ReXGradient-160K: A Large-Scale Publicly Available Dataset of Chest Radiographs with Free-text Reports

Xiaoman Zhang XIAOMAN_ZHANG@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Julián N. Acosta JULIAN_ACOSTA@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Josh Miller JOSH@GRADIENTHEALTH.IO
Gradient Health, Durham, NC, USA

Ouwen Huang OUWEN@GRADIENTHEALTH.IO
Gradient Health, Durham, NC, USA
Department of Statistical Science, Duke University, Durham, NC, USA
Laplace Institute, Durham, NC, USA

Pranav Rajpurkar PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Abstract

We present ReXGradient-160K, representing the largest publicly available chest X-ray dataset to date in terms of the number of patients. This dataset contains 160,000 chest X-ray studies with paired radiological reports from 109,487 unique patients across 3 U.S. health systems (79 medical sites). This comprehensive dataset includes multiple images per study and detailed radiology reports, making it particularly valuable for the development and evaluation of AI systems for medical imaging and automated report generation models. The dataset is divided into training (140,000 studies), validation (10,000 studies), and public test (10,000 studies) sets, with an additional private test set (10,000 studies) reserved for model evaluation on the ReXrank benchmark. By providing this dataset, we aim to accelerate research in medical imaging AI and advance the state-of-the-art in automated radiological analysis.

Data and Code Availability The dataset is available at https://huggingface.co/datasets/rajpurkarlab/ReXGradient-160K.

Institutional Review Board (IRB) This study was approved by the Harvard University IRB (22-0364) and IRB (22-0400) with waiver of consent for retrospective research on de-identified data.

1. Introduction

The increasing global demand for radiological expertise, coupled with uneven distribution of specialists and growing workloads, has created significant challenges in healthcare delivery. Recent years have seen remarkable advances in artificial intelligence (AI) applications for medical imaging, particularly in developing AI systems that can generate comprehensive radiology reports (Tanida et al., 2023; Zhou et al., 2024; Bannur et al., 2024; Pellegrini et al., 2023; Luo et al., 2023; Tanida et al., 2023; Pellegrini et al., 2023; Wu et al., 2023; Zhang et al., 2024a), aiming to enhance workflow efficiency and expand access to expert-level interpretations.

Several large-scale chest X-ray datasets with paired reports have been instrumental in advancing this field. The MIMIC-CXR dataset, comprising 227,835 studies from 65,379 patients treated at Beth Israel Deaconess Medical Center, provided one of the first large-scale collections of radiographs with corresponding free-text reports (Johnson et al., 2019). This was followed by CheXpert Plus, which enhanced the original CheXpert dataset with detailed radiology reports and metadata across 223,228 studies from 64,725 patients, offering improved capabilities for model development and evaluation (Chambon et al., 2024). The IU X-ray dataset, a smaller database with 7,470 image-report pairs, has been widely used for

early development and testing of report generation models (Demner-Fushman et al., 2016).

However, as AI-assisted medical reporting continues to evolve rapidly, there is a critical need for standardized benchmarks and comprehensive evaluation frameworks. Existing datasets often face limitations regarding consistent data splits, standardized evaluation metrics, and the ability to test model generalization across different clinical settings (Johnson et al., 2019; Chambon et al., 2024). Most notably, these datasets are typically collected from single institutions, which limits their ability to evaluate models' generalization capabilities across diverse healthcare settings. The Medical AI Data for All (MAIDA) initiative (Saenz et al., 2024) represents another important effort to address the generalizability challenge by establishing a framework for global medical-imaging data sharing across diverse clinical environments. However, its current data collection scale of approximately 100 scans per setting limits its ultimate size.

To address these challenges, we present ReXGradient-160K, a large-scale, multi-institutional dataset of chest radiographs with paired radiology reports (Figure 1). Our dataset contains 160,000 chest X-ray studies with associated reports from over 100,000 unique patients across multiple medical institutions, making it the largest publicly available multi-site chest X-ray dataset to date. This dataset is complemented by ReXGradient, a private evaluation set of 10,000 studies from 67 U.S. medical sites, which serves as the test set for the ReXrank benchmark (https://rexrank.ai) (Zhang et al., 2024b). Together, this comprehensive collection includes multiple images per study and detailed radiology reports, making it particularly valuable for developing and evaluating AI systems for medical imaging and automated report generation. The multi-institutional nature of our dataset provides a unique opportunity to assess and improve the generalization capabilities of AI models across different clinical settings and geographical locations.

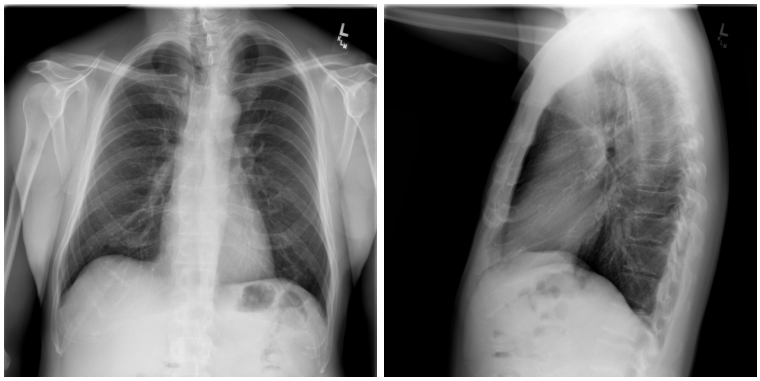
2. Related Works

Several chest X-ray datasets with paired reports have been developed to advance AI research in medical imaging. The MIMIC-CXR dataset (Johnson et al., 2019) contains 377,095 chest X-rays associated with 227,835 studies from 65,379 patients treated at Beth Israel Deaconess Medical Center (Boston, MA). This single-institution data set includes free text radiology reports and DICOM images and was one of the first large-scale collections to provide both image data and comprehensive reports. CheXpert Plus (Chambon et al., 2024) enhanced the original CheXpert dataset by adding radiological reports to 223,228 studies from 64,725 patients at Stanford Hospital. The dataset provides 14 labeled findings and preserved DICOM metadata, supporting research in both classification and report generation tasks. PadChest (Bustos et al., 2020), collected from the Hospital Universitario de San Juan in Spain, includes 160,861 images from 67,625 patients with reports written in Spanish, which limits its utility. Some smaller datasets like IU X-ray (Demner-Fushman et al., 2016) have contributed to advancing AI research in chest X-ray analysis, but limited in size and institutional diversity. Concurrent with dataset development, researchers have proposed domain-specific evaluation metrics, including RadCliQ (Jain et al., 2021), RaTEScore (Zhao et al., 2024), GREEN (Ostmeier et al., 2024), FineRadScore (Huang et al., 2024), and CheXPrompt (Zambrano Chaves et al., 2025) to better assess clinical accuracy beyond traditional NLP metrics.

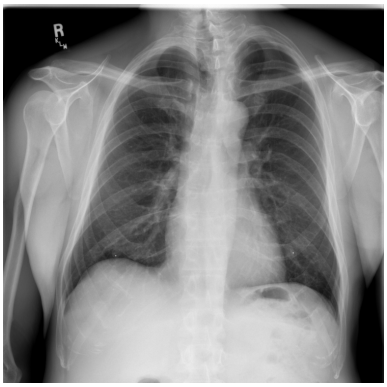
Our ReXGradient-160K dataset addresses key limitations in existing datasets. With 273,004 images from 160,000 studies and 109,487 patients, it represents the largest publicly available chest X-ray dataset in terms of the number of patients. Most significantly, ReXGradient-160K is collected from 4 distinct medical systems across the United States, providing unprecedented institutional diversity compared to existing single-institution datasets. This diversity is crucial for developing AI systems that can generalize across different clinical settings, radiograph acquisition protocols, and reporting styles.

3. Dataset Composition

ReXGradient-160K comprises 273,004 unique chest X-ray images from 160,000 radiological studies, collected from 109,487 unique patients across 3 U.S. health systems. The dataset is divided into three splits: training (140,000 studies), validation (10,000 studies), and public test (10,000 studies) sets. A single patient may be associated with multiple studies over time, and each study typically contains one or more images. Table 1 provides a comprehensive overview of the dataset statistics across different splits, including the number of studies, unique im-



Patient Demographic: 62Y, M
Study Date: 20130105
Indication: Preoperative respiratory examination for hernia repair. Ex-smoker.
Comparison: None.
Findings : The heart size and mediastinal contours are normal. There is a nodular density projecting inferior to the anterior aspect of the right fifth rib on the frontal examination. There is no definite corresponding finding on the lateral view. This may reflect a nipple shadow. A small pulmonary nodule is difficult to exclude. The lungs are otherwise clear. There is no pleural effusion. Osseous structures appear normal.
Impression: 1. No acute cardiopulmonary process. 2. Possible asymmetric nipple shadow on the right versus small pulmonary nodule. Repeat PA view with nipple markers recommended. These results will be called to the ordering clinician or representative by the Radiologist Assistant, and communication documented in the PACS Dashboard.



Patient Demographic: 62Y, M
Study Date: 20130105
Indication: Evaluate for pulmonary nodule versus asymmetric nipple shadow.
Comparison: Earlier the same date.
Findings : Repeat PA view with nipple markers demonstrates the nodular density overlapping the right lung base to correspond with the patient's right nipple shadow. There is no evidence of pulmonary nodule.
Impression: Repeat exam confirms the presence of a prominent right nipple shadow. No evidence of pulmonary nodule.

Figure 1: Example of multiple studies from a single patient in ReXGradient-160K. Left and Middle: Initial study with frontal and lateral views. Right: Follow-up study from the same patient demonstrating the dataset's longitudinal capability with preserved temporal relationships.

ages, patients, and the mean token counts for each report section.

3.1. Data De-identification

To ensure complete patient histories, we excluded patients whose first chronological study contained non-null comparison fields, as this indicated missing prior examinations in our collection. All radiological reports were processed through a standardized pipeline that verified the presence of four essential sections: indication, comparison, findings, and impression. Reports lacking clear section delineation were excluded. This comprehensive filtering and validation process, while reducing the initial cohort size, guaranteed that our final dataset of 160,000 studies represents complete, well-structured data suitable for robust deep earning research across diverse clinical environments.

3.2. Data De-identification

Our de-identification process is fully HIPAA compliant and is divided into two parts: text de-identification for reports and imagery metadata, and pixel de-identification for removing protected health information (PHI) specifically found in the imagery. In the cases of names, IDs, and dates, we resort to pseudonymization to add an additional layer of protection. Names are replaced, IDs are altered, and dates are shifted to within 365 days of the actual date. Studies for the same patient undergo the same date shift in order to maintain validity in comparing patient outcomes over time

3.3. Image Characteristics

All images in the dataset were originally in DICOM format and have been converted to PNG format using a standardized preprocessing pipeline. The process began with pixel data extraction from DICOM files using pydicom, followed by proper handling of bit depth and photometric interpretation. For MONOCHROME1 images, pixel values were inverted to maintain consistent representation across the dataset. Standard min-max normalization to the full 16-bit range (0-65535) was applied. In accordance with our data use agreement, images were downsam-

Table 1: Dataset Statistics. Study Statistics show the distribution of studies, unique images, and distinct patients. Report Statistics display the mean number of tokens for each report section (Indication, Comparison, Findings, and Impression).

Split	Study Statistics			Report Statistics (Mean Tokens)			
	Studies	Images	Patients	Indication	Comparison	Findings	Impression
Train	140,000	238,968	95,716	5.11	2.65	32.27	11.17
Validation	10,000	17,007	6,964	5.16	2.75	32.92	11.63
Test	10,000	17,029	6,807	5.13	2.78	32.48	11.52

Table 2: Demographic Statistics. Age Statistics show the percentage distribution across age ranges. Sex Statistics display the percentage distribution by gender.

Split	Age Statistics (%)					Sex Statistics (%)		
	0-20	20-40	40-60	60-80	80+	Male	Female	Unknown
Train	16.9	19.4	25.1	26.0	12.6	49.0	49.5	1.5
Validation	17.9	20.3	25.5	24.0	12.2	48.3	50.4	1.3
Test	17.3	19.6	25.4	24.5	13.2	47.9	50.3	1.8

pled to 25% of their original dimensions using cubic interpolation with anti-aliasing to maintain important structural details. Original images are commercially available through Gradient Health (Durham, NC, USA)

3.4. Report Structure

Each radiological report in our dataset is structured into four key sections:

- **Indication**: Provides relevant patient background and reason for examination
- **Comparison**: Provides relevant patient background and reason for examination
- **Findings**: Detailed radiological observations
- **Impression**: Summary of key conclusions and recommendations

To ensure consistent report formatting across all studies, we prompt GPT-4o to extract four key sections from the original reports. Additionally, we implemented a robust post-processing validation step to ensure all sections contained valid content, remove the cases where particular sections might be missing in the original reports.

3.5. Demographic Distribution

Table 2 presents the demographic distributions across training, validation, and testing splits. The age distribution shows consistency across all splits, with patients aged 40-80 years representing the largest proportion (approximately 50% of the dataset). The sex distribution exhibits a balanced ratio between male and female patients (approximately 49% male and 50% female) across all splits, with a minimal percentage (1.3-1.8%) of cases with unknown.

4. Data Records

All data will be available on Huggingface. Access to the ReXGradient-160K dataset requires signing our license on Huggingface. The dataset organizes images into subfolders, with each subfolder named according to the anonymous patient identifier and study ID. Each patient subfolder contains one or multiple studies, with each study containing one or multiple images. For the reports, we provide metadata including PatientID, AccessionNumber, PatientSex, EthnicGroup, PatientAge, PatientWeight, StudyDate, InstitutionName, and Manufacturer.

5. Discussion and Conclusion

The development of robust AI systems for radiology requires datasets that can evaluate model generalization across diverse clinical environments. Recent benchmarking work on chest X-ray report generation has revealed significant performance variations when models are evaluated across different institutional datasets. For example, comprehensive evaluations using the ReXrank framework show that state-of-the-art models (Zhou et al., 2024; Bannur et al., 2024; Chen et al., 2024) exhibit substantial performance drops when tested across different institutions. This finding underscores a fundamental limitation in current single-institution datasets: they inadequately assess whether AI systems can maintain performance across the varied clinical settings, imaging protocols, and patient populations in real-world deployment.

Our ReXGradient-160K dataset directly addresses this critical gap by providing the largest multi-institutional collection of chest radiographs available for public research. The dataset covers 79 medical sites across 3 major health systems, enabling researchers to develop and validate AI models that can generalize beyond single-institution constraints. It enables the research community to build more robust AI systems capable of maintaining consistent performance across diverse clinical environments, thereby advancing the goal of reliable AI-assisted radiology in real-world healthcare settings.

Acknowledgments

This work was supported by Biswas Family Foundation's Transformative Computational Biology Grant in Collaboration with the Milken Institute.

References

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*, 2024.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. *arXiv preprint arXiv:2405.20613*, 2024.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis

Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*, 2024.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023.

Agustina Saenz, Emma Chen, Henrik Marklund, and Pranav Rajpurkar. The maida initiative: establishing a framework for global medical-imaging data sharing. *The Lancet Digital Health*, 6(1):e6–e8, 2024.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442. Association for Computational Linguistics, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16 (1):3108, 2025.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond SL Ho. Libra: Leveraging temporal images for biomedical radiology analysis. *arXiv preprint arXiv:2411.19378*, 2024a.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexrank: A public leaderboard for ai-powered radiology report generation. *arXiv preprint arXiv:2411.15122*, 2024b.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. *medRxiv*, pages 2024–06, 2024.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.