream analytical utility, and disclosure or memorization risk, and the field still lacks consensus regarding which combination of metrics is most appropriate for specific biomedical settings [8, 10].

These concerns are magnified in liver injury research. Real-world HILI and DILI cohorts are typically small, class balance is often poor, and clinically meaningful structure resides not merely in univariate laboratory ranges but in multivariable relations among aminotransferases, alkaline phosphatase, bilirubin, latency, and injury pattern. Under these conditions, a synthetic generator can appear useful while in fact reproducing near-duplicate records, attenuating clinically relevant covariance, or performing well only because evaluation is insufficiently strict. In other words, the central problem is not simply how to generate more rows, but how to do so under a benchmark and evaluation design that makes fidelity, novelty, and utility simultaneously inspectable rather than selectively reported [2, 10].

The aim of this study was to construct a parser-audited public benchmark of HILI and DILI from LiverTox chapter materials and to use this benchmark to evaluate duplication-controlled synthetic tabular data generation under an agent-level split. Specifically, we assessed whether explicit duplication control could reduce near-duplicate synthesis while preserving enough fidelity and downstream utility to support reproducible toxicovigilance research.

Materials and Methods

Study Design and Source Corpus

This study was designed as a benchmark construction and evaluation study. Publicly available LiverTox portable document format (PDF) chapter files were used as the sole source of information to derive patient-level episodes of herbal-induced liver injury (HILI) and non-herbal drug-induced liver injury (DILI). LiverTox was selected because it is a dedicated, openly accessible resource for drug- and herb-induced liver injury and contains agent-specific chapters with clinical narratives and case materials relevant to structured toxicovigilance analysis [6].

Chapter eligibility was assessed by both authors using predefined criteria before case extraction. Chapters were eligible when they described herbal or dietary supplement exposure relevant to HILI, or conventional drug exposure relevant to DILI, and contained extractable patient-level case material with episode-level laboratory values. Chapters were excluded when they contained only general narrative summaries, lacked extractable liver chemistry values, duplicated cases already represented elsewhere, or did not permit mapping to a distinct patient episode. Herbal and dietary supplement chapters were assigned to the HILI source pool, whereas conventional drug chapters were assigned to the DILI source pool. The unit of analysis was the patient episode. The included agents are listed in Supplementary File S1, and excluded candidate records with exclusion reasons are listed in Supplementary File S2.

Case Extraction, Audit, and Benchmark Curation

A rule-based extraction pipeline was developed to convert chapter narratives and case tables into structured patient-episode records. Text was extracted from PDF chapter files using Python and pypdf; tabular and narrative outputs were then parsed with deterministic regular expressions and line-based table rules. The parser targeted age, sex, latency, ALT, AST, ALP, total bilirubin, INR, case number, agent name, and source chapter metadata. Age and sex were extracted from expressions such as “year-old woman” or “year-old man.” Latency values were converted to approximate days using fixed multipliers for days, weeks, months, and years. Laboratory values were accepted only when explicit numeric values and interpretable units were present; ALT, AST, ALP, total bilirubin, and INR were mapped to peak episode-level values when multiple values were available.

Ranges, approximate expressions, relative values such as “times the upper limit of normal,” and unresolved field mappings were routed to manual review rather than coerced into single analyzable values. A parser-audit layer was then applied because semi-structured toxicology documents cannot be assumed to yield reliable machine-extracted records without verification. The audit subset contained 17 patient-episode records selected to include both injury classes and both narrative- and table-derived extraction routes. A gold-standard extraction sheet was constructed by manual review and consensus. Within-tolerance agreement was calculated for numeric fields; formal inter-rater statistics were not computed because the reference sheet was consensus-curated rather than independently double-coded. Corrected values were propagated into the final benchmark release. Therefore, the released benchmark is parser-audited and corrected, not fully automated.

Cohort Definition, Variables, and Data Splitting

Episodes were included when they could be mapped to a single agent exposure and contained extractable core liver chemistry sufficient for benchmark use, with ALT and total bilirubin treated as required minimum fields. Records were excluded when case-level values remained irreconcilably ambiguous after review, represented duplicate extractions, lacked required core variables, or failed to meet the benchmark injury-screening criteria used for this study. Episodes were labeled as HILI when derived from herbal or dietary supplement chapters and as DILI when derived from conventional drug chapters.

The main continuous variables were ALT, AST, ALP, total bilirubin, age, and latency, when available. Derived variables included the De Ritis ratio (AST/ALT), R value, injury pattern category, and Hy’s law flag. R-value interpretation followed established clinical liver injury conventions [1, 2]. Data were split at the agent level, so that all records from the same causative agent were assigned to the same partition. This was done to reduce leakage and to make downstream evaluation reflect generalization across unseen agents, rather than repeated exposure to the same agent-specific profile.

Synthetic Data Generation and Evaluation Framework

Three synthetic data generators were compared: (1) bootstrap with jitter as an empirical resampling baseline, (2) an unconstrained Bayesian Gaussian mixture model (BGMM), and (3) a duplication-controlled quantile-bounded distance-to-closest-record rejection BGMM (QBound+DCR BGMM). The BGMM was fit class-conditionally using scikit-learn BayesianGaussianMixture with full covariance matrices, a maximum of five components per class or fewer when constrained by class size, a Dirichlet-process weight concentration prior of 0.01, regularization covariance of 1×10^{-6} , and a maximum of 1000 iterations. The QBound+DCR BGMM used the same class-conditional BGMM backbone, then rejected samples falling outside the 1st- to 99th-percentile class-specific empirical envelope expanded by 15% of the class-specific spread, samples closer than the class-specific 1st-percentile real-record distance threshold, and samples too close to already accepted synthetic records. The pooled epsilon threshold reported in the tables was 1.373 in standardized feature space.

Evaluation was organized along three dimensions recognized in the medical synthetic data literature: distributional fidelity, record-level novelty, and downstream utility [8, 10]. Fidelity was quantified using feature-wise 1-Wasserstein distance and correlation preservation, expressed as the mean absolute error between real and synthetic correlation matrices. Novelty was quantified using distance to the closest real training record and a near-

duplicate rule based on the epsilon threshold; these novelty metrics were interpreted as memorization-risk proxies rather than formal privacy guarantees [10]. Downstream utility was evaluated using train-on-synthetic, test-on-real (TSTR) classification and compared with a train-on-real, test-on-real (TRTR) reference [8, 11].

Statistical Analysis and Reproducibility

All generation experiments were repeated across 10 random seeds and summarized as mean $\pm$ standard deviation across runs. Numeric laboratory-like variables were log1p-transformed before modeling when appropriate, missing numeric values were imputed using training-set medians, and features were standardized using training-set parameters only. The TSTR and TRTR classifier was balanced logistic regression with maximum 2000 iterations and no hyperparameter search; no test-set information was used for preprocessing, model fitting, or synthetic data generation. Given the benchmark size, the analysis was intended to show stability and directionality rather than strong asymptotic inference.

All extraction, curation, splitting, synthesis, and evaluation procedures were implemented in Python using pandas, NumPy, scikit-learn, SciPy, and pypdf. Reproducibility was supported through archived source PDFs, parser-audit logs, curated cohort files, split definitions, model configuration files, and regenerated summary outputs. Pseudocode, configuration files, the included-agent list, the exclusion log, and corrected supplementary CSV files are provided as supplementary materials. The resulting study package should therefore be understood as a reproducible public benchmark and evaluation workflow for synthetic liver injury data, rather than as a claim of a universally superior generator.

Results

Benchmark Construction and Cohort Characteristics

The extraction and curation workflow yielded 74 candidate patient-episode records, of which 64 were retained in the final parser-audited benchmark and 10 were excluded during post-extraction curation. The final benchmark comprised 31 unique agents, including 20 HILI cases and 44 DILI cases. Under the agent-level split, the training set contained 45 cases, the validation set contained 10 cases, and the test set contained 9 cases; no agent appeared in more than one partition. The cohort construction workflow is shown in Figure 1.

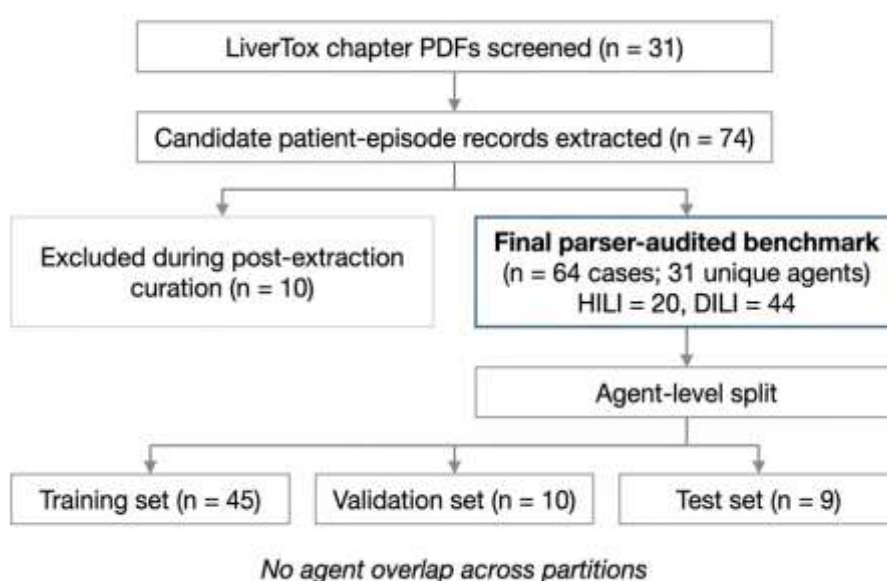


Figure 1. Cohort construction workflow for the parser-audited public LiverTox benchmark

The demographic and biochemical characteristics of the final benchmark are summarized in Table 1. Core liver chemistry availability was complete for ALT and total bilirubin and high for ALP, whereas AST and INR were less consistently extractable. The class-level biochemical patterns were clinically plausible: HILI cases showed higher ALT values and a more hepatocellular profile, whereas DILI cases were more heterogeneous across hepatocellular, mixed, and cholestatic patterns.

Table 1. Demographic and biochemical characteristics of the parser-audited LiverTox benchmark.

Group	Overall	HILI	DILI
Cases, n	64	20	44
Unique agents, n	31	13	18
Women, n (%)	32 (50.0%)	12 (60.0%)	20 (45.5%)
Men, n (%)	19 (29.7%)	4 (20.0%)	15 (34.1%)
Hy's law, n (%)	10 (15.6%)	4 (20.0%)	6 (13.6%)
Hepatocellular, n (%)	27 (42.2%)	13 (65.0%)	14 (31.8%)
Mixed, n (%)	15 (23.4%)	3 (15.0%)	12 (27.3%)
Cholestatic, n (%)	19 (29.7%)	3 (15.0%)	16 (36.4%)
Age, years	46.5 [28.2-57.8]	41.0 [33.0-55.0]	47.0 [26.0-60.0]
Age, years available, n (%)	54 (84.4%)	17 (85.0%)	37 (84.1%)
Latency, days	42.0 [24.0-90.0]	56.0 [28.0-90.0]	41.0 [21.0-90.0]
Latency, days available, n (%)	59 (92.2%)	19 (95.0%)	40 (90.9%)
ALT, U/L	927.5 [264.2-1569.5]	1553.5 [1322.0-2306.2]	635.5 [173.0-1112.5]
ALT, U/L available, n (%)	64 (100.0%)	20 (100.0%)	44 (100.0%)
AST, U/L	1219.0 [637.0-2001.0]	1425.5 [867.5-1803.0]	775.0 [367.5-2393.0]
AST, U/L available, n (%)	37 (57.8%)	18 (90.0%)	19 (43.2%)
ALP, U/L	456.0 [285.0-741.0]	500.0 [257.0-831.5]	454.0 [310.2-738.5]
ALP, U/L available, n (%)	61 (95.3%)	19 (95.0%)	42 (95.5%)
Total bilirubin, mg/dL	14.6 [8.4-24.1]	17.6 [11.0-26.7]	14.2 [7.3-23.7]
Total bilirubin, mg/dL available, n (%)	64 (100.0%)	20 (100.0%)	44 (100.0%)
INR	7.6 [3.9-15.4]	6.6 [6.0-11.7]	8.6 [3.1-15.4]
INR available, n (%)	16 (25.0%)	5 (25.0%)	11 (25.0%)

Data are presented as *n* (%) for categorical variables and as median (interquartile range) for continuous variables. The availability percentages refer to the proportion of cases with extractable values in the final parser-audited benchmark. HILI, herbal-induced liver injury; DILI, non-herbal drug-induced liver injury; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALP, alkaline phosphatase; INR, international normalized ratio; IQR, interquartile range.

These class-wise patterns suggest that the benchmark preserved clinically meaningful biochemical heterogeneity while remaining constrained by small sample size and incomplete availability of some laboratory fields.

Parser Audit and Benchmark Reliability

Parser audit showed that raw automated extraction was adequate for several core fields but insufficiently reliable for benchmark release without manual correction. Agreement was strongest for AST, sex, core-laboratory row completeness, ALT, and total bilirubin, but substantially weaker for ALP and INR, particularly when values were table-derived (Figure 2).

After audit-driven correction, all audited fields achieved complete coverage and exact or tolerance-based agreement in the reviewed subset. Accordingly, the final released benchmark should be interpreted as a parser-audited and corrected dataset rather than as a purely automated extraction product.

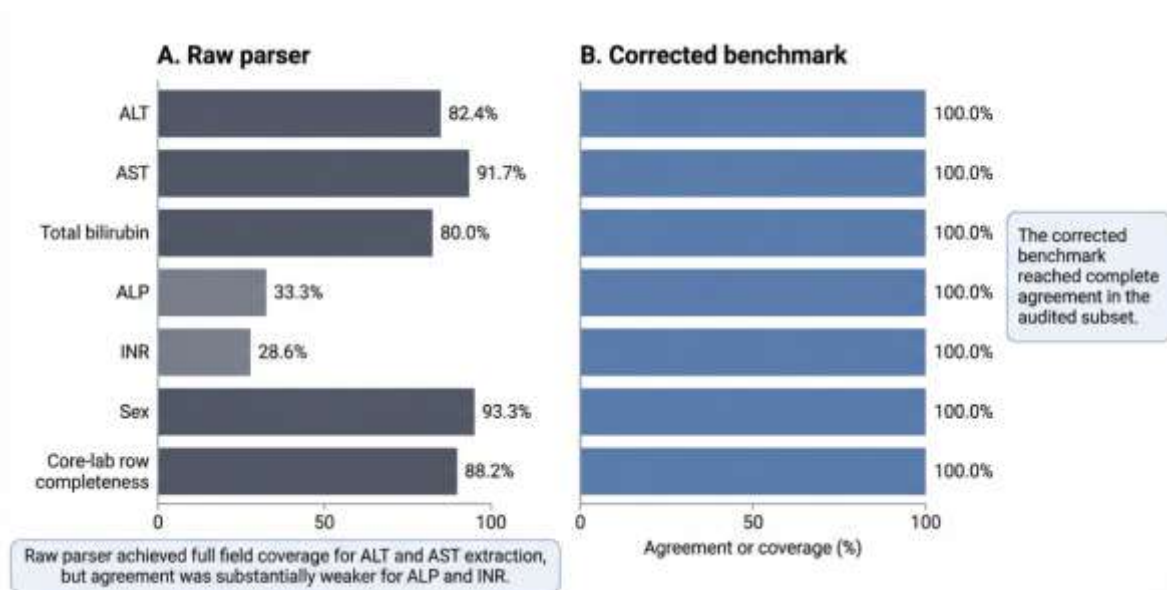


Figure 2. Parser audit profile: raw parser versus corrected benchmark

Distributional Fidelity of Synthetic Cohorts

Feature-wise Wasserstein distances and correlation preservation across the three synthetic generators are presented in Table 2.

Table 2. Feature-wise Wasserstein distance and correlation preservation across synthetic data generators.

Model	Runs	Corr. MAE (pooled)	WD ALT	WD AST	WD ALP	WD Bilirubin	WD Age	WD Latency	WD INR
BGMM baseline	10	0.171 ± 0.017	288.5 ± 72.9	382.8 ± 52.6	99.6 ± 29.5	3.7 ± 0.7	5.9 ± 1.0	98.3 ± 14.6	3.2 ± 0.3
Bootstrap+jitter	10	0.139 ± 0.020	247.2 ± 61.3	336.8 ± 78.5	71.3 ± 31.0	2.0 ± 0.6	4.1 ± 1.0	79.2 ± 49.1	3.2 ± 0.3
QBound+DCR BGMM	10	0.219 ± 0.021	350.8 ± 109.2	498.4 ± 52.1	149.4 ± 25.0	5.2 ± 1.7	7.6 ± 1.0	109.0 ± 39.6	3.4 ± 0.2

Values are reported as mean $\pm$ standard deviation across 10 repeated random-seed runs. Lower feature-wise Wasserstein distances indicate better distributional fidelity. Correlation preservation is expressed as the mean absolute error between the real and synthetic correlation matrices, where lower values indicate better preservation of the dependence structure. BGMM, Bayesian Gaussian mixture model; QBound+DCR BGMM, quantile-bounded distance-to-closest-record rejection Bayesian Gaussian mixture model.

Bootstrap with jitter showed the strongest correlation preservation, followed by BGMM, whereas QBound+DCR BGMM showed the weakest correlation preservation and generally larger feature-wise Wasserstein distances. These findings indicate that explicit duplication control imposed a measurable fidelity cost rather than improving conventional distribution matching.

Taken together, the fidelity results show that QBound+DCR BGMM should not be interpreted as a fidelity-optimized generator. Its role in this benchmark is better understood as a duplication-control model that deliberately moves synthetic records away from the closest real training records.

Record-Level Novelty and Near-Duplicate Generation

Record-level novelty differed markedly across generators (Table 3). Bootstrap with jitter remained closest to the real training records and produced near-duplicates by the study definition in every run. The BGMM model improved distance from the closest real record but still retained substantial near-duplicate generation. The QBound+DCR BGMM model produced the highest pooled DCR and the lowest pooled near-duplicate rate.

Table 2. Record-level novelty and near-duplicate generation across synthetic data generators.

Model	Pooled DCR median	epsilon threshold (pooled)	Near-duplicate rate (pooled)	Near-duplicate rate (HILI)	Near-duplicate rate (DILI)
BGMM baseline	1.265 ± 0.036	1.373 ± 0.000	60.4% ± 4.9%	85.7% ± 12.6%	34.2% ± 7.8%
Bootstrap+jitter	0.125 ± 0.008	1.373 ± 0.000	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
QBound+DCR BGMM	1.438 ± 0.036	1.373 ± 0.000	35.6% ± 6.3%	0.0% ± 0.0%	0.0% ± 0.0%

Values are reported as mean ± standard deviation across 10 repeated random-seed runs, unless otherwise indicated. Distance to the closest real record (DCR) was computed in the standardized feature space. The near-duplicate rate was defined as the proportion of synthetic records with DCR below the pooled epsilon threshold. Lower near-duplicate rates indicate greater record-level novelty. BGMM, Bayesian Gaussian mixture model; QBound+DCR BGMM, quantile-bounded distance-to-closest-record rejection Bayesian Gaussian mixture model.

Class-conditional novelty analysis showed that QBound+DCR BGMM eliminated near-duplicate generation under the class-specific thresholds used in this study, whereas BGMM retained near-duplicates in both classes. This result supports the intended function of the rejection mechanism while not implying formal privacy protection.

Downstream Utility under TSTR

Downstream classification performance under the train-on-synthetic, test-on-real framework is shown in Table 4.

Table 3. Downstream classification performance under the train-on-synthetic, test-on-real framework.

Model	BGMM baseline	Bootstrap+jitter	QBound+DCR BGMM
TSTR AUROC	0.183 ± 0.083	0.539 ± 0.280	0.233 ± 0.119
TSTR AUPRC	0.285 ± 0.027	0.520 ± 0.239	0.306 ± 0.053
TSTR Balanced accuracy	0.317 ± 0.102	0.558 ± 0.219	0.325 ± 0.133
AUROC vs TRTR	-0.039	0.317	0.011
AUPRC vs TRTR	-0.005	0.23	0.016

Values are reported as mean ± standard deviation across 10 repeated random-seed runs, except for the train-on-real, test-on-real reference, which is deterministic for the fixed split. Higher AUROC, AUPRC, and balanced accuracy indicate better downstream classification performance. TSTR, train-on-synthetic, test-on-real; TRTR, train-on-real, test-on-real; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve.

The unconstrained BGMM yielded poor downstream utility, whereas QBound+DCR BGMM performed slightly better and was close to the fixed-split TRTR reference. However, the TRTR AUROC was itself low, reflecting the severe small-sample and agent-level test-set constraints rather than a stable estimate of real-data classifier performance.

Bootstrap with jitter achieved the highest TSTR performance, but this result must be interpreted alongside its complete near-duplicate generation. In this benchmark, the highest apparent utility was therefore obtained by the least novel generator.

Overall, the results support a benchmark-paper interpretation. The central finding is not that one generator was uniformly superior, but that the parser-audited LiverTox benchmark exposes a measurable trade-off among fidelity, novelty, and downstream utility. Figure 3 summarizes this trade-off graphically.

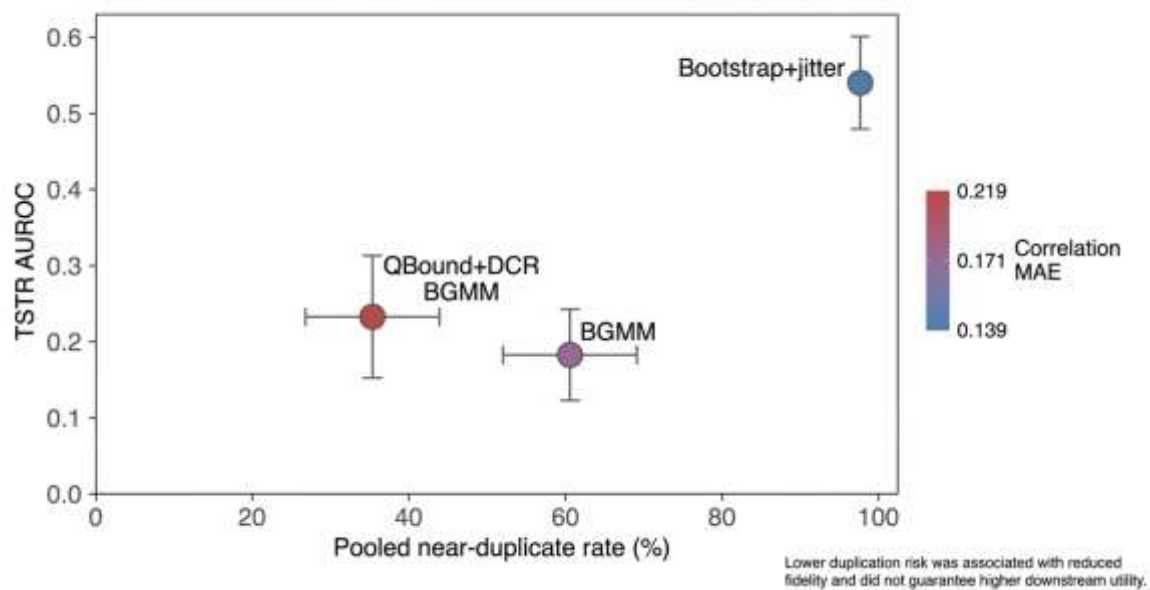


Figure 3. Fidelity-novelty-utility trade-off across synthetic data generators. The x-axis shows the pooled near-duplicate rate, the y-axis shows TSTR AUROC, and point color represents correlation mean absolute error. The figure illustrates that lower duplication risk did not automatically translate into better distributional fidelity or higher downstream utility.

Discussion

The main finding of this study is that a parser-audited public LiverTox benchmark can expose a measurable trade-off among distributional fidelity, record-level novelty, and downstream utility in synthetic liver injury data generation. The QBound+DCR BGMM model reduced near-duplicate generation substantially, but this gain was accompanied by weaker feature-wise fidelity and only modest downstream utility. This pattern is not a failure of the benchmark; it is precisely the type of trade-off that a benchmark for synthetic biomedical data should make visible.

A likely explanation is that distance-based rejection deliberately pushes accepted samples away from the local neighborhoods of real training records. This increases record-level novelty, but it can also distort class-specific covariance and weaken correlation preservation. The finding is consistent with broader clinical information extraction and synthetic-data literature showing that structured reuse of medical text and synthetic data evaluation both require task-specific validation rather than reliance on surface plausibility alone [12-15].

The results also caution against interpreting TSTR performance in isolation. Bootstrap with jitter achieved the strongest apparent utility, yet it generated near-duplicates by construction. Conversely, QBound+DCR BGMM was less duplicative but less faithful. Similar privacy-fidelity-utility tensions have been reported in synthetic clinical data studies, where no single generator consistently dominates across all evaluation axes [15-18].

For liver injury research, this matters because clinically relevant structure is sparse, heterogeneous, and tied to agent-specific patterns. A random record-level split or a single utility metric would probably overstate performance. The agent-level split used here is deliberately stringent, but the resulting test set was small; the low TSTR AUROC and high variability of TSTR estimates therefore limit any strong claim about downstream generalizability.

The more durable contribution is therefore the benchmark and evaluation workflow rather than the present lightweight generator. Recent reviews have emphasized that synthetic healthcare data studies remain heterogeneous in their validation practices and often lack comparable benchmarks [18]. This study contributes a narrowly focused, parser-audited benchmark in which future generators can be compared under the same source corpus, split design, and multi-axis evaluation framework.

Several limitations remain. The benchmark contains only 64 patient episodes from 31 agents, with an agent-level test set of nine cases, so performance estimates are inherently unstable. The parser is not fully automated and

depends on audit-driven correction, especially for ALP and INR. The present study did not add an epsilon-threshold sensitivity analysis because the small per-class pairwise distance structure makes such threshold sweeps unstable; this limitation is now explicit. Finally, the benchmark is derived from published LiverTox chapters rather than prospective case adjudication, so it should be used for computational evaluation rather than clinical incidence estimation.

Future work should expand the benchmark with additional LiverTox chapters, release versioned benchmark files and code in a public repository, and evaluate stronger generators using the same agent-level split. Future extraction workflows should also combine deterministic parsing with auditable human-in-the-loop verification. The present results support a cautious but useful conclusion: public liver injury case material can be transformed into a reproducible benchmark, and that benchmark can quantify the cost of duplication control rather than merely report synthetic data performance in the most favorable metric.

Conclusions

This study shows that a parser-audited public LiverTox benchmark can support reproducible, leakage-aware evaluation of synthetic liver injury data generation. Consistent with prior evidence that synthetic health data must be judged jointly for fidelity, novelty or privacy risk, and utility [10, 15, 18], the present results indicate that duplication control improves record-level novelty but does not remove the fidelity-utility trade-off. The benchmark and evaluation workflow therefore provide a practical basis for more transparent synthetic data assessment in toxicovigilance research.

List of Abbreviations: ALP - alkaline phosphatase; ALT - alanine aminotransferase; AST - aspartate aminotransferase; AUROC - area under the receiver operating characteristic curve; AUPRC - area under the precision-recall curve; BGMM - Bayesian Gaussian mixture model; DCR - distance to the closest real record; DILI - drug-induced liver injury; HILI - herbal-induced liver injury; INR - international normalized ratio; IQR - interquartile range; MAE - mean absolute error; PDF - portable document format; QBound+DCR BGMM - quantile-bounded distance-to-closest-record rejection Bayesian Gaussian mixture model; TSTR - train-on-synthetic, test-on-real; TRTR - train-on-real, test-on-real; WD - Wasserstein distance.

Author Contributions: T.Y.K.: Conceptualization, Methodology, Software, Formal analysis, Data curation, Validation, Visualization, Writing – original draft. J.H.K.: Conceptualization, Methodology, Supervision, Validation, Project administration, Writing – review and editing, Correspondence.

Funding: This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2026-25507270).

Ethics Statement: This study used only publicly available, previously published, de-identified case materials from the LiverTox database. There was no direct participant contact, intervention, or access to non-public medical records. In accordance with our institutional policy, formal ethical approval was not required for this study.

Data Availability Statement: The source materials used in this study are publicly available through LiverTox, a resource of the National Institute of Diabetes and Digestive and Kidney Diseases hosted by the NCBI Bookshelf. The parser-audited benchmark, included-agent list, exclusion log, corrected supplementary CSV files, model configuration files, and analysis scripts are provided as supplementary materials with this revised submission.

Acknowledgments: We acknowledge the assistance of Grammarly (<https://www.grammarly.com>) and Trinka (<https://www.trinka.ai>) for improving the linguistic quality of this manuscript. The generative AI engines embedded in these two platforms were used solely for language editing and correction of linguistic errors. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Conflict of Interest: The authors declare no conflict of interest.

References

1. EASL Clinical Practice Guidelines: Drug-induced liver injury. *J Hepatol.* 2019;70(6):1222-1261. <https://doi.org/10.1016/j.jhep.2019.02.014>.

2. Fontana RJ, Liou I, Reuben A, Suzuki A, Fiel MI, Lee W, et al. AASLD practice guidance on drug, herbal, and dietary supplement-induced liver injury. *Hepatology*. 2023;77(3):1036-1065. <https://doi.org/10.1002/hep.32689>.
3. Danan G, Teschke R. RUCAM in Drug and Herb Induced Liver Injury: The Update. *Int J Mol Sci*. 2015;17(1):14. <https://doi.org/10.3390/ijms17010014>.
4. Ballotin VR, Bigarella LG, Brandão ABM, Balbinot RA, Balbinot SS, Soldara J. Herb-induced liver injury: Systematic review and meta-analysis. *World J Clin Cases*. 2021;9(20):5490-5513. <https://doi.org/10.12998/wjcc.v9.i20.5490>.
5. Byeon J-H, Kil J-H, Ahn Y-C, Son C-G. Systematic review of published data on herb induced liver injury. *J Ethnopharmacol*. 2019;233:190-196. <https://doi.org/10.1016/j.jep.2019.01.006>.
6. Serrano J. LiverTox: An online information resource and a site for case report submission on drug-induced liver injury. *Clin Liver Dis (Hoboken)*. 2014;4(1):22-25. <https://doi.org/10.1002/cld.388>.
7. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019, Article No.: 659, pp.7335-7345.
8. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol*. 2020;20(1):108. <https://doi.org/10.1186/s12874-020-00977-1>.
9. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digit Med*. 2020;3(1):147. <https://doi.org/10.1038/s41746-020-00353-9>.
10. Kaabachi B, Despraz J, Meurers T, Otte K, Halilovic M, Kulynych B, et al. A scoping review of privacy and utility metrics in medical synthetic data. *NPJ Digit Med*. 2025;8(1):60. <https://doi.org/10.1038/s41746-024-01359-3>.
11. Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv preprint arXiv:1706.02633*. 2017.
12. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform*. 2017;73:14-29. <https://doi.org/10.1016/j.jbi.2017.07.012>.
13. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;77:34-49. <https://doi.org/10.1016/j.jbi.2017.11.011>.
14. El Emam K, Mosquera L, Fang X, El-Hussuna A. Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR Med Inform*. 2022;10(4):e35734. <https://doi.org/10.2196/35734>.
15. Adams T, Birkenbihl C, Otte K, Ng HG, Rieling JA, Näher A-F, et al. On the fidelity versus privacy and utility trade-off of synthetic patient data. *iScience*. 2025;28(5):112382. <https://doi.org/10.1016/j.isci.2025.112382>.
16. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes*. 2019;12(7):e005122. <https://doi.org/10.1161/CIRCOUTCOMES.118.005122>.
17. El Kababji S, Mitsakakis N, Fang X, Beltran-Bless A-A, Pond G, Vandermeer L, et al. Evaluating the utility and privacy of synthetic breast cancer clinical trial data sets. *JCO Clin Cancer Inform*. 2023;7:e2300116. <https://doi.org/10.1200/CCI.23.00116>.
18. Pezoulas VC, Zaridis DI, Mylona E, Androutsos C, Apostolidis K, Tachos NS, et al. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Comput Struct Biotechnol J*. 2024;23:2892-2910. <https://doi.org/10.1016/j.csbj.2024.07.005>.
19. Liu L, Blake V, Barman M, Gallego B, Churches T, Kennedy G, et al. Using natural language processing to extract information from clinical text in electronic medical records for populating clinical registries: a systematic review. *J Am Med Inform Assoc*. 2026;33(2):484-498. <https://doi.org/10.1093/jamia/ocaf176>.
20. Wiest IC, Wolf F, Leßmann M-E, van Treeck M, Ferber D, Zhu J, et al. A software pipeline for medical information extraction with large language models, open source and suitable for oncology. *NPJ Precis Oncol*. 2025;9(1):313. <https://doi.org/10.1038/s41698-025-01103-4>.