Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# MAUDE-Atlas: a deduplicated, leakage-controlled and terminology-coded derivative of the FDA MAUDE medical device adverse event reports

**Maxym Chekerskyi**

3dmonster3@gmail.com

Independent researcher https://orcid.org/0009-0004-9749-4867

**Research Article**

**Keywords:** medical device safety, post-market surveillance, adverse events, MAUDE, openFDA, deduplication, signal detection, benchmark dataset

**Posted Date:** November 23rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11789568/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare potential competing interests as follows: The author sells the full MAUDE-Atlas release and other datasets through https://sentineldata.com.ua. The free 100,003-report sample is deposited on Zenodo under CC BY 4.0.

# Abstract

MAUDE-Atlas v1.0.0 is a derivative of the public FDA Manufacturer and User Facility Device Experience (MAUDE) database, built from the openFDA device event endpoint on 30 August 2026. It contains 25,711,469 medical device adverse event reports, joined to patient records, report narratives, device identifiers and enforcement recall data, and coded against the official FDA MDR adverse-event terminology, which embeds the IMDRF annex A-G terms. Near-duplicate reports (MinHash/LSH, Jaccard >= 0.85) are grouped into 1,186,060 multi-report clusters and flagged, not removed. Reports are assigned to temporal training, validation and test partitions and to an out-of-distribution test set built from a 20% product-code holdout, with cluster-level leakage control. We report quality-control gates, a precision check on 200 sampled duplicate clusters, and three baseline results: severity classification (macro-F1 0.636 on the temporal test split), disproportionality signal detection (ROR AUC 0.561, PRR AUC 0.552) and link prediction (Hits@10 0.123). A free stratified sample of 100,003 reports is deposited on Zenodo under CC BY 4.0. The data support work on post-market surveillance, signal detection, text-based severity modelling and graph benchmarks. They are not validated for clinical decisions.


# Background

MAUDE collects adverse event and malfunction reports on medical devices from manufacturers, importers and device user facilities, whose reporting is mandatory, and from voluntary reporters such as clinicians and patients [1, 2]. It is a passive surveillance system, so it is open to under-reporting and reporting bias. The openFDA service exposes the reports through an API and bulk files [3]. Two properties of the raw files make secondary analysis harder. The same event is often reported more than once, for example by a manufacturer and by a user facility. And many problem descriptions are free text that has not been mapped to a controlled vocabulary. MAUDE-Atlas addresses both: it flags near-duplicate reports and codes problem terms by exact match against the official FDA MDR terminology. The release follows the Datasheets for Datasets structure [4].

# Data description

The full release v1.0.0 holds 25,711,469 event reports in 1,472 Apache Parquet files (Snappy compression), 18.73 GB in total. The public Zenodo deposit [9] holds a stratified sample of 100,003 reports. The sample is a true subset of the full release and preserves the distribution of event types across decades. It includes all linked rows from the device, patient, narrative, problem-code, cluster and split tables for the sampled reports, and the full terminology table. Table 1 lists the sample files.

Table 1
Files in the public sample.

| File | Rows | Content |
|---|---|---|
| reports.parquet | 100,003 | Canonical reports, event types, dates, deduplication flags |
| patients.parquet | 100,009 | Patient demographics and linkage to reports |
| mdr_texts.parquet | 247,259 | Narrative texts (event descriptions and other text fields) |
| patient_problem_codes.parquet | 210,946 | Patient problem codes, exact match to MDR terminology |
| event_clusters.parquet | 100,003 | Duplicate cluster assignment and canonical-report flag |
| splits.parquet | 100,003 | Assignment to train, validation, test or test_ood |

The sample also holds the terminology table (terminology_codes.parquet, 1,885 rows) and a manifest. It keeps only the canonical report of each duplicate cluster, so every sample row has is_canonical = true; 6,381 sample reports are still marked as near-duplicate members (dup_near) and 406 as exact duplicates (dup_exact). The device table (devices_enriched) and the product problem code table exist in the sample but have no rows. Of the 210,946 patient problem codes in the sample, 141,781 are exact matches and 69,165 are unclassified.

The full release has these tables (rows): reports 25,711,469; devices 25,753,190 (one row per device per report); patients 25,710,765; mdr_texts 58,791,403; device problem codes 52,552,238; patient problem codes 53,688,916; terminology codes 1,885; event clusters 25,711,469; splits 25,711,469; and reference tables (enforcement, 510(k), PMA, classification, recall, registration listing) with 672,896 rows in total. The clustering leaves 17,427,253 canonical unique events, so 32.2% of the reports are duplicates.

Table 2 gives the composition of the full release. Malfunctions are 62.1% of reports, injuries 36.4% and deaths 0.9%. The terminology table has 1,885 codes (annex A 487, B 31, C 160, D 43, E 784, F 81, G 299). Problem coding is reported per coded term occurrence: 85.52% of product problem terms and 67.3% of patient problem terms matched an official term exactly; the rest are kept and flagged as unclassified.

Table 2
Composition of the full release.

| Measure | Value |
|---|---|
| Event reports | 25,711,469 |
| Event type: death | 229,717 |
| Event type: injury | 9,346,364 |
| Event type: malfunction | 15,975,902 |
| Event type: other | 102,004 |
| Event type: unknown | 57,482 |
| Device mapping: UDI DI exact | 10,787,984 |
| Device mapping: brand exact | 2,516,924 |
| Device mapping: product code fallback | 12,427,640 |
| Device mapping: unmatched | 20,642 |
| Product problem terms: exact match / unclassified | 44,942,346 / 7,609,892 |
| Patient problem terms: exact match / unclassified | 36,131,617 / 17,557,299 |
| Multi-report duplicate clusters | 1,186,060 |
| Reports in such clusters | 9,470,276 |
| Split: train / validation / test / test_ood | 13,115,359 / 1,989,440 / 7,732,418 / 2,874,252 |

Device mapping is counted per device record. The devices table has 25,753,190 rows, 41,721 more than the number of reports, because a report that names several devices produces several rows. By share, 41.9% of devices were matched by UDI DI, 48.3% by product code, 9.8% by exact brand and 0.08% not at all.

Documentation is deposited with the data: a datasheet in English and a Russian-language sample datasheet.

## Experimental design, materials and methods

**Collection.** All source files were downloaded on 30 August 2026 from the openFDA bulk manifest (api.fda.gov/download.json, manifest saved) and from fda.gov (the MDR codes workbook). The device event data come from the openFDA device/event endpoint and were enriched with GUDID (UDI), enforcement recall, 510(k), PMA, classification and registration-listing data [3]. SHA-256 hashes of every source file are recorded in a manifest. The IMDRF website could not be reached from the build

environment, so the FDA MDR codes workbook, which embeds the IMDRF annex A-G terms, was used as the primary terminology. This is a documented deviation.

**Parsing and identifiers.** Source files were parsed as streams (ijson). Each report received a stable identifier, the first 16 hex characters of the SHA-256 of the string 'openfda_device_event:' followed by the MDR report key.

**Cleaning.** Dates were converted to ISO-8601 and given validity flags: future (after 2026-08-30), before 1990, invalid calendar date, and unparseable. Event types were standardised to death, injury, malfunction, other and unknown.

**Problem coding.** Product and patient problem terms were matched to the official terminology by exact match on the normalised term or code. Terms with no match are kept and flagged as unclassified. The unclassified remainder is dominated by legacy terms from before 2010 for which FDA publishes no crosswalk to current codes, and we did not invent one. Coding with language models was not part of v1.0.0 and is planned for v1.1.

**Deduplication.** Reports with the same report number are exact duplicates (754,664 reports), and reports with identical normalised narrative text are collapsed before clustering (7,527,796 reports). Near-duplicates are found with MinHash and locality-sensitive hashing [5] using 128 permutations in 32 bands of 4 rows, a Jaccard threshold of 0.85, and blocking by product code and event month. Candidate pairs were verified, and clusters were then refined with exact shingle Jaccard similarity: complete linkage for clusters of up to 100 report numbers, and a star-seed method for larger ones. Clusters are flagged, not deleted. Within a cluster the canonical report is chosen by source priority: manufacturer, then user facility, then voluntary reporter.

**Splits.** Reports dated 2022 or earlier form the training split, 2023 the validation split, and 2024 or later the test split. A further out-of-distribution split (test_ood) is made by holding out 20% of product codes (956 codes). Cluster-level leakage control was enforced and checked by assertion, so that no duplicate cluster spans two splits. In the public sample the year ranges are 1900–2022 (train), 2020–2023 (validation) and 2019–2026 (test); 0.4% of validation and 0.2% of test reports fall outside the nominal years because clusters that straddled a split boundary were moved as a whole to the later split (10,594 clusters, 946,435 reports in the full release). The test_ood split is not temporal and covers 1992–2026. The 288 product codes of test_ood in the sample appear in no other split, and no cluster spans two splits.

# Technical validation

**Quality-control gates.** Seven automated gates were run on the build. Three are hard gates and all three passed, so the overall result is a pass. One soft gate, G4 (unclassified problem terms), failed. G4 asked for at most 2% unclassified product problem rows and found 14.48% (7,609,892 of 52,552,238), for the legacy-term reason given above. For patient problem terms the unclassified share is 32.7%. G3 asked for

at least 95% of reports to carry a problem code and found 99.51%. We report G4 as it is and do not try to recover these terms in v1.0.0.

Table 3
Quality-control gates.

| Gate | Check | Type | Result |
|---|---|---|---|
| G1 | Record count matches the source | Hard | Pass |
| G2 | MDR report key is complete | Hard | Pass |
| G3 | Reports with a problem code | Soft | Pass |
| G4 | Unclassified problem terms | Soft | Fail |
| G5 | Event type distribution | Soft | Pass |
| G6 | Monthly anomalies | Soft | Pass |
| G7 | Deduplication consistency | Hard | Pass |

**Duplicate clusters.** From the 1,186,060 multi-report clusters, 200 were sampled at random and every pairwise Jaccard similarity was recomputed from the raw narratives. This is a programmatic check, not a manual review. In 95.0% of the clusters the minimum pairwise similarity was at least 0.85, the clustering threshold; the mean of the per-cluster minima was 0.9155. The 10 clusters below the threshold all group follow-up submissions that share one report number and whose narratives were rewritten between submissions. All 190 clusters formed by text similarity passed.

**Baselines.** Three baselines were run to show that the data can be used for the tasks they are meant for. They are reference points, not tuned models.

Table 4
Baseline results.

| Task | Method | Result |
|---|---|---|
| Severity classification | Predict the standardised event type from the narrative: TF-IDF (1–2 grams, 50,000 features) and logistic regression, trained on 2,000,001 reports and evaluated on 200,000 reports per split | Macro-F1: validation 0.6594, test 0.6362, test_ood 0.6696 |
| Signal detection | ROR and PRR [6, 7] for pairs of product code and problem code (counts from the training split, minimum 3), scored against recalls (45,113 pairs, 1,038 positive) | AUC: ROR 0.5612, PRR 0.5524 |
| Link prediction | Adamic-Adar [8] on the graph of the top 500 product codes and their problem codes, filtered ranking, 63,541 training edges | Hits@10 0.1229, MRR 0.0597 (4,921 test edges) |

Severity classification uses four classes (death, injury, malfunction, other). Test F1 per class is 0.710 for death, 0.883 for injury, 0.952 for malfunction and 0.0 for the rare class other, which the baseline does not

learn (0.183 on test_ood). Accuracy is 0.953 on validation, 0.930 on test and 0.891 on test_ood. For signal detection, a pair counts as positive if its product code has at least one record in the openFDA device recall endpoint whose recall reason, root cause or product description contains the problem term. The enforcement endpoint was rejected as a reference because its records carry no product code. The area under the precision-recall curve is 0.029 for ROR and 0.027 for PRR, against a positive rate of 2.3%. For link prediction, the task is to rank the problem codes of a product code, with 4,921 edges held out for testing.

The signal-detection AUCs are only slightly above 0.5. Simple disproportionality scores are only slightly better than chance against this reference. We report this as a starting point for better methods.

## Limitations

- MAUDE is a passive surveillance system. Under-reporting and reporting bias are present, and report counts are not incidence rates.
- The openFDA data are unvalidated. The release must not be used for clinical decisions.
- Problem coding in v1.0.0 is exact-match only. 14.48% of product problem terms and 32.7% of patient problem terms are unclassified, and gate G4 failed for this reason.
- Deduplication is conservative. Near-duplicates are flagged, not removed, and 5.0% of the 200 clusters checked contained a pair below the 0.85 similarity threshold.
- Narratives are published by FDA and may still contain personal identifiers.
- The deposit on Zenodo is the 100,003-report sample, which keeps one canonical report per duplicate cluster and has empty device and product problem code tables. The results in Tables 2–4 refer to the full release and cannot be fully reproduced from the sample.

## Declarations

# Ethics statement

The work uses public reports published by the U.S. Food and Drug Administration. No human participants or animals were studied and no new data were collected, so ethical approval was not required. FDA removes direct personal identifiers before publication, but narratives can contain residual identifiers, and users of the data should not attempt to identify individuals.

# CRediT author statement

Maksym Chekerskyi: Conceptualization, Data curation, Methodology, Software, Validation, Writing - original draft, Writing - review and editing.

# Funding

This work received no external funding.


# Declaration of competing interests

The author sells the full MAUDE-Atlas release and other datasets through https://sentineldata.com.ua. The free 100,003-report sample is deposited on Zenodo under CC BY 4.0.

# Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Claude (Anthropic) to draft and edit the text of this article. The dataset and its build code were produced in a session of a separate AI agent (Biomni, Phylo) directed by the author. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

# Data availability

The stratified sample (100,003 reports) and the datasheet are available on Zenodo at https://doi.org/10.5281/zenodo.22204785 under the CC BY 4.0 licence. The full release v1.0.0 (25,711,469 reports) is distributed by the author at https://sentineldata.com.ua. The underlying source data are public U.S. government data from openFDA. The build scripts (fixed random seed 20260830, pandas 2.3.1, pyarrow 21.0.0) are included with the full release.

## References


1. U.S. Food and Drug Administration. Manufacturer and User Facility Device Experience (MAUDE) database. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm
2. U.S Code of Federal Regulations, Title 21, Part 803: Medical Device Reporting
3. U.S. Food and Drug Administration. openFDA: device adverse event API and downloads. https://open.fda.gov/
4. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daume H (2021) III, K. Crawford. Datasheets for datasets. Commun ACM 64(12):86–92
5. Broder AZ (1997) On the resemblance and containment of documents. Proceedings of Compression and Complexity of Sequences 21–29
6. Evans SJW, Waller PC, Davis S (2001) Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf

10(6):483–486

7. van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG (2002) A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 11(1):3–10
8. Adamic LA, Adar E (2003) Friends and neighbors on the Web. Soc Networks 25(3):211–230
9. Chekerskyi M, CODE SENTINEL DATA) (2026) (. MAUDE-Atlas v1.0.0 - Sample (100,003 FDA Medical Device Reports). Zenodo https://doi.org/10.5281/zenodo.22204785

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Specificationstable.docx