

Data Readiness and Selection Bias in a Public Chest Radiograph Corpus for Generative AI Reporting: A Reproducible Audit

Mark Andre Balao Adallo
World Citi Colleges, Philippines

Corresponding Author

Abstract

Background: Public radiology report corpora are frequently reused in artificial intelligence research, but report-section completeness, redaction artifacts and near-duplicate text may affect the construction and interpretation of generative AI benchmarks. **Objective:** To quantify data-readiness limitations and selection effects in a public chest radiograph report collection using a reproducible, text-only audit. **Methods:** We audited 3,955 XML records from the Indiana University/National Library of Medicine chest X-ray collection. Eligibility required nonempty Findings and Impression sections, minimum lengths of 10 and two words, absence of XXXX redaction tokens, and no duplicated Findings after normalized exact matching and within-metadata-stratum character n-gram near-duplicate detection (cosine ≥ 0.99). We compared retained proportions for records coded normal and non-normal by MeSH metadata, with a retention risk ratio and 95% confidence interval. We also measured source-section lengths, lexical similarity, and simple negation and uncertainty cues; these are descriptive markers, not diagnostic errors. **Results:** Of 3,955 reports, 1,237 (31.3%) met strict criteria. Exclusions comprised 536 with incomplete sections, 21 below minimum length, 1,606 with redaction placeholders, 553 with repeated Findings and two near-duplicates. The retention proportion was 25.3% among normal-coded and 34.5% among non-normal-coded records (risk ratio 0.73; 95% CI 0.66–0.81). Among retained reports, median Findings and Impression lengths were 27 and five words, respectively; median paired lexical cosine was 0.027. **Conclusion:** Strict benchmark construction materially reduced usable reports and altered case-metadata composition. Transparent exclusion logs, duplicate audits, and explicit separation of lexical from clinically validated fidelity measures are needed before comparing generative AI reporting systems. No language model or clinical diagnostic performance was evaluated.

Keywords: Chest Radiography; Dataset Audit; Generative Artificial Intelligence; Radiology Reports; Reproducibility.

I. INTRODUCTION

Generative language models are being explored for radiology reporting, but experimental claims depend on the corpus from which prompts and reference texts are drawn. A benchmark may contain readable report narratives while concealing substantial variation in section coverage, de-identification placeholders, repeated sentences, and source-label composition. Reusing such data without a documented screening process risks confounding apparent model performance with the properties of the evaluation set. Data quality must therefore be analyzed before the outputs of an artificial intelligence model can be interpreted [1–4].

The publicly distributed Indiana University (IU) chest X-ray report collection and the National Library of Medicine Open-i architecture helped establish open biomedical image–text retrieval research [1,2,8]. Modern language-generation evaluations use paired reports and automated scores, yet the clinical meaning of text similarity remains limited without expert judgment. Yu et al. showed why radiology report evaluation benefits from clinically informed measures rather than lexical resemblance alone [5], and Smit et al. demonstrated the importance of expert-informed clinical label extraction from radiology text [6]. The FLAIR consensus guidelines, published in 2026, further emphasize reproducible data provenance, model and prompt details, risk statements and transparency in radiology LLM research [7]. Recent evaluations of AI-assisted radiograph reporting emphasize practical workflow and multicohort validation [9,10].

This study examines source-data readiness as a distinct empirical question. Rather than generating impressions and attributing unvalidated text changes to hallucination, we audit the original XML distribution and quantify the consequences of explicit screening criteria. Primary outcomes are retention and exclusion counts and differences in metadata-defined report composition. Secondary measurements summarize document lengths, normalized lexical similarity, repeated Findings with variant historical Impressions, and heuristic negation and uncertainty expressions. The aim is an auditable report-corpus profile that can support—but does not substitute for—future blinded AI-reporting evaluations.

II. MATERIALS AND METHODS

2.1 Design and source archive

We performed a retrospective computational audit of the originally uploaded IU/NLM chest X-ray XML report archive, NLMCXR_reports.tgz, provided to the investigators for secondary analysis. The archive contains 3,955 XML report files rather than the 3,996 reports described in the foundational publication [1]; all denominators in this paper refer to the uploaded file. Archive integrity was documented by SHA-256 (8fb6de7eec73d8c3665067ad4bb003ccd57f971ae316d2642e1627ac7268667a). The report XML includes labeled text sections, a report identifier and Medical Subject Headings (MeSH) assigned to records. No linked radiographic image pixels, independent clinical adjudications, patient identifiers, or machine-generated outputs were analyzed.

The uploaded XML files report a CC BY-NC-ND 4.0 URI. The study reports aggregate derivatives and computational procedures rather than redistributing the original narrative text. Investigators should confirm permissions for downstream processing and institutional requirements before external submission, code release or model-provider transmission.

2.2 Deterministic screening and duplication control

XML reports were parsed locally using Python's standard XML library. After retaining only reports containing both Findings and Impression, we excluded records whose Findings section contained fewer than 10 whitespace-delimited words or whose Impression contained fewer than two. Next, records containing the whole-word de-identification token XXXX in either section were excluded. Text was normalized by Unicode case folding, removal of non-word characters and collapse of whitespace. Repeated normalized Findings were removed deterministically by ascending report number, keeping the first occurrence. Within each MeSH metadata stratum, character 3–5-gram TF-IDF vectors and cosine similarity were used to identify additional near-identical Findings (threshold ≥ 0.99), retaining the lower-ID report. This threshold is an operational duplicate screen; it does not certify semantic equivalence or patient-level independence.

All exclusions were counted sequentially. The strict no-placeholder rule is intentionally conservative: a placeholder may sometimes affect only an incidental name or date, and excluding the entire report can unnecessarily remove useful text. We therefore report an alternate permissive numerator after the two section checks and before redaction and duplication screening to quantify the sensitivity of cohort size to these choices. No report was selected because of an AI model's output.

2.3 Source metadata strata and retention estimand

A record was categorized as "normal-coded" if the XML's MeSH major terms were exclusively normal, and as "non-normal-coded" otherwise. These categories reflect provided metadata, not an independent radiologist's verdict or ground-truth disease status. For each stratum, we calculated the proportion of original XML reports retained after strict screening. The prespecified composition contrast was the ratio of the normal-coded retention proportion to the non-normal-coded retention proportion; its 95% confidence interval used the log-risk-ratio normal approximation. A two-sided Pearson chi-square test assessed independence between the original stratum and strict eligibility. Because this is a complete audit of a finite release, inferential results should be interpreted as a descriptive measure of association under a hypothetical sampling model, not population generalization.

2.4 Descriptive text metrics

For retained reports we calculated Findings and Impression token counts, ratios of Impression to Findings length, and TF-IDF cosine similarity between each paired section. One common word-unigram-and-bigram vocabulary was fitted across the two sections, with sublinear term frequency. The cosine measure (0 to 1) is a surface lexical-overlap index; it does not measure semantic correctness, clinical equivalence, omission severity or hallucination. We also flagged the presence of simple rule-based negation cues (e.g., "no," "not," or "without") and uncertainty cues (e.g., "possible," "may," or "cannot exclude") in each section. These binary cues were descriptive only, without a clinical concept scope/negation parser and without expert adjudication. Repeated Findings groups were additionally identified in the section-complete, sufficiently long and no-placeholder pool prior to deduplication. We counted groups in which the historical Impression text varied. This identifies divergent phrasing paired with identical Findings, not necessarily contradictory clinical meaning. Exploratory comparison of lexical-cosine distributions between metadata strata used a two-sided Mann-Whitney test and a 2,000-resample within-stratum percentile bootstrap interval for the difference in medians. Those analyses were not adjusted for multiplicity and were not prespecified in the original AI-generation protocol.

2.5 Reproducibility and information governance

All analyses were completed offline using standard Python XML/tar/hash libraries, NumPy, SciPy and scikit-learn. The accompanying source code recreates exclusions, deduplication, comparison metrics, graphical outputs and anonymized text-free derived data from the official archive. The code sets seed 20261009 for bootstrap resampling. Original report narratives and unique source identifiers were not placed into the public-facing aggregate results package. This work involves secondary material

and no newly enrolled patients; any required local ethics determination remains the responsibility of the submitting investigator. Comparable attention to transparent corpus boundaries and reproducible selection is illustrated by Osma's scientometric analysis in a separate research field [11]; that study is cited as a methodological parallel, not evidence for radiology-specific screening criteria.

III. RESULTS AND DISCUSSION

3.1 Corpus completeness and sequential screening

Among 3,955 XML reports, 3,419 contained both sections. A further 21 records failed the minimum word-length screen, leaving 3,398. The strict redaction filter then excluded 1,606 records; 1,792 passed the initial text-eligibility steps. Removal of 553 exact repeated Findings and two additional near-duplicates produced 1,237 final eligible reports (31.3%; Wilson 95% interval 29.9%–32.7%). Independently of the screening sequence, 1,855/3,955 (46.9%) original reports contained the token XXXX in either section; this higher raw prevalence includes reports excluded at an earlier step. Figure 1 and Table 1 preserve the denominator and exclusion order for reproducibility.

Table 1. Sequential screening flow applied to the uploaded XML report archive.

Screening stage	Excluded at this stage	Remaining
Original XML files	—	3,955
Missing Findings or Impression	536	3,419
Minimum text length not met	21	3,398
XXXX placeholder in either section	1,606	1,792
Exact normalized Findings duplicate	553	1,239
Character n-gram near-duplicate (≥ 0.99)	2	1,237

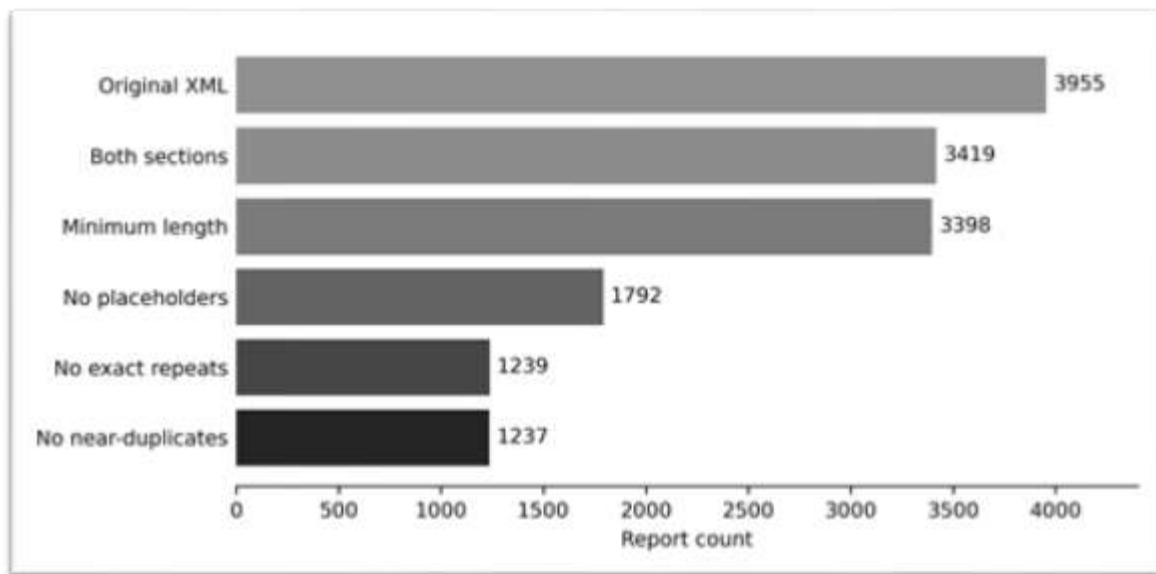


Figure 1. Number of reports retained after each deterministic screening stage. Counts represent the provided archive, not the larger source collection described in historical literature.

3.2 Metadata-defined selection imbalance

The original archive comprised 1,391 normal-coded and 2,564 non-normal-coded reports. Strictly eligible counts were 352 and 885, respectively. The corresponding retention proportions were 25.3% and 34.5%, giving a retention risk ratio of 0.73 (95% CI 0.66–0.81; Pearson $\chi^2 = 35.59$, $P < .001$). The resulting retained set was 28.5% normal-coded, compared with 35.2% in the original archive. This demonstrates measurable selection associated with the provided metadata, but no inference can be drawn about independently diagnosed disease prevalence or the direction of subsequent model-performance bias.

Table 2. Strict report eligibility by original MeSH-derived metadata category.

Metadata category	Original reports	Eligible reports	Retention (%)
Normal-coded	1,391	352	25.3
Non-normal-coded	2,564	885	34.5
Total	3,955	1,237	31.3

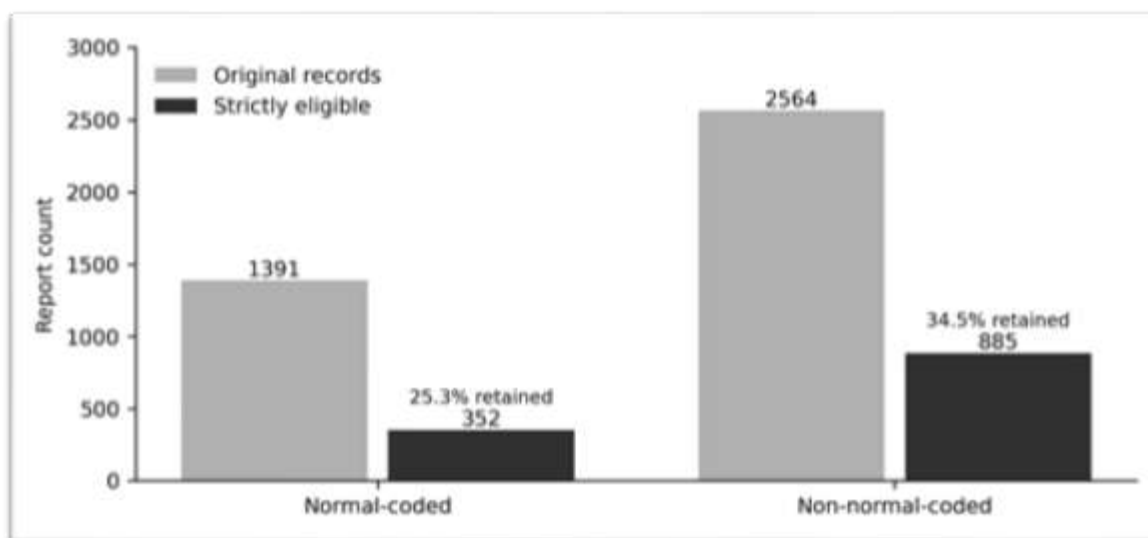


Figure 2. Original and retained source-report counts by the original MeSH metadata coding; percentages refer to within-stratum retention, not clinical classification accuracy.

3.3 Report-length and lexical features

Across the 1,237 retained reports, median Findings length was 27 words (IQR 21–34) and median Impression length was five words (IQR 4–7). The median Impression-to-Findings word-count ratio was 0.190. Normal-coded reports had shorter median Findings (22 versus 29 words). The pooled median word-unigram/bigram TF-IDF cosine similarity between paired sections was 0.0266 (IQR 0.0085–0.0951), consistent with substantially compressed, differently worded impression text. The median was 0.0201 in normal-coded and 0.0362 in non-normal-coded records (exploratory difference -0.0161 ; bootstrap 95% interval -0.0225 to -0.0103 ; exploratory Mann–Whitney $P = .001$). These small lexical-overlap values should not be interpreted as poor radiologist reporting or semantic disagreement.

Table 3. Section-length and paired lexical-similarity characteristics of strictly eligible reports.

Characteristic	All (n=1,237)	Normal-coded (n=352)	Non-normal-coded (n=885)
Findings, median words (IQR)	27 (21–34)	22 (17–28)	29 (23–36)
Impression, median words (IQR)	5 (4–7)	4 (4–5)	5 (4–8)
Impression/Findings median word ratio	0.190	0.192	0.188
TF-IDF cosine, median	0.0266	0.0201	0.0362

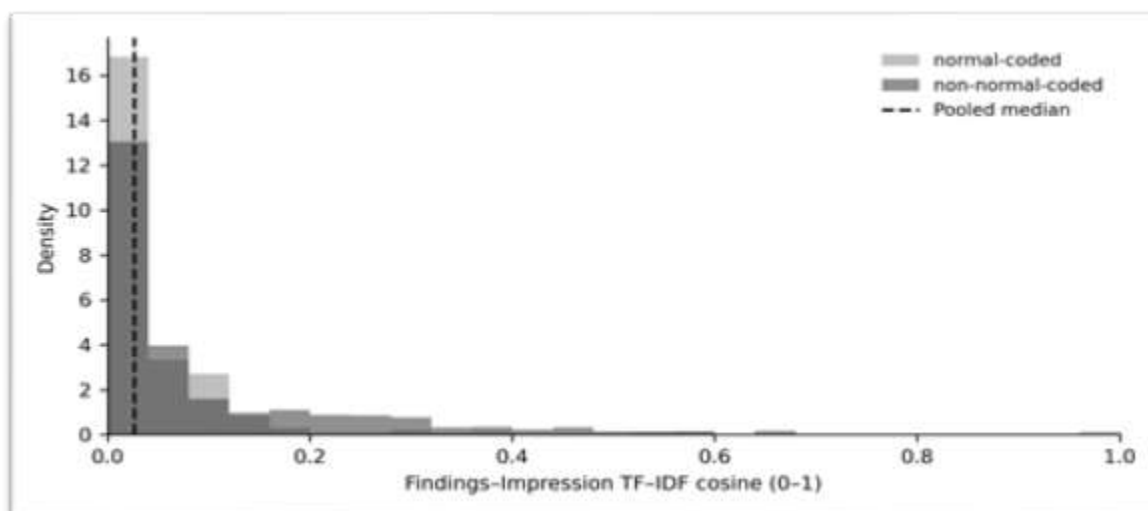


Figure 3. Distribution of paired section lexical similarity in the strictly eligible cohort. TF-IDF cosine indexes surface word overlap only and is not a clinically validated fidelity measure.

3.4 Duplicated source text and heuristic linguistic cues

Among the 1,792 pre-deduplication no-placeholder reports, 101 groups contained at least two reports with identical normalized Findings, comprising 654 records. In 34 such groups (337 records), the accompanying historical Impressions were not identically worded. These observations show that one Findings string may be paired with multiple distinct reference wordings, highlighting why random report-level partitioning without grouping may introduce leakage or unstable reference matching. The discordant wording should not be labeled diagnostic contradiction without clinical adjudication.

Simple negation markers appeared in Findings for 1,134/1,237 retained reports (91.7%) and in Impressions for 1,018 (82.3%). Markers occurred in both sections for 935 reports, only in Findings for 199, only in Impressions for 83, and in neither for 20. Uncertainty cues appeared in Findings for 60 reports (4.9%) and in Impressions for 53 (4.3%); 20 reports showed such cues in both sections. This variation is descriptive and may reflect appropriate condensation of negative statements; differences in word-level cue presence are not validated omission or hallucination rates.

Table 4. Source-section linguistic screening flags in the 1,237 strictly eligible reports.

Indicator	Count	Percentage of 1,237
Negation cue in Findings	1,134	91.7%
Negation cue in Impression	1,018	82.3%
Negation cue in both sections	935	75.6%
Negation cue only in Findings	199	16.1%
Negation cue only in Impression	83	6.7%
Uncertainty cue in Findings	60	4.9%
Uncertainty cue in Impression	53	4.3%

3.5 Interpretation of the empirical contribution

The principal result of this audit is that strict preparation for an AI report-generation evaluation removed more than two-thirds of records in the supplied IU chest radiograph report archive, with nonrandom retention across the source's MeSH-coded categories. The combined causes—section incompleteness, placeholder redaction, repeated Findings and near-duplicate report language—affect the material that model developers may treat as clean paired examples. Because selection differed by metadata category, comparative reporting studies that silently clean their corpora may not evaluate the same case mix even when they cite an identical public collection. This is a dataset comparability result rather than evidence that any particular language model is safer or less safe.

A second observation is that repeated Findings strings may be associated with different historical Impression text. Identical inputs with differing free-text references create ambiguity for naive sentence-overlap metrics and permit nearly identical test and training instances when reports are randomized at record level. Kapoor and Narayanan discuss broader pathways by which leakage can weaken the reproducibility of machine-learning research [4]. For a future generative radiology benchmark, grouping on Findings duplicates before partitioning and preserving the original report identifiers are more defensible than randomizing all texts independently. The current exercise implemented deterministic de-duplication but cannot detect patient-level overlaps because patient identifiers were not available.

The observed low lexical similarity between Findings and Impression should likewise not be equated with clinical disagreement. Impression sections intentionally summarize higher-level conclusions, while Findings sections often record numerous negated normal features. An Impression can remain clinically faithful while repeating few individual tokens. The distinction aligns with prior demonstrations that text similarity does not fully capture report correctness [5,6]. Negation and uncertainty cue differences in this study are examples of why unvalidated regex markers must be treated as flags requiring qualified expert review rather than errors in the original source.

3.6 Practical implications for future generative AI evaluation

The 1,237 retained reports constitute a feasible audited pool for subsequent text-to-text experimentation. A separately locked subset of 320 primary and 20 pilot cases had been specified in the earlier study protocol, with paired prompts designed to test source-grounded versus ordinary instruction. Those prompts are not assessed here. A full hallucination-safety evaluation requires outputs from a version-locked model, independent and blinded radiologist adjudication, prespecified severity anchors and paired uncertainty analysis; its results cannot be inferred from the historical sections or the present audit. FLAIR provides a relevant reporting framework when such model experiments are performed [7].

Transparency should also include a sensitivity analysis of redaction handling. Excluding every report with any placeholder greatly simplifies the input but may unnecessarily remove relevant text where the placeholder affects only metadata or a non-diagnostic time reference. Future comparisons should preregister a controlled placeholder-replacement policy, document the legal basis for input processing, and compare resulting case composition. The strong filtering in this paper is one reproducible operational choice, not an optimal universal policy.

3.7 Limitations

The audit assesses one historical, de-identified US radiograph corpus and is not a probability sample of modern diagnostic practice. The observed archive has fewer XML reports than the original publication describes, and differences between distributions could arise from acquisition or release history not available here. The text-length cutoffs, strict redaction exclusion and near-duplicate threshold were design choices, not clinically validated quality criteria. Although the deterministic procedure improves reproducibility, it can exclude useful information and changes report composition. Metadata normal/non-normal categories are not verified diagnoses. Automated TF-IDF and regex features capture surface wording rather than anatomical meaning, negation scope or error severity. Exact or near-exact shared Findings need not imply the same patient; conversely, paraphrased records may arise from the same patient. No images were inspected, no clinician reviewed case-level fidelity, and no generative model was executed. Consequently this paper makes no claim about AI hallucination prevalence, reporting accuracy, radiation dose, patient outcomes, or clinical deployment.

IV. CONCLUSION

In a reproducible audit of 3,955 IU/NLM chest radiograph XML reports, only 1,237 met a strict pairwise-text readiness specification. Placeholder artifacts and repeated Findings were substantial, and cleaning retained a different proportion of normal-coded and non-normal-coded records. Textual similarity and negation markers also varied across report sections but are not clinical error measures. Before investigating the quality of AI-generated radiology impressions, researchers should publish data-source hashes, complete screening flows, record-level duplication controls, selection effects and a separation between automatic lexical proxies and expert-validated diagnostic fidelity.

ACKNOWLEDGMENT

The author acknowledges the U.S. National Library of Medicine and the contributors to the Indiana University Chest X-ray Collection for making de-identified radiology reports available for secondary research. Their contribution to public biomedical research resources is recognized without implying their endorsement of the present analysis.

Conflict of Interest

The author declares that there are no conflicts of interest related to this study.

Funding

This research received no external or institutional funding.

Ethics Statement

Ethics approval and informed consent were not applicable because this computational study used only publicly available, de-identified secondary radiology reports, with no participant recruitment or access to identifiable records. Institution-specific ethics requirements should be confirmed before submission.

Data Availability

The study used the Indiana University/NLM Open-i chest X-ray report collection (XML files identify a CC BY-NC-ND 4.0 license). No raw reports are redistributed; offline analysis code and text-free aggregate results are available in the reproducibility package.

AI-Assistance Disclosure

Paperpal was used to check grammar, spelling, and language. Generative AI also assisted early drafting, code scaffolding, and document formatting; it did not generate the reported statistical findings. The author must verify all analyses, references, and disclosures before submission.

REFERENCES

- [1]. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc.* 2016;23(2):304–310. doi:10.1093/jamia/ocv080.
- [2]. Demner-Fushman D, Antani S, Simpson M, Thoma GR. Design and development of a multimodal biomedical information retrieval system. *J Comput Sci Eng.* 2012;6(2):168–177. doi:10.5626/JCSE.2012.6.2.168.
- [3]. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc.* 2013;20(1):144–151. doi:10.1136/amiajnl-2011-000681.
- [4]. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns.* 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.

- [5]. Yu F, Endo M, Krishnan R, et al. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*. 2023;4(9):100802. doi:10.1016/j.patter.2023.100802.
- [6]. [6] Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *Proceedings of EMNLP*. 2020:1500–1519. doi:10.18653/v1/2020.emnlp-main.117.
- [7]. Kottlors J, Iuga AI, Bluethgen C, et al. Guidelines for reporting studies on large language models in radiology: an international Delphi expert survey. *Radiology*. 2026;318(2):e250913. doi:10.1148/radiol.250913.
- [8]. Demner-Fushman D, Antani SK, Simpson M, Rahman MM. Combining Text and Visual Features for Biomedical Information Retrieval. Report to the Board of Scientific Counselors. US National Library of Medicine; 2010.
- [9]. Huang J, Wittbrodt MT, Teague CN, et al. Efficiency and quality of generative AI–assisted radiograph reporting. *JAMA Netw Open*. 2025;8(6):e2513921. doi:10.1001/jamanetworkopen.2025.13921.
- [10]. Hwang EJ, Lee JH, Lim WH, et al. Clinical validation of a generative artificial intelligence model for chest radiograph reporting: a multicohort study. *Radiology*. 2025;316(3):e250568. doi:10.1148/radiol.250568.
- [11]. Osma AN. From e-commerce to governance: a triple helix scientometric mapping of ASEAN’s digital trade research evolution (2015–2023). *J Scientometr Res*. 2026;15(1):133–150. doi:10.5530/jscires.20262074.