

ReXGradient-160K: 一个大型公开可用的胸部 X 光片数据集，带有自由文本报告

Xiaoman Zhang¹, Julián N. Acosta¹, Josh Miller², Ouwen Huang^{2,3,4} and Pranav Rajpurkar ¹

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, ²Gradient Health, Durham, NC, USA,

³Department of Statistical Science, Duke University, Durham, NC, USA, ⁴Laplace Institute, Durham, NC, USA

我们介绍了 ReXGradient-160K，这是目前公开的胸部 X 光数据集中患者数量最大的一个。该数据集包含来自美国三个医疗系统（79 个医疗机构）的 109,487 名独特患者的 160,000 份胸部 X 光检查及其配对的放射学报告。这个全面的数据集包括每项研究中的多张图像和详细的放射学报告，使其特别适合于医学影像 AI 系统的开发与评估以及自动化报告生成模型的研发。数据集被划分为训练（140,000 项研究）、验证（10,000 项研究）和公共测试（10,000 项研究）集，并保留了一个额外的私人测试集（10,000 项研究），用于在重排序级别基准上的模型评估。通过提供这个庞大的数据集，我们旨在加速医学影像 AI 的研究，并推进自动化放射学分析的前沿技术。我们的数据集将在 <https://huggingface.co/datasets/rajpurkarlab/ReXGradient-160K> 开源。

1. 介绍

全球对放射学专业知识的 demand 不断增加，加上专家分布不均和工作量的增加，给医疗服务带来了重大挑战。近年来，在医学影像领域的人工智能（AI）应用取得了显著进展，特别是在开发能够生成全面放射学报告的 AI 系统方面，旨在提高工作效率并扩大获得专家级解读的机会 (Bannur et al., 2024; Pellegrini et al., 2023; Tanida et al., 2023; Zhou et al., 2024)。

几个包含配对报告的大规模胸部 X 光数据集在推进该领域方面发挥了重要作用。MIMIC-CXR 数据集由贝斯以色列女执事医疗中心的 65,379 名患者的 227,835 项研究组成，提供了第一批大规模带有相应自由文本报告的放射图像 (Johnson et al., 2019)。随后是 CheXpert Plus，它通过在 64,725 名患者的 223,228 项研究中加入详细的放射学报告和元数据来增强原始的 CheXpert 数据集，为模型开发和评估提供了改进的能力 (Chambon et al., 2024)。IU X 光数据集是一个较小的数据库，包含 7,470 个图像-报告对，广泛用于早期报告生成模型的发展和测试 (Demner-Fushman et al., 2016)。

然而，随着 AI 辅助医疗报告的快速发展，迫切需要标准化基准和全面的评估框架。现有的数据集通常在一致的数据划分、标准化评估指标以及测试模型在不同临床环境中的泛化能力方面存在局限性 (Chambon et al., 2024; Johnson et al., 2019)。特别值得注意的是，这些数据集通常是单一机构收集的，这限制了它们评估模型在多种医疗环境中泛化能力的能力。Medical AI Data for All (MAIDA) 倡议 (Saenz et al., 2024) 代表了另一个重要的努力，旨在通过建立全球医学影像数据共享框架来解决泛化挑战。然而，其目前每个设置大约收集 100 次扫描的数据规模限制了最终的规模。

为应对这些挑战，我们介绍了 ReXGradient-160K，一个大规模、多机构的胸部 X 光片数据集，并附有对应的放射学报告。我们的数据集包含来自多个医疗机构的超过 10 万名独特患者的 16 万份胸部 X 光研究及其相关报告，使其成为迄今为止最大的公开可用的多站点胸部 X 光数据集。该数据集由 ReXGradient 补充，这是一个来自 67 个美国医疗站点的 1 万项研究的私有评估集，作为 ReXrank 基

Table 1 | 数据集统计。研究统计展示了研究、独特图像和不同患者的比例分布。报告统计显示了每个报告部分（指征、比较、发现和印象）的平均标记数量。

分割	研究统计			报告统计（平均标记数）			
	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

准测试 (<https://rexrank.ai>) (Zhang et al., 2024) 的测试集。综合来看，这一全面的数据集合包括每项研究的多幅图像和详细的放射学报告，特别有利于开发和评估用于医学影像和自动报告生成的人工智能系统。我们数据集的多机构性质提供了独特的机会，可以评估并改进 AI 模型在不同临床环境和地区的一般化能力。

2. 数据集组成

ReXGradient-160K 包含来自 160,000 项放射学研究的 273,004 张独特的胸部 X 光图像，这些研究来自于跨越美国三个医疗系统的 109,487 名独特患者。数据集被分为三部分：训练（140,000 项研究）、验证（10,000 项研究）和公共测试（10,000 项研究）集合。单个患者可能会随着时间与多项研究相关联，而每个研究通常包含一张或多张图像。表 1 提供了跨不同划分的数据集统计的综合概述，包括研究数量、独特图像数、患者数量以及每份报告部分的平均标记计数。

2.1. 数据去标识化

我们的去识别化流程完全符合 HIPAA 规定，并分为两个部分：报告和影像元数据的文本去识别，以及通过像素去识别来移除影像中发现的受保护健康信息 (PHI)。对于姓名、ID 和日期，我们采用假名化以增加一层保护。姓名被替换，ID 被修改，日期则调整为实际日期前后 365 天内的某个时间点。同一患者的各项研究会经历相同的时间偏移，以便在比较患者随时间的健康结果时保持有效性。

2.2. 图像特征

数据集中的所有图像最初都是 DICOM 格式，并已使用标准化的预处理管道转换为 PNG 格式。该过程始于使用 pydicom 从 DICOM 文件中提取像素数据，随后正确处理了位深度和光度解释。对于 MONOCHROME1 图像，反转了像素值以在整个数据集中保持一致表示。应用了标准最小-最大归一化到完整的 16 位范围 (0-65535)。根据我们的数据使用协议，图像使用具有抗锯齿功能的三次插值缩小到原始尺寸的 25%，以保持重要的结构细节。

2.3. 报告结构

我们数据集中的每个放射学报告被分为四个关键部分：

- **指示**：提供相关患者背景和检查原因

Table 2 | 人口统计。年龄统计显示各年龄段的百分比分布。性别统计展示按性别的百分比分布。

分割	年龄统计 (%)					性别统计 (%)		
	0-20	20-40	40-60	60-80	80+	M	F	U
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

- **比较**：提供相关患者背景和检查原因
- **发现**：详细的放射学观察
- **印象**：关键结论和建议的总结

为了确保所有研究的报告格式一致，我们提示 GPT-4o 从原始报告中提取四个关键部分。此外，我们实施了一个强大的后期处理验证步骤，以确保所有部分包含有效内容，并删除了原始报告中某些部分可能缺失的情况。

2.4. 人口分布

表 2 显示了训练、验证和测试分割中的人口分布情况。年龄分布显示所有分割中的一致性，40-80 岁的患者占最大比例（约占数据集的 50%）。性别分布在所有分割中男性与女性患者的比率较为均衡（约 49% 为男性，50% 为女性），并且未知性别信息的比例极小（1.3%-1.8%）。

3. 数据记录

所有数据均可在 Huggingface 上获得。访问 ReXGradient-160K 数据集需要在 Huggingface 上签署我们的许可协议。该数据集将图像组织到子文件夹中，每个子文件夹根据匿名患者标识符和研究 ID 进行命名。每位患者的子文件夹包含一个或多个研究，每个研究包含一张或多张图像。对于报告，我们提供了包括 PatientID（患者 ID）、AccessionNumber（登记号）、PatientSex（患者性别）、EthnicGroup（种族组别）、PatientAge（患者年龄）、PatientWeight（患者体重）、StudyDate（研究日期）、InstitutionName（机构名称）和 Manufacturer（制造商）在内的元数据。

致谢

本工作得到了 Biswas 家族基金会与 Milken 研究所合作的变革性计算生物学基金的支持。

披露声明

O.H 和 J.M 是 Gradient Health 的创始人，并持有该公司股权，Gradient Health 是一家专注于提高健康数据可获取性和可用性的私营公司。Gradient Health 提供了本研究中使用的数据集，并未为此项研究提供资金，也未参与其设计、执行或发表。

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.
- 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.
- 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.
- 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.
- 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, 2023.
- 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*, 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.