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.

A regulatory medical device dataset with risk labels and an image-linked subset from the NMPA registry

Yu Han

yu.han@eng.ox.ac.uk

University of Oxford

Aaron Ceross

University of Birmingham

Jeroen H.M. Bergmann

University of Oxford

data-descriptor

Keywords:

Posted Date: March 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9168242/v1

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

Additional Declarations: No competing interests reported.

A regulatory medical device dataset with risk labels and an image-linked subset from the NMPA registry

Yu Han [1,*], Aaron Ceross [2], and Jeroen H.M. Bergmann [1,3]

[1]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[2]Birmingham Law School, University of Birmingham, Birmingham, United Kingdom
[3]Department of Technology and Innovation, University of Southern Denmark, Odense, Denmark
[*]Corresponding author: Yu Han (yu.han@eng.ox.ac.uk)

Abstract

We present NMPA-MedDevice, a regulatory dataset derived from China's National Medical Products Administration (NMPA) Unique Device Identification (UDI) registry. The release comprises four components: (1) a frozen raw snapshot of the NMPA UDI registry (66,472 records, July 2024); (2) a reproducibly cleaned text-and-metadata corpus of approximately 52,000 unique device records with risk class labels deterministically derived from the ninth character of the NMPA registration number; (3) a curated image-linked subset of 1,005 devices (Class I/II/III = 39/462/504) with precomputed text and image feature embeddings; and (4) an external temporal validation set of 300 devices from a later registry update (October–November 2025). All textual data, derived labels, the cleaned corpus, preprocessing scripts, dataset splits, and precomputed features are publicly deposited. Raw product images are not redistributed due to copyright restrictions; precomputed embeddings and image retrieval scripts are provided instead.

Background & Summary

Medical device risk classification is a central component of regulatory oversight. Regulatory authorities, including the U.S. Food and Drug Administration (FDA), European Union (EU) national competent authorities under the Medical Device Regulation (MDR), and China's National Medical Products Administration (NMPA), assign devices to tiered risk categories (typically Class I, II, and III) based on intended use, invasiveness, and clinical criticality[1,2]. These classifications determine market approval pathways, the stringency of clinical evidence requirements, and post-market surveillance obligations[3]. Manual classification by regulatory experts is labour-intensive, time-consuming, and susceptible to inter-rater variability[4], particularly as the volume and diversity of medical device submissions continue to grow with the proliferation of digital health products and AI-based software as medical devices[5].

Existing computational studies in medical device classification have relied predominantly on text-only inputs from single jurisdictions, with limited reproducibility and no standardised benchmarks[4].

Publicly available datasets combining regulatory device descriptions with structured metadata and risk classification labels from the Chinese NMPA registry have not previously been reported.

Here we describe NMPA-MedDevice, a regulatory dataset constructed from the NMPA UDI Database Full Release (July 2024)[6]. The Unique Device Identification (UDI) system, mandated under international frameworks such as the International Medical Device Regulators Forum (IMDRF) guidelines[8] and implemented in China since 2019[7], requires manufacturers to assign standardised identifiers to all registered medical devices. The NMPA UDI registry is publicly accessible and covers all medical devices registered in China.

The release is organised into four components. The first is the frozen raw registry snapshot (66,472 records). The second is a cleaned text-and-metadata corpus of approximately 52,000 unique device records, obtained after removing duplicates, incomplete submissions, and placeholder entries; risk class labels are deterministically derived from the registration number and included in the released file. The third is a curated image-linked subset of 1,005 devices with paired textual descriptions, regulatory metadata fields, verified risk classification labels, and precomputed multimodal feature embeddings, validated in a published classification study[11]. The fourth is an external temporal validation set of 300 devices from a later registry update. The cleaned corpus preserves the original NMPA text fields.

The release includes fixed cross-validation splits and random seeds to support reproducible experimentation. The large cleaned corpus can be used for fully supervised training (labels are included) or, by withholding labels, as a pool for semi-supervised and self-training experiments[12,13]. Risk class labels are derived programmatically from registration numbers, so annotation does not depend on subjective expert judgement. The accompanying code release includes the preprocessing pipeline, feature extraction scripts, and image retrieval utilities.

The dataset was developed as part of a broader research program on AI-assisted regulatory classification of medical devices. A companion modeling study using the curated subset has been reported elsewhere.[11].

While the dataset was initially compiled in the context of a machine learning study, the released resource is intended to support a broad range of future research including regulatory science, natural language processing of Chinese regulatory text, multimodal device classification, and cross-jurisdictional regulatory analysis.

Methods

Figure 1 provides an overview of the dataset construction pipeline, from raw registry download through cleaning, label derivation, subset curation, and final release.

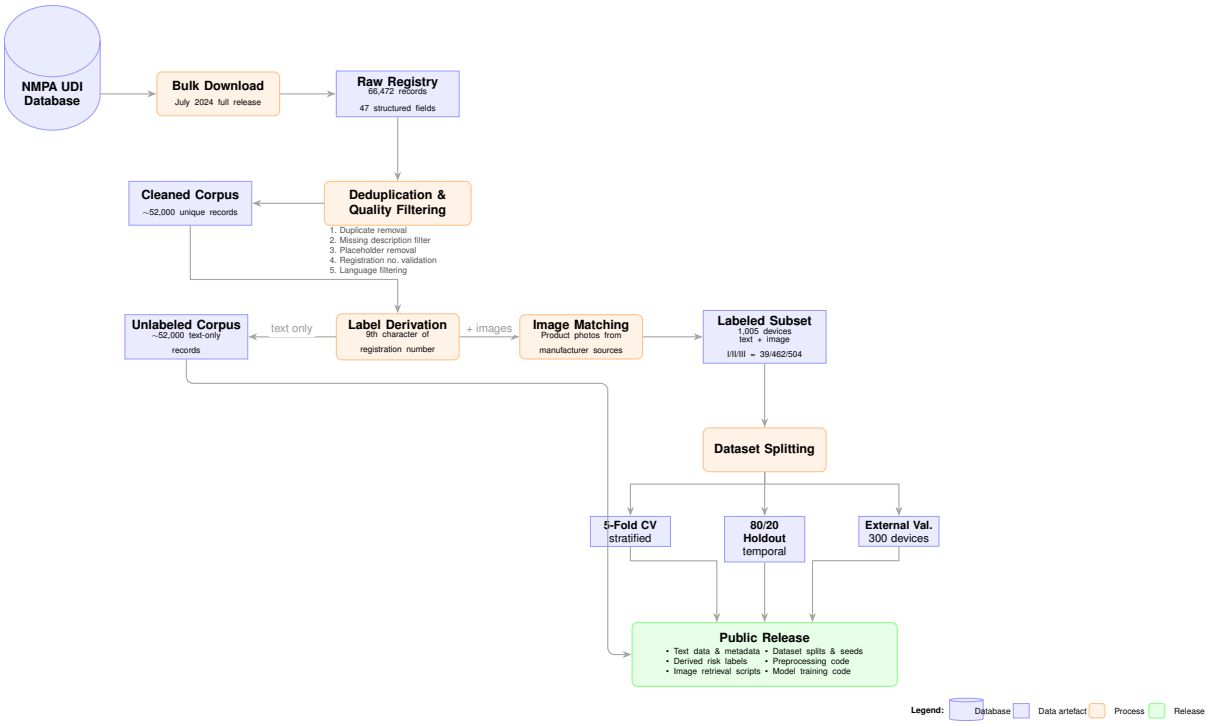


Figure 1: Overview of the NMPA-MedDevice dataset construction pipeline. The process begins with a bulk download from the NMPA UDI Database and proceeds through deduplication, quality filtering, label derivation, image matching, and dataset partitioning. Rectangles represent data artefacts; rounded boxes represent processing steps; the cylinder denotes the source database.

Data Source

The primary data source is the NMPA Unique Device Identification (UDI) Database, a publicly accessible registry maintained by the National Medical Products Administration of China[6]. The UDI system was established under the NMPA's 2019 regulation on medical device unique identification[7] and follows the International Medical Device Regulators Forum (IMDRF) UDI guidance[8], which recommends globally harmonised device identification to improve traceability, facilitate post-market surveillance, and support supply chain integrity. The U.S. FDA[9] and the European Union[2] have implemented analogous systems, making the NMPA UDI registry part of a growing international ecosystem of structured device registries.

Specifically, we used the `UDID_FULL_RELEASE_20240701` bulk download, which contains structured records for all medical devices registered in China as of 1 July 2024. The registry is publicly available at https://udi.nmpa.gov.cn/download.html and permits bulk download in comma-separated values (CSV) format encoded in UTF-8. Each record in the raw registry contains 47 structured fields spanning six functional categories: device identification (12 fields), product information (7 fields), regulatory classification (4 fields), manufacturer information (4 fields), safety and clinical information (7 fields), production identifiers (7 fields), and version control metadata (6 fields). The complete field listing is provided in Table 4. The raw download contains 66,472 records.

Data Cleaning and Deduplication

The raw registry data underwent a multi-stage cleaning process designed to remove noise while preserving the natural distribution of device types and risk classes. The cleaning pipeline consists of four

sequential steps, with record counts after each step reported in Table 1. First, entries with empty or null product description fields (`cpms`) were excluded, as the product description is the primary text field for downstream NLP tasks. Second, records containing generic or template text rather than actual device descriptions were identified through pattern matching (e.g., descriptions consisting solely of the string "暂无" [not yet available] or containing fewer than 5 Chinese characters) and removed. Third, entries with predominantly non-Chinese product descriptions (less than 30% Chinese characters, e.g., English-only or mixed-script entries) were removed to ensure linguistic consistency of the text corpus. Fourth, near-duplicate records sharing identical product description text were identified and deduplicated, retaining only one record per unique description. Registration number validity was verified after cleaning: all 52,251 retained records have well-formed registration numbers with valid ninth-character risk codes (1, 2, or 3), so no additional exclusion for malformed registration numbers was necessary.

Table 1 summarises the record counts after each cleaning step. The final cleaned corpus contains 52,251 unique device records with complete textual fields. All cleaning scripts are provided in the code release to ensure full reproducibility.

Table 1: Data cleaning audit trail. Each row shows the cumulative effect of one cleaning step applied sequentially to the raw registry download.

Step	Operation	Removed	Remaining
0	Raw registry download (July 2024)	—	66,472
1	Empty / null product description removal	2	66,470
2	Placeholder and short text removal ($<$ 5 Chinese characters)	5,588	60,882
3	Non-Chinese text removal ($<$ 30% Chinese characters)	1,841	59,041
4	Near-duplicate description removal (exact `cpms` match)	6,790	52,251

Records with malformed registration numbers were not encountered after deduplication; all 52,251 retained records have well-formed registration numbers with valid ninth-character risk codes.

Label Derivation

Risk classification labels were derived programmatically from the NMPA registration number encoding scheme. Chinese medical device registration numbers follow a standardised format defined by the NMPA[10]. The ninth character of the registration number field (`zczbhhzbapzbh`) encodes the device's regulatory risk class according to the NMPA's internal coding system. For example, a registration number of the form 国械注准`20161xxxxxx` maps to Class I (low risk), 国械注准`20162xxxxxx` to Class II (moderate risk), and 国械注准`20163xxxxxx` to Class III (high risk). Figure 2 illustrates the structure of the registration number and the mapping from the ninth character to the three-tier risk classification.

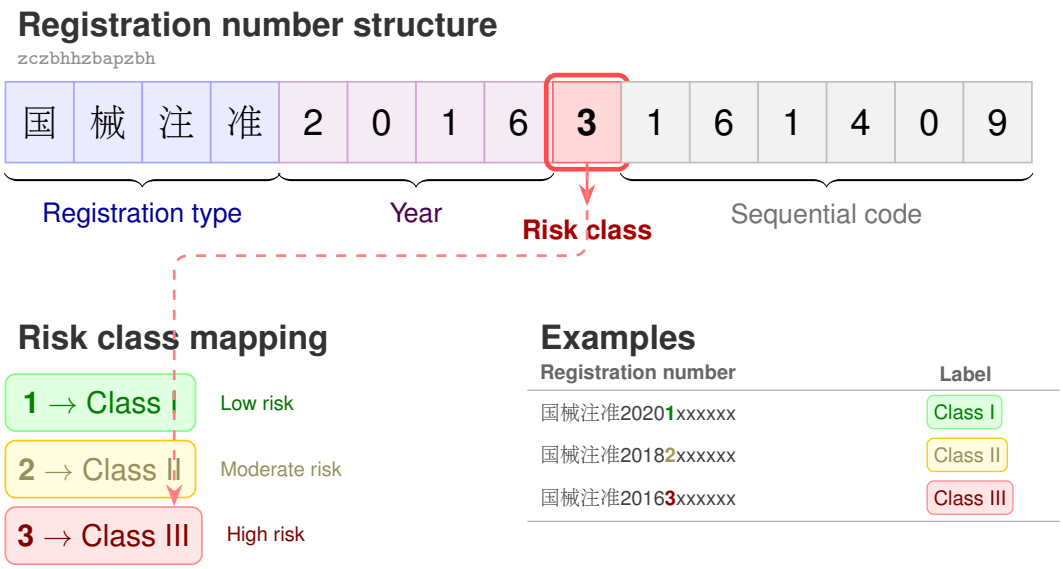


Figure 2: Derivation of risk class labels from NMPA registration numbers. The ninth character of the standardised registration number (`zczbhhzbapzbh`) encodes the device's regulatory risk class: 1 = Class I (low risk), 2 = Class II (moderate risk), 3 = Class III (high risk). This mapping is deterministic and reproducible.

Devices were assigned to Class I, Class II, or Class III based on this character. All 52,251 retained records have valid ninth-character risk codes, so no additional exclusion was necessary. The label derivation script (`scripts/label_derivation.py`) is included in the release.

Labeled Subset Curation

From the cleaned corpus, a curated subset of 1,005 devices was assembled for supervised multimodal classification. The subset contains 39 Class I (3.9%), 462 Class II (46.0%), and 504 Class III (50.1%) devices, reflecting the natural skew of the registry toward higher-risk classes, though with some distributional shift relative to the full registry due to image availability constraints (see Selection Bias Assessment). Class I devices remain underrepresented relative to Class II and III, though their proportion in the curated subset (3.9%) is higher than in the full registry (1.5%), likely because Class I devices with available images tend to be distinctive products with prominent manufacturer web presence. The selection criteria for inclusion in the labeled subset were: (1) complete and non-trivial product description text, (2) a valid and unambiguous registration number encoding the risk class, and (3) availability of at least one product image from publicly accessible manufacturer sources.

The labeled subset spans 22 aggregated NMPA device categories, including imaging and diagnostic equipment, surgical instruments, orthopaedic implants, cardiovascular devices, in vitro diagnostic reagents, rehabilitation equipment, and consumable supplies.

Image Collection and Processing

Product images were manually retrieved from manufacturers' official websites, online product listings, and publicly accessible regulatory submission portals. Each image was matched to its registry entry via the unique device identifier (`zxxsdycpbs`) field, with the matching index stored in `matched_products.csv`.

The matching process was performed independently by two research assistants, and discrepancies were resolved by a third reviewer.

The resulting paired samples consist of the NMPA-derived product description text and a corresponding product image. The image types comprise product photographs from manufacturer websites and regulatory submissions (approximately 72%), technical schematics including engineering drawings and exploded views (approximately 14%), contextual usage photographs showing devices in clinical settings (approximately 4%), and other categories such as packaging photographs and close-up views of components (approximately 10%). Over 95% of the images are high-resolution studio photographs with clean backgrounds and adequate lighting. Image quality was assessed using three metrics: pixel resolution, Laplacian variance (a measure of image sharpness), and visual clarity. Based on these metrics, images were categorised into three quality tiers (Table 2).

Table 2: Image quality distribution in the labeled subset ($N = 1{,}005$).

Quality Tier	**Count (%)**	**Criteria**
High	409 (40.7%)	Resolution $\geq$ 500k pixels, Laplacian variance $\geq$ 100
Medium	546 (54.3%)	Resolution $\geq$ 100k pixels, Laplacian variance $\geq$ 50
Low	50 (5.0%)	Below medium thresholds

Figure 3 visualises the image type and quality distributions.

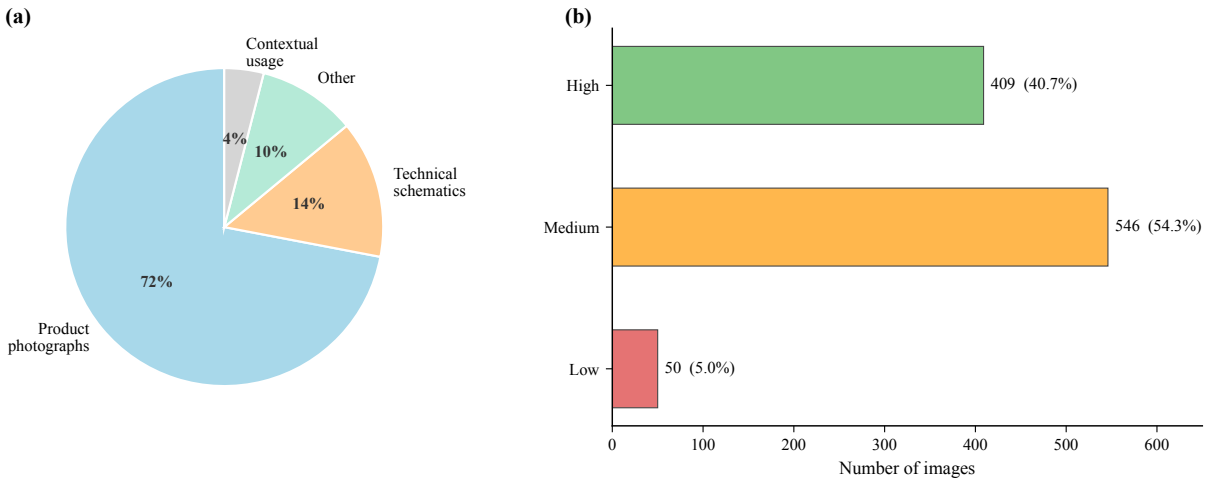


Figure 3: Image characteristics in the labeled subset ($N = 1{,}005$). (a) Distribution of image types by source category. (b) Image quality tier distribution based on resolution and sharpness metrics.

All images were resized to 256×256 pixels and centre-cropped to 224×224 pixels, with pixel values normalised using ImageNet channel-wise means and standard deviations[17]. Training-time augmentation details are described in the companion study[11].

Due to copyright restrictions on manufacturer-provided product images, raw images are not redistributed with the dataset. Instead, following established practice for sharing visual data without distributing copyrighted images[20,21], the release includes: (1) **pre-extracted feature embeddings** in HDF5 format (BERT-base-Chinese [CLS] token embeddings, 768 dimensions, for the text modality; EfficientNet-B5 average-pool embeddings, 2,048 dimensions, for the image modality), which enable reproducibility of downstream classification experiments without requiring access to the original images;

(2) **record–image linkage metadata** (`matched_products.csv`); and (3) a **best-effort image retrieval script** (`scripts/image_retrieval.py`) that attempts to download product images from publicly accessible manufacturer websites using the device identifiers in the linkage file. Because manufacturer web pages may change or become unavailable over time, image reconstruction is not guaranteed and may degrade due to link rot.

Cleaned Corpus Preparation

The cleaned corpus comprises the 52,251 device records retained after the cleaning pipeline described above (Table 1). These records contain complete textual fields and deterministically derived risk class labels, but do not have associated images.

In the companion study[11], the derived labels for this corpus were withheld to serve as a pool for self-training experiments. In this release, the derived labels are included for all retained records, so the corpus can be used for either fully supervised or semi-supervised experiments.

Text Preprocessing

Product descriptions are written in Chinese and vary in length and vocabulary. The preprocessing pipeline applies Unicode normalisation (NFC form), removal of formatting artefacts, and punctuation standardisation, followed by Chinese word segmentation using Jieba[14]. Tokenisation for feature extraction uses the `bert-base-chinese` tokenizer[15]. The companion study[11] applied additional synonym-based augmentation during training; this augmentation is not applied to the released text data. All preprocessing scripts are included in the code release.

Feature Extraction

Features were extracted from both modalities using pre-trained models. Text embeddings (768-dimensional) were obtained from `bert-base-chinese`[15] [CLS] token representations. Image embeddings (2,048-dimensional) were obtained from EfficientNet-B5[16] average-pool features using frozen ImageNet-pretrained weights. The extracted features are provided as pre-computed HDF5 files in the `features/` directory, along with the extraction script (`scripts/extract_features.py`). Following practice in other datasets[20,21], pre-extracted embeddings are distributed instead of raw images.

Dataset Splitting

To support reproducible benchmarking, we provide three types of fixed dataset splits. The first is a five-fold stratified cross-validation scheme in which the labeled subset ($N = 1{,}005$) is divided into five folds, each preserving the overall class distribution ($\sim$3.9% Class I, $\sim$46.0% Class II, $\sim$50.1% Class III); fold assignments are stored in `splits/cv_fold_indices.json` with fixed random seeds. The second is a temporal holdout split (80/20), an internal validation partition based on data acquisition timestamps with 813 training samples and 192 validation samples, designed to simulate prospective evaluation; due to the small number of Class I devices ($N = 39$), all Class I samples are retained in the training set, meaning that the temporal holdout evaluates prospective generalisation for Class II and III only. Split assignments are stored in `splits/split_metadata.json`. The third is an external temporal validation set comprising 300 devices from the NMPA weekly update (`UDID_WEEKLY_UPDATE_20251027_20251102`), released

after the main dataset was constructed; this set provides an independent evaluation of cross-temporal generalisation, with class distribution (Class I = 12, Class II = 138, Class III = 150) approximately matching the labeled subset. All split indices, metadata, and random seeds are provided in the release to ensure exact reproducibility. The provided splits allow direct comparison of results across studies.

Data Records

The NMPA-MedDevice dataset is deposited at Figshare (https://doi.org/10.6084/m9.figshare.31322272)[22]. The release comprises four distinct data components and an accompanying script directory, summarised in Table 3. Below we describe each component in detail.

Table 3: Overview of released data components.

Component	Records	Labels	Images	Release Form
Raw registry snapshot	66,472	Derivable	No	Direct deposit (CSV)
Cleaned corpus	∼52,000	Included	No	Direct deposit (CSV)
Curated image-linked subset	1,005	Included	Embeddings only	Direct deposit (CSV + HDF5)
External validation set	300	Included	Embeddings only	Direct deposit (XLSX + HDF5)
Scripts & split indices	—	—	—	Direct deposit

Labeled Subset (`labeled/`)

`sampled_data.csv` — A comma-separated values file containing 1,005 rows (one per device) and 50 columns. The first 47 columns correspond to the original NMPA registry fields described in Table 4. Three additional derived columns are appended: `risk_class` (the assigned risk classification label: "I", "II", or "III"), `n_row` (the original row index in the full 66,472-record registry), and `n_number` (a sequential sample number within the labeled subset). The file is encoded in UTF-8 with a header row. All text fields preserve the original Chinese content without translation.

`matched_products.csv` — A CSV file providing the matching index between device records and product images. Columns include the device identifier (`device_id`), risk class, target product name, matched image filename, matched product name, and text similarity scores used during the matching process. This file links each device record to its corresponding product image and is required to reconstruct the multimodal dataset when used in conjunction with the image retrieval script.

Pre-Extracted Features (`features/`)

Pre-computed deep learning embeddings are stored in HDF5 format so that classification experiments can be reproduced without raw images:

`labeled_features.h5` (11 MB) — Contains BERT-base-Chinese text embeddings (shape: $1{,}005 \times 768$, [CLS] token pooling, max sequence length 128) and EfficientNet-B5 image embeddings (shape: $1{,}005 \times 2{,}048$, average pooling, ImageNet-pretrained weights, input size 456×456). Additional datasets include a boolean `has_image` mask, risk class labels, device indices, and registration numbers. All extraction parameters are stored as HDF5 attributes.

`external_validation_features.h5` (2.1 MB) — Contains text and image embeddings for the 300 external validation devices, using the same models and parameters as the labeled features file.

Samples without available images have zero-valued image embedding vectors; the `has_image` boolean mask should be used to identify them. Feature extraction details are also recorded as HDF5 file-level attributes for provenance.

Raw Registry and Cleaned Corpus (`corpus/`)

`nmpa_full_66k.csv` — The complete NMPA UDI registry as downloaded on 1 July 2024, containing 66,472 rows and 47 columns. This file includes all records prior to deduplication and quality filtering. Researchers may apply the provided cleaning scripts (`scripts/preprocessing.py`) to reproduce the cleaned corpus from scratch, or they may apply custom cleaning pipelines suited to their specific research objectives.

`nmpa_cleaned_52k.csv` — The cleaned corpus produced by the preprocessing pipeline described in the Methods section, containing approximately 52,000 rows. Each row retains the original 47 NMPA registry fields plus a derived `risk_class` column (Class I, II, or III), deterministically extracted from the ninth character of the registration number. This file is deposited directly so that users need not re-execute the cleaning pipeline to access the core text-and-metadata corpus.

Dataset Splits (`splits/`)

`cv_fold_indices.json` — A JSON file containing five-fold stratified cross-validation indices. The file structure is a dictionary with keys "fold_0" through "fold_4", each mapping to a dictionary with "train" and "test" keys containing lists of integer indices into `sampled_data.csv`. The random seed used for fold generation is recorded in the file metadata.

`split_metadata.json` — A JSON file containing metadata for the 80/20 temporal holdout split, including the split criterion (acquisition timestamp), random seed, and class distribution per partition.

`train_files.csv` and `validation_files.csv` — CSV files listing the device identifiers assigned to the training ($N = 813$) and validation ($N = 192$) partitions of the temporal holdout split, respectively.

External Validation (`external_validation/`)

`external_validation_product.xlsx` — An Excel file containing 300 device records from the NMPA weekly update covering 27 October to 2 November 2025. These records were not available during the construction of the main dataset and serve as an independent temporal validation set. The file contains the same 47 NMPA registry fields as the main corpus, plus a derived `risk_class` column.

Scripts (`scripts/`)

`image_retrieval.py` — A best-effort Python script that attempts to download product images from publicly accessible manufacturer websites. The script uses product names and manufacturer identifiers from `sampled_data.csv` to construct search queries and retrieve matching images. It accepts `matched_products.csv` as a validation reference and outputs images in JPEG format. Because manufacturer web pages may change or become unavailable, successful retrieval is not guaranteed. Usage instructions and dependency requirements are documented in the repository README.

`preprocessing.py` — The complete text cleaning and deduplication pipeline, implementing all five steps described in the Methods section. Input: raw CSV from the NMPA UDI Database. Output: cleaned corpus in CSV format.

`label_derivation.py` — A script that extracts risk class labels from the ninth character of NMPA registration numbers. Input: any CSV file containing a `zczbhhzbapzbh` column. Output: the same file with an appended `risk_class` column.

`extract_features.py` — The feature extraction pipeline used to generate the pre-computed embeddings in `features/`. Extracts BERT-base-Chinese [CLS] token embeddings (768-d) from text fields and EfficientNet-B5 average-pool features (2,048-d) from product images, saving results as compressed HDF5 files. Requires `torch`, `torchvision`, `transformers`, and `h5py`.

Field Descriptions

Table 4 describes all 50 data fields (47 original NMPA registry fields plus 3 derived fields), grouped by functional category with English translations. The primary text field for NLP tasks is `cpms` (free-text product description in Chinese). The primary label field is `risk_class` (available in the labeled subset and external validation set).

Table 4: Data fields in the NMPA-MedDevice dataset, grouped by function. Fields marked with † are derived and available only in the labeled subset.

Field	Type	Description
Device Identification (12 fields)		
`zxxsdycpbs`	String	Primary product identifier (UDI-DI)
`cpbsbmtxmc`	String	Identifier coding scheme (e.g. GS1, MA code)
`cpbsfbrq`	Date	Identifier issue date
`zxxsdyzsydydsl`	Integer	Total number of UDI identifiers for this device
`sydycpbs`	String	Secondary / alternate product identifier
`bszt`	String	Identifier carrier type (barcode / QR code)
`sfyzcbayz`	Boolean	Whether the device has a regulatory record
`zcbacpbs`	String	Registration record product identifier
`sfybtzjbs`	Boolean	Whether there is a packaging identifier
`btcpbsyzxxsdy-cpbssfyz`	Boolean	Whether packaging ID matches primary ID
`btcpbs`	String	Packaging product identifier
`deviceRecordKey`	String	Internal database record key
Product Information (7 fields)		
`cpmctymc`	String	Generic / common product name (Chinese)
`spmc`	String	Trade / commercial product name
`ggxh`	String	Specifications and model number

Continued on next page

Table 4: (continued)

Field	Type	Description
cpms	String	Free-text product description (Chinese); **primary text field for NLP tasks**
cphhhbh	String	Product catalogue / batch number
sfwblztlcp	Boolean	Whether the device contains biological material
cplb	String	Product category (consumable / equipment)
Regulatory Classification (4 fields)		
yflbm	String	Legacy / original classification code
qxlb	String	Device type category
flbm	String	NMPA classification catalogue code (e.g. 13-01-02)
zczbhhzbapzbh	String	Full NMPA registration / record number; 9th character encodes risk class
Manufacturer Information (4 fields)		
tyshxydm	String	Unified Social Credit Code (manufacturer ID)
ylqxzcrbarmc	String	Manufacturer / registrant name (Chinese)
ylqxzcrbarywmc	String	Manufacturer / registrant name (English)
ybbm	String	Standardised identifier code
Safety and Clinical Information (7 fields)		
cgzmraqxgxx	String	MR safety information status
sfbjwycxsy	Boolean	Whether for single use only
zdcfsycs	Integer	Maximum number of reuses
sfwwjbz	Boolean	Whether aseptic packaging is required
syqsfxyjxmj	Boolean	Whether sterilisation is needed before use
mjfs	String	Sterilisation method
qtxxdwzlj	String	URL for additional information
Production Identifiers (7 fields)		
tsrq	Date	Special / additional date field
scbssfbhph	Boolean	Whether production ID includes batch number
scbssfbhxlh	Boolean	Whether production ID includes serial number
scbssfbhscrq	Boolean	Whether production ID includes manufacturing date
scbssfbhsxrq	Boolean	Whether production ID includes expiry date
tscchcztj	String	Special storage / transport conditions
tsccsm	String	Special storage / transport notes
Version Control (6 fields)		
versionNumber	Integer	Record version number
versionTime	Date	Version timestamp

Continued on next page

Table 4: (continued)

Field	Type	Description
`versionStauts`	String	Version status (new / updated)
`correctionNumber`	Integer	Number of corrections
`correctionRemark`	String	Correction notes
`correctionTime`	Date	Correction timestamp
Derived Fields (labeled subset only)		
`risk_class` †	String	Risk class label (Class I / II / III), derived from 9th character of `zczbhhzbapzbh`
`n_row` †	Integer	Original row index in full registry
`n_number` †	Integer	Sequential sample number in labeled subset

Table 5: Class distribution across dataset components.

Component	Class I	Class II	Class III	Total	Fields
Raw registry snapshot	Labels derivable from registration number			66,472	47
Cleaned corpus	Labels included (derived)			∼52,000	47 + 1 derived
Curated image-linked subset	39 (3.9%)	462 (46.0%)	504 (50.1%)	1,005	47 + 3 derived
External validation	12 (4.0%)	138 (46.0%)	150 (50.0%)	300	47 + 1 derived

Technical Validation

Label Accuracy Verification

Risk class labels were derived deterministically from the ninth character of the NMPA registration number, which is assigned by the NMPA during the official device registration process. To verify the reliability of this derivation, we performed a manual audit of 100 randomly sampled records from the labeled subset. Two independent reviewers cross-referenced the derived labels against the NMPA online registry portal (https://udi.nmpa.gov.cn/) and observed 100% agreement in the audited sample, supporting the reliability of the derivation procedure. This spot check is consistent with the ninth-character encoding being applied consistently, though it does not constitute an exhaustive verification of every record in the corpus.

Data Completeness and Integrity

We assessed data completeness across both dataset components. Table 6 summarises the key statistics.

Table 6: Data completeness assessment.

Metric	Labeled Subset	Full Registry
Total records	1,005	66,472
Non-empty `cpms`	100.0%	95.3%
Valid registration number	100.0%	97.8%
Description length, median [IQR]	45 [22–75] chars	35 [20–70] chars
Description length, mean ± SD	65 ± 70 chars	59.7 ± 74.7 chars
Unique manufacturers	387	12,841
NMPA device categories represented	22	22

In the curated image-linked subset ($N = 1{,}005$), all records contain non-empty product descriptions (`cpms`) and valid registration numbers (`zczbhhzbapzbh`). In the cleaned corpus (∼52,000 records after cleaning), all entries passed the minimum description length filter and language validation checks. No duplicate registration numbers remain after deduplication. All retained records have well-formed registration numbers with valid ninth-character risk codes, and the derived `risk_class` labels are included.

Descriptive Statistics

Figure 4 presents the distribution of product description lengths and class frequencies across the dataset components.

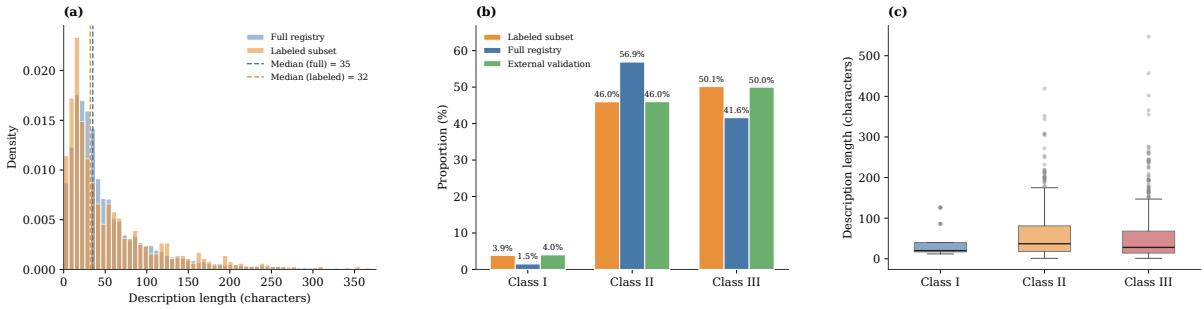


Figure 4: Descriptive statistics of the NMPA-MedDevice dataset. (a) Distribution of product description lengths in the labeled subset (orange) and full registry (blue); vertical dashed lines indicate medians. (b) Class distribution across dataset components. (c) Description length by risk class in the labeled subset.

Selection Bias Assessment

Because the labeled subset requires image availability from manufacturer websites, it may overrepresent manufacturers with stronger digital presence. We quantified this bias by comparing the labeled subset against the full NMPA registry (66,472 devices) along three dimensions[11]. In terms of web presence, 100% of the labeled subset had retrievable images, compared to only 34.2% of a random 500-device sample from the full registry ($\chi^2(1) = 287.6$, $p < 0.001$). Regarding device category distribution, significant differences were observed across the 22 aggregated NMPA device categories ($\chi^2(21) = 47.3$, $p < 0.001$), with imaging devices (18.2% vs. 12.1%) and surgical instruments (24.6% vs. 22.3%) overrepresented, and low-cost consumables (9.4% vs. 15.7%) underrepresented in the labeled subset. The class distribution in the curated subset also differs from the full registry: Class III devices are overrepresented (50.1%

vs. 41.6%) and Class II devices underrepresented (46.0% vs. 56.9%), likely because higher-risk devices tend to have more prominent manufacturer web presence. A modest shift toward longer descriptions was also observed in the labeled subset (median 45 vs. 35 characters; KS statistic $= 0.089$, $p < 0.01$; $t = 2.34$, $p < 0.05$). These biases are inherent to the image-retrieval process and should be considered when interpreting results from the labeled subset. The cleaned corpus, which does not require image availability, provides a more representative sample of the full registry. Table 7 summarises the comparison.

Table 7: Representativeness assessment: labeled subset ($N = 1{,}005$) vs. full NMPA registry ($N = 66{,}472$).

Dimension	**Labeled Set**	**Full Registry**	**Significance**
Image availability	100.0%	34.2%	$p < 0.001$
Class I proportion	3.9%	1.5%	
Class II proportion	46.0%	56.9%	$p < 0.001$
Class III proportion	50.1%	41.6%	
Imaging & diagnostics	18.2%	12.1%	
Surgical instruments	24.6%	22.3%	$p < 0.001$
Consumables	9.4%	15.7%	
Description length (med.)	45 chars	35 chars	$p < 0.01$

Benchmark Classification Results

To illustrate the utility of the dataset for machine learning research, we report representative benchmark results obtained using the curated subset. These results are drawn from a companion methodological study. Table 8 reports selected classification results from the companion study[11], obtained using five-fold stratified cross-validation on the provided splits.

Table 8: Selected benchmark results on the curated subset under five-fold stratified cross-validation. ST = with self-training. Full results are reported in the companion study[11].

Modality	**Model**	**Accuracy**	**Macro-F1**	**AUROC**	**Brier**
Text-only	SVM	0.784	0.846	0.919	0.287
Image-only	SVM	0.515	0.335	0.484	0.539
Multimodal	Cross-attn. SVM	0.803	0.782	0.914	0.298
Multimodal	Cross-attn. SVM + ST	**0.916**	**0.875**	**0.989**	**0.108**

Text-only classification outperformed image-only classification. Detailed results are reported in the companion study[11].

External Temporal Validation

The external validation set of 300 devices from the NMPA weekly update (October–November 2025) was used to assess cross-temporal generalisation. The best-performing text-only configuration from the companion study achieved 70.4% accuracy (95% CI: 0.649–0.756) on this external set, compared to 91.6% under internal cross-validation, consistent with distribution shift between the July 2024 snapshot and the October 2025 update. Full results are in the companion study[11].

Usage Notes

Recommended Use Cases

The dataset supports text-based and multimodal medical device risk classification using supervised, semi-supervised, or self-training approaches. The cleaned corpus (∼52,000 records with labels) can be used for Chinese-language NLP tasks such as text classification or information extraction; structured metadata fields (e.g., device category, sterilisation method) are available as auxiliary labels. Pre-extracted BERT and EfficientNet embeddings in `features/` enable multimodal experiments without image acquisition; raw images can be reconstructed using the provided retrieval scripts. The NMPA risk classification scheme is structurally comparable to FDA [18] and EU MDR [2] frameworks.

Getting Started

Researchers can begin using the dataset by downloading it from the repository (https://doi.org/10.6084/m9.figshare.31322272). For text-only experiments, `labeled/sampled_data.csv` should be loaded with the `cpms` column as input and the `risk_class` column as the target variable, using the cross-validation indices from `splits/cv_fold_indices.json` for evaluation. Multimodal experiments can be conducted using the pre-extracted features by loading `features/labeled_features.h5` with `h5py` and concatenating the `text_embeddings` (768-d) and `image_embeddings` (2,048-d) arrays, using the `has_image` mask to handle samples without images. Alternatively, for multimodal experiments with raw images, `scripts/image_retrieval.py` can be used to download product images, with `labeled/matched_products.csv` providing the image-to-record matching; features may then be re-extracted using `scripts/extract_features.py`. For semi-supervised experiments, `corpus/nmpa_cleaned_52k.csv` can be used directly; researchers may withhold the included `risk_class` column to simulate an unlabeled pool, or use it for fully supervised training on the full corpus. External validation can be performed on `external_validation/external_validation_product.xlsx`.

Limitations and Considerations

Users should be aware of several limitations. The labeled subset overrepresents manufacturers with stronger digital presence, as image availability from manufacturer websites was a prerequisite for inclusion (see Selection Bias Assessment above); results obtained on this subset may therefore not generalise to the full population of NMPA-registered devices, though the cleaned corpus provides a more representative sample. All product descriptions are in Chinese, and cross-lingual applications will require translation or multilingual models such as XLM-RoBERTa [19] or multilingual BERT. The dataset covers only NMPA-registered devices; cross-jurisdictional generalisation has not been validated, although the risk classification framework is conceptually aligned with FDA and EU MDR systems. Class I devices are underrepresented (3.9% of the curated subset), reflecting real-world registry distributions, and researchers should consider class-weighted losses, oversampling strategies, or evaluation metrics robust to imbalance (e.g., macro-F1, balanced accuracy). The main corpus reflects the NMPA registry as of July 2024, with the external validation set covering October–November 2025; device registrations, descriptions, and classifications may change over time. The multimodal component of the dataset is substantially less stable as a reusable public resource than the text and metadata components: raw images are not re-

distributed, and the retrieval script depends on continued availability of manufacturer web pages, which may change or go offline over time. Researchers who wish to use the image modality are advised to download and archive images promptly after dataset access. The precomputed embeddings provide a frozen alternative that does not depend on external web availability.

Ethical and Legal Considerations

All data in this release are derived from publicly accessible government registries and manufacturer websites. No personally identifiable information is included. The NMPA UDI Database is a public resource made available by the Chinese government for transparency in medical device regulation[6]. Product descriptions and registration numbers are publicly searchable on the NMPA portal. The dataset does not contain protected health information or clinical data. Image retrieval scripts access only publicly available web pages and respect manufacturer robots.txt policies.

Data Availability

The NMPA-MedDevice dataset is publicly available at Figshare (https://doi.org/10.6084/m9.figshare.31322272)[22]. The deposit includes the raw registry snapshot (`corpus/nmpa_full_66k.csv`), cleaned corpus (`corpus/nmpa_cleaned_52k.csv`), curated image-linked subset (`labeled/sampled_data.csv`), external temporal validation set (`external_validation/external_validation_product.xlsx`), pre-extracted feature embeddings (`features/`), fixed dataset splits (`splits/`), and all preprocessing and utility scripts (`scripts/`). The source registry is publicly accessible at the NMPA UDI Database (https://udi.nmpa.gov.cn/download.html).

Code Availability

All code used in dataset construction, preprocessing, and feature extraction is publicly available as part of the Figshare deposit (https://doi.org/10.6084/m9.figshare.31322272) in the `scripts/` directory. This includes data cleaning and deduplication scripts, label derivation code, the text preprocessing pipeline, and the image retrieval utility. Classification model code for the companion study[11] is available at https://github.com/RegAItool/Multimodal-regulatory.
The code was developed using Python 3.9 with PyTorch 1.12, Scikit-learn 1.1, Transformers 4.24, and `efficientnet_pytorch` 0.7. Dependency specifications (`environment.yml`, `requirements.txt`) are included in the GitHub repository.

Acknowledgements

We thank the National Medical Products Administration (NMPA) of China for making the UDI Database publicly accessible.

Funding

This work was supported by the National Institute for Health and Care Research (NIHR) HealthTech Research Centre for Community Healthcare at Oxford Health NHS Foundation Trust. The views expressed

are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Author Contributions

Y.H. conceived and designed the study, collected and curated the dataset, developed the preprocessing pipeline and classification models, conducted all experiments, and wrote the manuscript. A.C. contributed to the regulatory analysis and reviewed the manuscript. J.H.M.B. supervised the project and reviewed the manuscript. All authors approved the final version.

Competing Interests

The authors declare no competing interests.

References

1. World Health Organization. Medical device regulations: global overview and guiding principles. (WHO, Geneva, 2003).

2. European Parliament and Council. Regulation (EU) 2017/745 on medical devices. *Off. J. Eur. Union* **L117**, 1–175 (2017).

3. Palojoki, S., Saranto, K. & Lehtonen, L. Reporting medical device safety incidents to regulatory authorities: an analysis and classification of technology-induced errors. *Health Informatics J.* **25**, 731–740 (2019).

4. Ceross, A. & Bergmann, J. A machine learning approach for medical device classification. *Proc. 14th Int. Conf. Theory and Practice of Electronic Governance*, 332–338 (ACM, 2021).

5. Gerke, S., Minssen, T. & Cohen, G. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artif. Intell. Healthcare* 295–336 (Academic Press, 2020).

6. National Medical Products Administration. UDI Database. https://udi.nmpa.gov.cn/download.html (accessed 25 September 2024).

7. National Medical Products Administration. Rules for medical device unique identification system (NMPA Order No. 66). (NMPA, Beijing, 2019).

8. International Medical Device Regulators Forum. UDI guidance: unique device identification (UDI) of medical devices. IMDRF/UDI WG/N7FINAL:2013 (2013).

9. U.S. Food and Drug Administration. Unique device identification system. https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/unique-device-identification-system-udi-system (accessed 15 January 2025).

10. National Medical Products Administration. Rules for the classification of medical devices (NMPA Order No. 15). (NMPA, Beijing, 2015).

11. Han, Y., Ceross, A. & Bergmann, J.H.M. Toward automated regulatory decision-making: AI-assisted medical device risk classification with multimodal transformers and self-training. *Expert Syst. Appl.* **318**, 131669 (2026). https://doi.org/10.1016/j.eswa.2026.131669

12. Xie, Q., Luong, M.-T., Hovy, E. & Le, Q.V. Self-training with Noisy Student improves ImageNet classification. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 10687–10698 (2020).

13. Zoph, B. et al. Rethinking pre-training and self-training. *Adv. Neural Inf. Process. Syst.* **33**, 3833–3845 (2020).

14. Sun, J. Jieba: Chinese text segmentation. https://github.com/fxsjy/jieba (accessed 15 January 2025).

15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. NAACL-HLT*, 4171–4186 (2019).

16. Tan, M. & Le, Q.V. EfficientNet: rethinking model scaling for convolutional neural networks. *Proc. Int. Conf. Mach. Learn.*, 6105–6114 (2019).

17. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).

18. U.S. Food and Drug Administration. Classify your medical device. https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device (accessed 15 January 2025).

19. Conneau, A. et al. Unsupervised cross-lingual representation learning at scale. *Proc. ACL*, 8440–8451 (2020).

20. Chen, R.J. et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**, 850–862 (2024).

21. Schuhmann, C. et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* **35**, 25278–25294 (2022).

22. Han, Y., Ceross, A. & Bergmann, J.H.M. NMPA-MedDevice: A Dataset of Chinese Medical Devices with Risk Labels and an Image-Linked Subset. *Figshare* https://doi.org/10.6084/m9.figshare.31322272 (2026).