MIMIC-CXR-DB: An Admission-Linked, Ontology-Grounded Multimodal Resource Built on MIMIC-CXR and MIMIC-IV

Houcemeddine Turki[1], Lukman E. Ismaila[2,3], Ahmed Ben Salem[1], Ahmed Nebli[4], Mahdi Kchaou[1], Abdulhameed Abiola Dere[5], Anas Alzahrani[6], and Makram Koubaa[1]

[1] Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia
turkiabdelwaheb@hotmail.fr
[2] Johns Hopkins University School of Medicine, Baltimore, MD, USA
[3] Kennedy Krieger Institute, Baltimore, MD, USA
[4] Independent Researcher, Jülich, Germany
[5] University of Ilorin, Ilorin, Nigeria
[6] King Abdulaziz University, Jeddah, Saudi Arabia

Abstract. Groups working with MIMIC still have to rebuild the same link between chest radiographs and the patient context held in the electronic health record. We release MIMIC-CXR-DB, a relational database that joins MIMIC-CXR-JPG radiographs to MIMIC-IV at the level of the hospital admission. Under a common key it exposes a curated cardiopulmonary laboratory panel, ICD-10 primary and secondary diagnoses, and an ontology-grounded triple representation of both radiology and discharge notes against twelve biomedical vocabularies. Existing MIMIC derivatives annotate free text against a fixed list of target findings. Our grounding fixes no such list in advance, which yields 14,863 distinct discharge concepts but costs precision that consumers have to filter. Of 377,095 radiographs (227,827 studies), 146,333 (38.8%) fall within an inpatient admission window in MIMIC-IV v3.1. We keep the rest with a null `hadm_id` and compare the two subsets, since our observations suggest the linkage rule enriches for admitted, sicker patients. We describe how the database was built, what the physicians reviewed, and what an automated audit shows about the coverage and the concept specificity of the released triples. The database and the build code fall under the existing PhysioNet credentialed access of the parent projects.

Keywords: Multimodal Database · Chest Radiography · Electronic Health Records · Tabular and Imaging Fusion · Ontology Grounding · MIMIC-IV · MIMIC-CXR

1 Introduction

Chest radiography is one of the most frequently performed imaging examinations and a first-line tool for cardiopulmonary disease [10]. The community has invested heavily in large public collections such as ChestX-ray8 [17], CheXpert [8], and PadChest [3], and the MIMIC-CXR family [10,11] stands out because its studies were collected as part of routine care and come with paired free-text reports.

Putting a MIMIC-CXR study back into its clinical context is what still takes custom engineering. The labs drawn around the time of the film, the comorbidities the patient brought to the admission, and the eventual discharge summary all sit in MIMIC-IV [9], spread across modules designed for a hospital-wide record. Pulling out the cardiopulmonary slice means writing and validating a join across several of those modules, harmonizing units, and reconciling the result against MIMIC-CXR-JPG study identifiers. None of that is novel or publishable, yet each group redoes it with its own cohort filtering, lab harmonization, and report parsing, usually undocumented. Those hidden choices are what make results hard to compare across papers. Models trained without the surrounding context also tend to latch onto shortcut features [13].

We provide that join as a reusable resource for multimodal learning over chest radiographs and tabular clinical context. We are not proposing a new note-extraction method. What we offer is a curated database other groups can adopt without repeating the integration, built as a seven-table relational schema that ties radiographs, admissions, a cardiopulmonary laboratory panel, ICD-10 diagnoses, and ontology-grounded note triples together through the admission key. We characterize the linked and unlinked radiographs, keep every unlinked image in the table, describe what the clinical reviewers actually checked, and audit what the released triples contain, including how few of them carry a clinically specific concept.

2 Related Work

MIMIC-CXR [10] introduced the original chest radiograph collection paired with free-text reports, and MIMIC-CXR-JPG [11] added the CheXpert label layer. Chest ImaGenome contributes scene-graph annotations [18], and REFLACX contributes radiologist eye-tracking and speech for a subset [1]. Table 1 sets MIMIC-CXR-DB against its closest neighbors. The nearest precedent is Symile-MIMIC [14], the only prior release pairing MIMIC-CXR images with both ECG and laboratory data, though it covers a small subset and exposes neither ICD diagnoses nor anything derived from the notes. MIMIC-eye reuses the MIMIC-CXR study identifiers and a selection of MIMIC-IV elements, but targets the eye-tracking modality from REFLACX and not broad EHR coverage. MIMIC-IV Note supplies raw text keyed to `hadm_id` and leaves the NER, the grounding, and the join to the user. MIMIC-CXR-DB brings all four together in one release.
Open vocabulary against closed schema. One difference does not show up in the columns of Table 1. Existing derivatives structure MIMIC free text against a target list fixed in advance, either the fourteen CheXpert findings or the vocabulary of Chest ImaGenome, and ask of each item whether the text asserts it. Precision on those enumerated concepts is high, but the approach says nothing about the rest of the note, which is most of it. Medication histories,

Table 1. Comparison with the closest existing MIMIC derivatives on the dimensions relevant to an admission-level multimodal resource. "Subset" means only a subset of MIMIC-CXR studies is covered. "12 sources" means the grounding pipeline queries twelve distinct biomedical ontologies as candidate targets; nine contribute at least one link to the released triples (Section 4.3).

Resource	CXR	Admis.	Labs	ICD	Notes	Ontology
MIMIC-CXR-JPG [11]	full	–	–	–	raw text	–
MIMIC-IV Note [9]	–	yes	via IV	via IV	raw text	–
Chest ImaGenome [18]	full	–	–	–	scene graph	–
REFLACX/MIMIC-eye [1]	subset	partial	–	–	raw text	–
Symile-MIMIC [14]	subset	yes	few	–	–	–
MIMIC-CXR-DB	full	38.8% linked	36 curated	primary + secondary	triples	12 sources, 9 matched

comorbidities, procedures, and the reasoning recorded in a discharge summary all fall outside any fixed radiographic label set. Our pipeline fixes no list beforehand. It extracts whatever the text mentions and resolves each mention against twelve ontologies, which is how the tables come to span 14,863 discharge and 6,009 radiology concepts. Precision is the price, and Section 4.3 measures it instead of assuming it away. The trade seems right for a released resource, since extraction is expensive and happens once, while filtering to a task-specific subset is a single predicate and depends on a downstream question we cannot anticipate.

3 Construction

The pipeline routes three modalities through parallel per-modality stages (imaging and ICD-10 diagnoses; laboratory harmonization; NER and ontology grounding) and emits the seven relational tables of Fig. 1, all organized around the hospital admission key.

3.1 Sources, integration key, and imaging table

The EHR content comes from MIMIC-IV v3.1 [9] and the radiographs and reports from MIMIC-CXR-JPG v2.1.0 [11], all pulled through authenticated BigQuery queries. Everything hinges on `hadm_id`, the hospital admission identifier that MIMIC-IV assigns to a single inpatient stay, where one value means one row of `admissions` spanning one admit-to-discharge interval for one patient. We build the schema around it because it is the coarsest identifier at which radiographs, laboratory measurements, diagnoses, and discharge notes all co-occur. A radiograph links to an admission when its study timestamp falls inside that admission's window, and otherwise keeps a null `hadm_id`. Unlinked images belong mostly to outpatient and emergency-department-only encounters. The primary table joins the CheXpert-labeled radiographs with DICOM metadata, admissions, demographics, and ICD-10 diagnoses, which we consolidate per admission into one primary diagnosis (`seq_num`=1) and a pipe-separated list of secondary diagnoses.

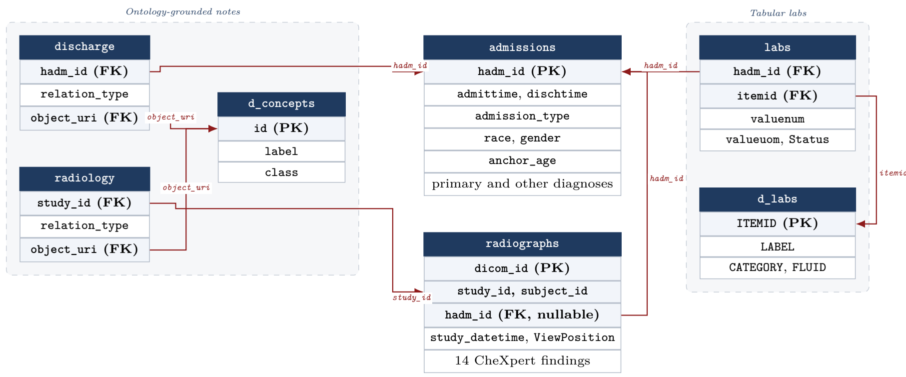

Fig. 1. Entity-relationship diagram of the seven released tables. Everything resolves to `hadm_id` at the centre, apart from radiology triples, which key on `study_id`. Bold fields are keys and arrows run from a foreign key to the primary key it references. Rows in `radiographs` are individual images, so `study_id` repeats across views, and `hadm_id` is null on the 61.2% of images with no overlapping inpatient window (Section 4.1).

3.2 Laboratory panel

The review team of Section 3.4 chose the panel of 36 laboratory `ITEMID`s by applying three criteria to the `d_labitems` dictionary. First, the analyte has to be used in routine cardiopulmonary assessment or in the differential diagnosis of the CheXpert finding set. Second, it has to be recorded for a reasonable share of inpatient admissions, so that the column is not almost entirely null. Third, where MIMIC-IV carries the same analyte under several `ITEMID`s from different laboratories, we keep each one separately and impose no unit or assay harmonization of our own. That third rule is why `d_labs` lists D-Dimer twice and carries blood-gas duplicates of hemoglobin, hematocrit, sodium, potassium, chloride, and glucose. The panel covers Chemistry (19), Hematology (9), and Blood Gas (8), taking in CBC components, core electrolytes and creatinine, cardiac markers (Troponin I/T, NTproBNP), inflammatory markers (CRP, Sedimentation Rate), D-Dimer, and arterial blood gases. The team worked by consensus and used no formal scoring instrument, so we make no claim that the panel is exhaustive. The full list with a per-analyte rationale sits in the code repository, and the extraction can be re-run with a different list.

Measurements come from `mimiciv_3_1_hosp.labevents`. For each admission and analyte the `labs` table keeps the earliest in-admission value (`valuenum`), its unit (`valueuom`), and a category (`Status`) in {Higher, Normal, Lower} computed against documented reference ranges, with `d_labs` resolving each `itemid` to a label, category and fluid. We keep the numeric value next to the category so that users can re-derive patient-adjusted bands, since the released category is population-level.

3.3 NER and ontology grounding

Free-text radiology and discharge notes are parsed by a section-aware hierarchical parser keyed on the conventional MIMIC headers (e.g., `IMPRESSION:`, `PHYSICAL EXAM:`, `Past Medical History`). Named entity recognition is run on each section with the `en_core_sci_lg` SciSpaCy model [12], and the NegEx pipeline removes negated mentions so the retained entities reflect confirmed findings.

Each entity is then grounded against twelve biomedical ontologies queried in a set order. SNOMED CT [2] comes first, resolved locally through ROBOT and `rdflib` [16], followed by OCHV, MeSH, PathLex, RadLex, SOHO, GALEN, LOINC, UPHENO, DCM, SYMP, and IOBC from BioPortal. Matching uses RapidFuzz similarity at a threshold of 90% or above against every lexical form of each concept. If nothing matches, we drop only the modifiers that POS-tag as adjectives or determiners and query again, so no arbitrary tokens are truncated. The pipeline emits triples (`hadm_id`, `relation_type`, `object_uri`) for discharge notes and (`study_id`, ...) for radiology, where `relation_type` records the source section, 35 labels for discharge and 9 for radiology. We treat it as a provenance label and not a semantic predicate. The dictionary `d_concepts` maps each `object_uri` to a label and a high-level class, and Table 3 shows what a row looks like.

3.4 Physician review of the schema

The clinical review carried out for this release covers the *schema* and not the individual entity-to-concept associations, and we scope the claim to match. A three-person team, made up of a board-certified radiologist, a family medicine physician, and a medical student with a computer science background, reviewed and signed off on the column names and data types of every released table, on the 36-analyte laboratory panel and its criteria (Section 3.2), and on how note section titles map onto `relation_type` values and concept classes in `d_concepts`. The section-title mapping took the most iteration, because MIMIC discharge summaries use inconsistent headers that have to be collapsed onto a stable vocabulary.

No one adjudicated individual associations, and we report no precision estimate for the triples. They are the raw output of the pipeline in Section 3.3, filtered only by NegEx negation removal. Section 4.3 describes what that output contains, and a physician-annotated correctness evaluation is the first item of planned work. The rubric, the reviewer instructions, and the section-title mapping are in the code repository.

4 The Released Database

MIMIC-CXR-DB is organized into the seven relational tables of Fig. 1. In the release, `admissions` holds 40,301 rows and `radiographs` holds 377,095 images (227,827 studies) over 65,379 patients, of which 146,333 images link to 40,172

Table 2. Linked (L, 146,333) vs. unlinked (U, 230,762) radiographs. Counts are images, not studies. ViewPosition as % within subset; CheXpert as % positive among images labeled for that finding. Demographics are not comparable: unlinked images have no admission record.

ViewPosition	L	U	CheXpert finding	L	U
PA	11.7%	34.2%	Cardiomegaly	67.5%	63.1%
AP	69.8%	19.5%	Pleural Effusion	66.5%	49.2%
LATERAL/other	18.5%	46.2%	Consolidation	56.4%	26.4%
			Pneumonia	31.4%	23.3%
			Pneumothorax	20.9%	15.1%

admissions covering 26,481 patients. The `labs` table holds 38,264 rows across 36 analytes, and the two triple tables hold 3,380,692 discharge and 1,578,323 radiology rows drawing on 53,564 concepts. Apart from the diagnosis consolidation described above, `radiographs` and `admissions` come from the parent projects unchanged, with the 14 CheXpert findings encoded as 1.0/0.0/−1.0. One join detail is worth knowing: `radiographs.hadm_id` carries a trailing decimal, so it needs a numeric cast to match `admissions`. Every foreign key resolves once it is applied, leaving one orphan admission and four orphan study identifiers. No imaging pixels are stored here, and the JPG and DICOM files stay on PhysioNet, reachable through the `dicom_id` path [11,5].

4.1 Linked vs. unlinked radiographs

Each row of `radiographs` is a single image keyed by `dicom_id`, and one study can contribute several views. Of 377,095 CheXpert-labeled images, 146,333 (38.8%) fall inside an inpatient admission window in MIMIC-IV v3.1, while the other 230,762 belong to outpatient and emergency-department-only encounters. We keep every row and set `hadm_id` to null on the unlinked ones, so a user can work on the linked subset or the full collection without re-running the join. We observe that AP projections dominate the linked subset at 69.8% against 19.5%, whereas the unlinked side carries more PA and lateral films, which may reflect a difference between bedside inpatient and ambulatory acquisition. Acute findings also appear more often on the linked side (Table 2). Taken together, these patterns suggest the linkage rule selects for sicker, admitted patients, and results obtained on the linked subset may therefore not carry over to the full collection. We cannot compare demographics across the two, since unlinked images carry no admission record.

4.2 Cohort and content overview

The linked cohort is mostly older adult inpatients. Patients aged 60 or above account for 59.9% of admissions and ages 40 to 59 for a further 29.6%, while the cohort is 68.7% self-reported White and 47.8% female. The most heavily populated discharge sections are *Past Medical History* (449,833 concepts), *Imaging and Diagnostics.General* (365,477), and *Physical Examination.Systems* (345,817), and on the radiology side *IMPRESSION* (644,452), *INDICATION* (417,785), and

Table 3. Released triple format for one radiology study (`study_id` 53889406; 11 triples total), joined to `d_concepts`. Rows 1–2 are clinically typed and survive a class filter; rows 3–5 are the laterality, generic, and stem-artifact matches that dominate by volume.

relation_type	object_uri	label	class
IMPRESSION	dicom:112147	Air space	Clinical_Activities_and_Procedures
IMPRESSION	loinc:LA14959-3	Improves	Clinical_Findings_and_Diseases
IMPRESSION	ochv:C0205090	right	Domain_Specific
IMPRESSION	ochv:1553	atelectasi	Domain_Specific
INDICATION	ochv:C0439234	year	Domain_Specific

FINDINGS (383,611). Within the laboratory panel, 61.3% of categorized values fall in *Normal*, 21.4% in *Lower*, and 17.3% in *Higher*.

Because the tables share an admission key, a finding can be traced to the diagnoses recorded for the same stay. Cardiomegaly is positive on 16,895 linked admissions, Pleural Effusion on 15,401, and Pneumonia on 6,839. We observe co-occurrences along expected clinical lines, with Support Devices alongside Cardiomegaly (59.8%), Atelectasis alongside Pleural Effusion (65.7%), and Pleural Effusion alongside Pneumonia (60.4%). Unspecified septicemia is the most frequent ICD-10 primary diagnosis for all three, and since the same code tops every row this may reflect the base-rate prevalence of septicemia coding in an acute inpatient cohort rather than any finding-specific association. These numbers are descriptive and unadjusted, and the full matrix is in the code repository.

4.3 Automated audit of the triple tables

With no human review of individual associations, this audit is the only evidence we can offer about what the triple tables hold, so we report the unflattering parts as well. Every `object_uri` in `discharge.csv` (3,380,692 triples) and `radiology.csv` (1,578,323) resolves to a `d_concepts` entry, making URI resolvability 100% by construction. The two tables draw on 14,863 and 6,009 distinct concepts, with exact duplicates at 4.9% and 1.0%. Nine of the twelve queried vocabularies contribute a match. OCHV supplies most links (88.0% discharge, 86.8% radiology), followed by SNOMED CT (8.1%, 6.6%) and LOINC (2.1%, 3.8%), with MeSH, IOBC, DCM, OBO, PathLex, and SYMP in the tail. OCHV likely comes out ahead because its lay-term coverage matches the surface forms SciSpaCy extracts, while the clinically specific concepts sit in SNOMED CT and LOINC.

Grouping the triples by `d_concepts` class suggests that specificity, and not resolvability, is the binding constraint here. Only 5.5% of discharge and 5.7% of radiology triples carry a clinically typed concept, meaning findings and diseases, anatomy, substances, procedures, or measurements. Most of the rest falls into the OCHV *Domain_Specific* class (88.1%, 87.2%) and *Qualifiers* (6.4%, 5.6%), and we observe correspondingly generic concepts at the top by volume: *admits*, *discharge*, *patient*, *no*, *left* for discharge notes, and *year*, *chest*, *right*, *atelectasi* for radiology. Inspection also turns up error modes we would expect of a threshold-based lexical matcher. Stemming artifacts get matched as concepts, as in *atelectasi* and *pulmonaries*, and abbreviations land on an unintended expansion, so `PLT` resolves to OCHV *primed lymphocyte test* instead of platelet count on 0.47% of discharge triples. Table 3 shows what filtering looks like, since selecting on concept class, vocabulary, or `relation_type` takes a single predicate.

5 Discussion

What this resource removes is a recurring engineering step, and it introduces no new method. Once the tables share an admission key, a model can draw on the radiograph together with the laboratory panel, demographics, ICD-10 codes, and the note triples. Fusing imaging with EHR variables reportedly improves mortality prediction and length-of-stay estimation over unimodal baselines [15,7], and the schema supplies the aligned inputs. The triples also offer concept-level supervision for report generation [4], phenotyping [6], and knowledge graphs [19].

Planned evaluation. Two evaluations sit outside this release. The first would measure the extraction against a physician-annotated reference, sampling associations across sections and vocabularies and having the review team adjudicate them under a five-label rubric of correct mapping, partially correct, wrong ontology, wrong concept, and spurious entity, with inter-rater agreement and a third reviewer for disagreements. That would supply the precision number this audit leaves open. The second would test whether the integrated context adds signal beyond the image, comparing multi-label CheXpert prediction with and without the harmonized labs and the discharge concepts on a held-out admission split.

Limitations. Only 38.8% of radiographs are admission-linked, and the unlinked ones cannot be reattached, because MIMIC-IV v3.1 holds no matching admission record. That makes the linked subset the analyzable one, and Section 4.1 suggests it is enriched for admitted, sicker patients. The `labs` table keeps a single earliest in-admission value per analyte, so a baseline on these columns sees the state a patient arrived in and not how it evolved. Tasks whose signal is a trend instead of a level, such as deterioration or treatment response, will therefore be understated. The panel came out of consensus and not a formal instrument (Section 3.2), and we release the extraction code so the list can be changed.

The triple tables are the unreviewed output of an automated pipeline. Nobody adjudicated individual associations and we report no precision estimate, so their error rate is unknown. Section 4.3 shows a composition weighted toward generic and qualifier concepts along with systematic lexical error modes. Users should treat the triples as a recall-oriented index over the notes and not as a ground-truth EHR, and should not read twelve-vocabulary grounding as twelve vocabularies' worth of depth, since most links rest on a consumer-health resource.

6 Conclusion

MIMIC-CXR-DB joins MIMIC-CXR-JPG radiographs to MIMIC-IV at the level of the hospital admission and adds a cardiopulmonary laboratory panel, structured ICD-10 diagnoses, and an ontology-grounded triple representation of both radiology and discharge notes. We characterize the linked and unlinked subsets and retain every unlinked image, document the construction pipeline and the clinical review applied to the schema, and report a structural audit of the released triples that quantifies their coverage alongside their limited concept specificity. The open-vocabulary design responsible for that low precision floor is also what lets the tables span concepts no fixed label set would have anticipated, and filtering the output remains the consumer's task, reduced by the schema to a single predicate.

Releasing the database and the build code means the admission-level join no longer has to be rebuilt for each new study. We hope this lowers the entry cost for work that needs a radiograph and its clinical context together, and that fixing the cohort and harmonization choices in one documented schema makes results easier to compare.

Acknowledgments. We thank Ahmed Dhia El Euch and Aiman Ghrab for their contributions to this work. This work is supported by the Saudi National Institute of Health (Saudi NIH) under Grant PRI01-2401-KAU19-46641119.

Data Availability Statement. MIMIC-CXR-DB inherits the PhysioNet Data Use Agreement of MIMIC-IV and MIMIC-CXR-JPG [5]. Upon approval by the PhysioNet editors, the database will be released on PhysioNet.

Disclosure of Interests. The authors have no competing interests.

References

1. Bigolin Lanfredi, R., Zhang, M., Auffermann, W.F., Chan, J., Duong, P.A.T., Srikumar, V., Tasdizen, T.: REFLACX: A dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. Scientific Data **9**(1), 350 (2022)
2. Bodenreider, O., Cornet, R., Vreeman, D.J.: Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm. Yearbook of Medical Informatics **27**(1), 129–139 (2018)
3. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: PadChest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis **66**, 101797 (2020)
4. Delbrouck, J.B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., Langlotz, C.: Improving the factual correctness of radiology report generation with semantic rewards. In: Findings of ACL: EMNLP. pp. 4348–4360 (2022)

5. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource. Circulation **101**(23), e215–e220 (2000)
6. Hripcsak, G., Albers, D.J.: Next-generation phenotyping of electronic health records. J. Am. Med. Inform. Assoc. **20**(1), 117–121 (2013)
7. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning. npj Digital Medicine **3**(1), 136 (2020)
8. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Ng, A.Y.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proc. AAAI Conf. Artif. Intell. vol. 33, pp. 590–597 (2019)
9. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data **10**(1), 1 (2023)
10. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
11. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Peng, Y., Horng, S.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
12. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: Fast and robust models for biomedical natural language processing. In: Proc. BioNLP Workshop. pp. 319–327 (2019)
13. Rafferty, A., Ramaesh, R., Rajan, A.: Limitations of public chest radiography datasets for artificial intelligence: Label quality, domain shift, bias and evaluation challenges. arXiv preprint arXiv:2509.15107 (2025)
14. Saporta, A., Puli, A.M., Goldstein, M., Ranganath, R.: Contrasting with Symile: Simple model-agnostic representation learning for unlimited modalities. In: Advances in Neural Information Processing Systems. vol. 37, pp. 56919–56957 (2024)
15. Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J.: Multimodal deep learning for biomedical data fusion: A review. Brief. Bioinform. **23**(2), bbab569 (2022)
16. Turki, H., et al.: Automating epidemiology report generation using SNOMED CT and SQL. Research Square (2025)
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2097–2106 (2017)
18. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Moradi, M.: Chest ImaGenome dataset for clinical reasoning. In: Proc. NeurIPS Datasets and Benchmarks Track. vol. 1 (2021)
19. Xiang, X., Wang, Z., Jia, Y., Fang, B.: Knowledge graph-based clinical decision support system reasoning: A survey. In: Proc. IEEE Int. Conf. Data Sci. Cyberspace. pp. 373–380 (2019)