Benchmarking Foundation Models for Cervical Cancer CT Reporting in Zambia

Kangwa E. Mukuka[1,2]✉[0009−0001−5130−507X], Lena Lambart[3], Festus Mwape[1,2][0009−0005−9325−2198], Lighton Phiri[1,2][0000−0003−3582−9866], Concetto Spampinato[5][0000−0001−6653−2577], Karim Lekadir[6][0000−0002−9456−1612], and Marawan Elbatel[4][0009−0008−5021−4281]

[1] Department of Computing and Informatics, University of Zambia, Lusaka, Zambia
[2] DataLab Research Group, University of Zambia, Lusaka, Zambia
✉ kangwa1mukuka@cs.unza.zm
[3] Department of Oncology, Cancer Diseases Hospital, University Teaching Hospitals, Lusaka, Zambia
[4] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
mkfmelbatel@connect.ust.hk
[5] Department of Electrical, Electronic and Computer Engineering, University of Catania, Catania, Italy
[6] Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain

Abstract. Cervical cancer is the fourth most common malignancy among women worldwide, yet Africa bears incidence and mortality seven to eight times those of high-income regions. Its staging is inherently multi-organ under FIGO, yet automated reporting remains unexplored in low-resource settings. To address this gap, we introduce the first African benchmark for cervical cancer CT diagnosis and report generation. To support this, we curated 100 patients and 902 DICOM series with paired reports from a tertiary Zambian cancer hospital, spanning abdominopelvic and thoracic acquisitions. Under a unified protocol, we benchmarked four 3D CT foundation models with patient-level bootstrap intervals. Experiments show Pillar-0 attains the strongest performance, achieving an average AUC of 71.4% across 10 organs, yet no model meaningfully outperforms a trivial always-positive baseline on the cervical mass itself, revealing a domain gap. By releasing this benchmark publicly, we aim to catalyze research into accessible cervical cancer imaging. Code is available at: https://github.com/xmed-lab/TriALS-Report.

Keywords: Computed Tomography (CT) · Cervical Cancer (CaCx) · Radiology Report Generation · Low-Resource Settings

1 Introduction

Cervical cancer is the fourth most common malignancy among women worldwide in both incidence and mortality, yet its burden is profoundly inequitable [2, 5, 8].

An estimated 662,301 new cases and 348,874 deaths occurred globally in 2022 [11, 17, 27], and Africa carries incidence and mortality rates seven to eight times those of high-income regions [5]. Computed tomography (CT) is central to managing this burden, because treatment allocation depends on the FIGO stage, and the findings determining stage are distributed across the body rather than confined to the cervix [5]. Hydronephrosis from ureteric obstruction defines stage IIIB, retroperitoneal nodal involvement stage IIIC, and hepatic, pulmonary, or osseous metastases stage IVB. Staging one patient therefore requires a radiologist to interpret and report abdominopelvic and thoracic acquisitions together. This demand collides with a workforce constraint, since cross-sectional imaging volumes have grown while radiologist numbers have not [15, 16], and report authoring consumes a substantial share of the workflow beyond interpretation alone [25]. Staff shortages and reporting backlogs are most acute in precisely the settings where cervical cancer is most prevalent.

This local context has been thoroughly documented by the Enterprise Medical Imaging in Zambia (EMIZ) project[7]. Baseline studies conducted at the University Teaching Hospitals (UTH) revealed monotonous workflows and interrelated challenges exacerbated by a critical deficit of medical imaging experts, driving the urgent need for modern Enterprise Imaging (EI) implementations [29]. Specifically, with fewer than 15 trained radiologists in the Zambian public sector, the EMIZ project has actively laid the groundwork for AI-guided solutions by characterizing what constitutes a high-quality report in Southern Africa. Recent work under this umbrella established a comprehensive radiology report terminology, aggregating 3,522 terms from international templates and literature to standardize metadata and archetypes for structured reporting [19]. Automated report drafting is thus an opportunity for artificial intelligence to relieve a bottleneck that falls hardest on low-resource care, building directly upon this newly standardized reporting vocabulary [15, 22].

Despite this opportunity, the foundation models that could plausibly fill the gap have neither been trained on nor evaluated against African cohorts. Recent progress on 3D volumetric CT has produced several such models, Merlin [6], M3D [3], Pillar-0 [1] and SPECTRE [7], in contrast to a field that had concentrated on two-dimensional imagery [3, 26]. Every one of these systems was developed on North American, European, or Asian data acquired with modern scanners and consistent protocols, and where abdominal reporting benchmarks exist they target populations and pathologies far from this setting [10]. Consequently, no prior work has established whether these models transfer to a Sub-Saharan cervical cancer staging workflow, and no public dataset exists that would allow the question to be asked.

To bridge this gap, we introduce the first African benchmark for multi-label disease classification and automated report generation from cervical cancer CT. The foundational dataset curation and initial experimental framework for this benchmark were established as part of a computer science capstone project (CSC4004) at the University of Zambia by Mukuka and Mwape [20]. Expanding

[7] https://emi.org.zm

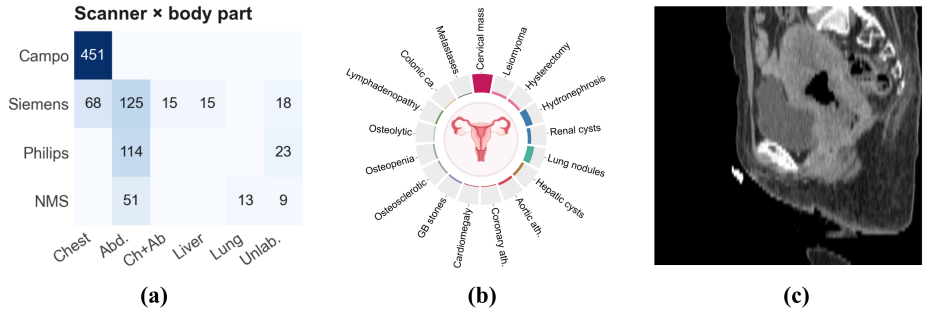

Fig. 1. The curated Zambian cervical cancer cohort. (a) Scanner against body-part protocol across the 902 DICOM series, showing the scanner heterogeneity of routine regional care. (b) The staging landscape: the cervical primary sits at the centre, and the 17 findings our benchmark evaluates orbit it, grouped and coloured by organ, with bar height proportional to positive exams in the test set. Several set FIGO stage in their own right, including hydronephrosis (IIIB), lymphadenopathy (IIIC), and distant metastases (IVB), which is why cervical cancer reporting cannot be reduced to the cervix alone. (c) A representative sagittal pelvic CT slice from the cohort. The centre illustration in (b) is a schematic of pelvic anatomy, not patient data.

upon this work, we curate a cohort of 100 patients and 902 DICOM series with paired radiology reports from a tertiary Zambian cancer hospital, spanning the abdominopelvic and thoracic acquisitions FIGO staging requires, and we lay out the resulting staging landscape in Figure 1. Under a unified protocol that leaves every encoder frozen, we evaluate four contemporary 3D CT foundation models, reporting patient-level bootstrap confidence intervals so that conclusions drawn from a modest cohort remain statistically honest. Our results carry a message a single aggregate number would conceal. Averaged across 10 organs and 17 staging-relevant findings, the strongest model reaches an AUC of 71.4%, comfortably above chance; yet on the cervical mass itself three of the four models are indistinguishable from chance, and none meaningfully surpasses a trivial baseline that always predicts the finding present. Aggregate competence therefore coexists with failure on the disease defining the cohort, which makes regionally representative evaluation a precondition rather than an afterthought for deployment in African radiology.

2 Methodology

2.1 Overall Benchmark

Our benchmark follows a four-stage workflow: (1) Patient Dataset Collection, curating paired CT series and radiology reports from a tertiary Zambian cancer hospital; (2) Structured Label Extraction, parsing staging-relevant findings from free-text reports into a traceable multi-label target space; (3) Model Development, evaluating four 3D CT foundation models with their encoders held frozen; and

Table 1. CaCx dataset characteristics ($n = 100$ patients; 902 DICOM series).

Feature	**Value**
Patients (Train/Test/Val)	100 (60/30/10)
DICOM Series (Total)	902
Scanner Manufacturer, n (%)	
Campo-Imaging	451 (50.0)
Siemens	241 (26.7)
Philips	137 (15.2)
Neusoft Medical Systems (NMS)	73 (8.1)
Body-Part Protocol, n (%)	
Chest	519 (57.5)
Abdomen	290 (32.2)
Other[a]	43 (4.8)
Unlabelled	50 (5.5)
Acquisition Phase, n (%)	
Contrast-enhanced[b]	256 (28.4)
Ancillary[c]	70 (7.8)
Other / Unlabelled	576 (63.9)
Volumetric Geometry	
Matrix Size	512×512
Slice Thickness (mm), mean	2.90 ($n = 730$)
Volume Depth (slices), median/max	192/933 ($n = 690$)

[a] Chest+abdomen, liver, lung.
[b] Arterial, venous, delayed phases.
[c] Bolus tracking, topogram/scout, non-contrast, exam summary.

(4) Model Evaluation, covering multi-label disease classification and zero-shot report generation with patient-level uncertainty quantification. The framework answers one question: whether representations learned on high-income cohorts transfer to an African cervical cancer staging workflow.

2.2 Dataset

Ethical clearance was granted by the University of Zambia Biomedical Research and Ethics Committee - UNZABREC (Reference Number: 2731-2022) and the National Health Research Authority - NHRA (Reference Number: NHRA000024/10/05/2022). Additionally, formal permission was granted by the Cancer Diseases Hospital, University Teaching Hospitals - CDH-UTHs (Reference Number: CDH/101/14/1). Before conducting any computational analysis, we de-identified all patient data in accordance with the Data Protection Act (No. 3 of 2021) of the Republic of Zambia [24].

Since no public CT dataset exists for cervical cancer in an African population, we curated a retrospective cohort of 100 patients comprising 902 unique DICOM

series with paired radiology reports, accrued between July 2025 and June 2026, as summarized in Table 1. Patients ranged from 24 to 80 years of age (median 53). The cohort was partitioned at the patient level into 60 training, 30 test, and 10 validation cases, with no patient contributing series to more than one split.

The cohort reflects routine regional care rather than a curated research protocol, and is heterogeneous along every axis we measured. Series span four manufacturers, dominated by Campo-Imaging (50.0%), and the anatomical breadth FIGO staging requires, with 57.5% chest and 32.2% abdominal protocols. Critically, 63.9% carry no usable acquisition-phase label, a consequence of inconsistent legacy DICOM tagging that places a hard ceiling on protocol-aware preprocessing, and spatial resolution is likewise variable, with a mean slice thickness of 2.90 mm and depths reaching 933 slices. This heterogeneity motivates the series-level quality control we apply before evaluation, and is itself a defining property of the setting this benchmark represents.

Preprocessing and Label Extraction. DICOM series were converted to NIfTI, and reports from DOCX to plain text. Out of 1,072 series which successfully matched DICOM-to-NIfTI conversions, duplicate entries arose from repeated rescan passes. The UIDs across near-identical visit folders were consolidated to yield the 902 unique series reported in Table 1. Thereafter the findings and impressions were parsed into JSON. We utilized LLM-based label-extraction, specifically Qwen2.5-14B-Instruct via vLLM per-finding evidence-quote-then-yes/no prompting with temperature=0. Structured clinical information was extracted using an anatomy-specific feature configuration spanning 17 anatomical categories, with each feature paired to its supporting sentence so that every label remains traceable to the text that produced it. This yields a candidate space of 232 binary findings. Because a 30-patient test cohort cannot support estimation for findings that never occur, we restrict quantitative analysis to the subset of those findings observed at least once, which numbers 17 and maps onto 10 organ groups; the remainder are retained in the released label space for future use.

2.3 Model Development

All four models are 3D CT architectures, and we keep every encoder frozen: the question is what these representations already carry about a Zambian cohort, not how far fine-tuning could take them. **Merlin** [6] and **Pillar-0** [1] target the abdomen. Merlin aligns an inflated 3D ResNet152 encoder with reports and health records; Pillar-0 reads the full bit depth of CT by projecting each volume patch into several intensity windows. **M3D** [3] and **SPECTRE** [7] are generalists, the first feeding a 3D vision transformer into a large language model backbone, the second pairing local and global transformers pretrained by self-supervised learning and aligned to report-level context. Only Merlin and M3D generate reports; the other two are scored on classification alone. We keep to models whose pretraining covers abdominopelvic anatomy, leaving chest-only systems such as CT2Rep [12] and BTB3D [13, 14] for future work.

2.4 Benchmark Tasks and Model Evaluation

The benchmark formalizes two tasks, multi-label disease classification and automated report generation, and decouples linguistic metrics from clinical efficacy so that diagnostic performance is never inferred from text overlap alone.

Classification Evaluation. Each frozen encoder produces a mean-pooled volume representation, over which we train a lightweight linear probe per finding on the 60 training patients and evaluate on the 30 held-out test patients; performance therefore reflects the linear separability of each finding in the pretrained feature space. We report **AUC** and **F1**, macro-averaged over the findings within each organ group and then over the 10 organ groups, so that organs with many findings do not dominate. Given only 30 test patients, point estimates alone would mislead, so every value carries a 95% confidence interval from patient-level bootstrap resampling with 1,000 iterations under a fixed seed, resampling exams rather than findings so that within-patient correlation is preserved. We further report two reference baselines: predicting each finding at its observed prevalence, and a trivial policy that always predicts the finding present. The latter is essential here, because several findings are prevalent enough to admit a deceptively high F1 from a constant predictor that computes nothing.

Report Generation Evaluation. Report generation is evaluated strictly zero-shot, with no adaptation, isolating what pretrained vision language alignment transfers to this population. We report **GREEN** [21], and **RadGraph-XL** (Entity, Partial, and Complete) [9] metrics emphasizing clinical correctness over rudimentary text-overlap. Additionally, we report traditional Natural Language Generation (NLG) metrics such as **BLEU** [23], **ROUGE** [18], **METEOR** [4], and **BERTScore** [28] strictly as descriptive endpoints. Text-overlap metrics are systematically flawed in the medical domain, since they reward templated boilerplate and fail to penalize catastrophic clinical errors such as hallucinated findings or reversed negations. Clinically grounded metrics such as GREEN [21] and RadGraph-XL [9] address this by scoring entity and relation correctness.

3 Results

Disease Diagnosis Performance. The diagnostic performance of the four frozen foundation models is detailed in Table 2. Averaged over the 10 organ groups spanning 17 staging-relevant findings, Pillar-0 attains the strongest discrimination with an AUC of 71.4% (95% CI [59, 80]), followed by Merlin at 59.5% ([52, 67]). The remaining two models are not distinguishable from chance: M3D reaches 48.1% ([38, 62]) and SPECTRE 44.9% ([36, 59]), and both intervals comfortably contain 50. Half of the contemporary 3D CT foundation models we evaluated therefore carry no usable linear signal about this cohort, a result that point estimates without intervals would have obscured.

The F1 column demands equal caution, and it is why we anchor it to a trivial always-positive baseline. Averaged over organ groups, only Pillar-0 exceeds that reference (25.5% against 19.1%), while Merlin, M3D, and SPECTRE fall at or

Table 2. Disease diagnosis performance on the Zambian cervical cancer test set (30 patients), reported as percentages with 95% bootstrap confidence intervals (1,000 iterations, patient-level resampling). All encoders are frozen and a linear probe is trained per finding. GU denotes genitourinary. Organ columns macro-average over the findings in that organ; *Average* macro-averages over all 10 organ groups spanning 17 findings, including organs not shown individually. Best result per column in bold. *Random* predicts each finding at its observed prevalence p. † marks an AUC whose interval includes 50, i.e. not distinguishable from chance. ‡ marks an F1 at or below the trivial always-positive baseline, $2p/(1+p)$.

Model	Cervix		GU		Average (10 organs)	
	AUC	**F1**	**AUC**	**F1**	**AUC**	**F1**
Random	*50.0*	*70.0*	*50.0*	*20.0*	*50.0*	*11.8*
Pillar-0 [1]	64.6† [42,86]	**82.6** [68,94]	**79.0** [64,91]	30.8‡ [10,43]	**71.4** [59,80]	**25.5** [10,30]
Merlin [6]	58.2† [36,80]	75.6‡ [58,88]	67.4 [52,84]	30.0‡ [0,54]	59.5 [52,67]	14.2‡ [8,20]
M3D [3]	58.2† [31,81]	79.1‡ [63,91]	58.5† [38,80]	**32.1**‡ [8,52]	48.1† [38,62]	16.0‡ [7,23]
SPECTRE [7]	**74.1** [51,93]	79.1‡ [63,91]	50.4† [32,74]	23.1‡ [0,40]	44.9† [36,59]	18.1‡ [7,23]

below a constant predictor that performs no computation whatsoever. Reporting F1 against an implicit chance level of zero, as is common practice, would have presented three failing models as partially competent.

Nowhere is this more consequential than on the cervix. That column carries the highest F1 values in the table, peaking at 82.6% for Pillar-0, which in isolation reads as strong detection of the primary tumour. It is not. Cervical masses are present in 21 of the 30 test patients, so a classifier answering "present" unconditionally scores 82.4%. Pillar-0 exceeds this trivial policy by roughly a quarter of a percentage point, and the other three fall below it. The AUC column confirms the diagnosis: three of the four cervix intervals span chance, Pillar-0 among them at 64.6% ([42, 86]), and the fourth clears it only barely, with SPECTRE at 74.1% ([51, 93]) resting a lower bound of 50.6 on the chance threshold while that same model is indistinguishable from chance on average. No model here detects the disease that defines the cohort.

Against that failure, what Pillar-0 does recover is clinically coherent. Its strongest reported result falls precisely on a secondary finding driving FIGO stage: genitourinary disease, which includes the hydronephrosis defining stage IIIB, reaches an AUC of 79.0% ([64, 91]), an interval excluding chance and sitting well above its own cervix performance, with Merlin following the same ordering at 67.4% ([52, 84]). The competence of these models is thus real but displaced. They recognize the generic obstructive findings populating their pretraining corpora, whereas the cervical primary, comparatively rare in the cohorts these models were built on and radiologically subtle on the heterogeneous legacy acquisitions

Table 3. Zero-shot radiology report generation on the held-out test set ($n = 30$ patients), evaluated with standard NLG metrics as well as GREEN and RadGraph-XL. No model is adapted to the cohort. BLEU-4, METEOR, BERTScore-F1, and ROUGE-L are reported as descriptive endpoints only; text-overlap and embedding-similarity scores are unreliable proxies for clinical correctness, and BERTScore in particular saturates on fluent radiological prose irrespective of factual accuracy.

Region	Model	BLEU-4 ↑	METEOR ↑	BERTScore-F1 ↑	ROUGE-L ↑	GREEN ↑	RadGraph-XL ↑		
							E	P	C
Abd/Pelvis	Merlin [6]	0.2808	5.7006	80.7354	13.6803	27.7844	22.6901	18.1423	15.9442
	M3D [3]	0.1248	5.1968	81.4756	9.0304	0.4762	3.7147	2.9909	1.3197
Chest	M3D [3]	0.5253	6.5172	82.2979	9.3812	1.4483	2.8104	2.3840	1.5448

of Table 1, remains invisible. Pillar-0's margin is consistent with its abdomen-pelvis pretraining domain and its use of the full CT bit depth, plausibly more robust to this cohort's scanner variability than single-window encoding. Notably, the F1 ordering does not track AUC: M3D posts the nominally best genitourinary F1 (32.1%) while sitting at chance in AUC, a further symptom of F1's sensitivity to prevalence rather than evidence of discrimination.

Report Generation. Table 3 reports zero-shot report generation on the test set for the two models exposing a generative interface, and the results are uniformly far below clinical usability. Note that since Pillar-0 and SPECTRE do not support report generation, they are evaluated exclusively on clinical classification metrics. BLEU-4 remains below 0.6 in every region, ROUGE-L spans roughly 9 to 14, and METEOR roughly 5 to 6.5, indicating that generated text shares little more than incidental vocabulary with the references. BERTScore alone appears healthy at approximately 81 to 82, illustrating the pathology motivating our design, since embedding similarity saturates on fluent radiological prose regardless of whether its findings are correct. Read against the classification results this is unsurprising: models that cannot linearly separate the cervical primary have no basis on which to describe it. GREEN roughly ranges between 0.4 to below 28 while RadGraph-XL ranges from 1.3 to below 23 across Entity, Partial, and Complete metrics. The GREEN and RadGraph-XL values indicate that Merlin reflects a modest but genuine clinical signal invisible to BERTScore. Although Merlin performs higher than M3D in both GREEN and RadGraph-XL, these results remain far below clinical usability, confirming that report generation has failed to transfer to this cohort as thoroughly as classification did.

4 Discussion

This study introduces the first African benchmark for disease classification and automated report generation from cervical cancer CT, curated from a tertiary Zambian cancer hospital and released to the community. Its central finding is a dissociation that a single headline number would have hidden. Averaged across 10 organ groups, the strongest model recovers a statistically defensible signal, with Pillar-0 reaching an AUC of 71.4% ([59, 80]); yet on the cervical primary that

defines the cohort, three of the four models are indistinguishable from chance and none meaningfully exceeds a predictor that answers "present" every time. Two of the four models fail to beat chance even on average. What the models do recover is displaced onto secondary findings well represented in their pretraining corpora, most clearly the genitourinary disease carrying the hydronephrosis that defines FIGO stage IIIB. Representations learned on high-income cohorts thus transfer for generic abdominal pathology but not for the disease that matters most here, making regionally representative data a precondition for deployment rather than a refinement of it.

These findings exist within boundaries we state plainly. The test cohort comprises 30 patients and our intervals are correspondingly wide; we report intervals rather than withhold the benchmark, since the alternative in a data-scarce region is that no evaluation is published at all, and where intervals overlap we refrain from ranking models. The cohort is single-centre, confounding scanner and population effects. Holding encoders frozen isolates what pretraining encodes but does not establish what fine-tuning on African data could achieve, so the gap is not a fixed ceiling. Lightweight strategies such as LoRA adaptation or deeper, non-linear probing heads are natural candidates for testing this directly. Our report generation evaluation rests on NLG metrics, and extends to GREEN, and RadGraph-XL to overcome text-overlap. These metrics were not validated against radiologist assessment on this cohort, leaving it to future work. Finally, 63.9% of series carry no usable phase label, precluding phase-aware modelling, a faithful property of the setting rather than a defect of curation.

The clinical implication follows directly. These models cannot presently support cervical cancer staging here, because staging begins with the primary tumour and that is precisely where they fail; their competence on secondary findings suggests only a narrower role, flagging the sequelae that alter stage once a radiologist has established the primary. We release this benchmark to catalyze regionally representative cervical cancer imaging research.

Acknowledgments. This study was conducted as part of the much larger "Enterprise Medical Imaging in Zambia" project (EMI x Z), funded by generous grants from Google Research, Data Science Africa, and The University of Zambia. Additionally, this study was partially supported by the European Union – Next Generation EU, Mission 4 Component 2 Line 1.3, through the PNRR MUR project PE0000013 – FAIR "Future Artificial Intelligence Research" (CUP E63C22001940006).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, K.K., Liu, L., Lian, L., Nercessian, M., Yala, A., et al.: Pillar-0: A new frontier for radiology foundation models. arXiv preprint arXiv:2511.17803 (2025)
2. Arroyo Mühr, L.S., Dillner, J.: Challenges and possible solutions towards reaching global cervical cancer elimination. eClinicalMedicine p. 103890 (2026)

3. Bai, F., Du, Y., Huang, T., et al.: M3D: Advancing 3D medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005), https://aclanthology.org/W05-0909/
5. Bhatla, N., Aoki, D., Sharma, D.N., et al.: Cancer of the cervix uteri: 2025 update. International Journal of Gynecology & Obstetrics **171**(S1), 87–108 (2025)
6. Blankemeier, L., Kumar, A., Cohen, J.P., Chaudhari, A., et al.: Merlin: a computed tomography vision–language foundation model and dataset. Nature **652**(8112), 1318–1328 (2026). https://doi.org/10.1038/s41586-026-10181-8
7. Claessens, C., Viviers, C., , et al.: Scaling self-supervised and cross-modal pretraining for volumetric CT transformers. arXiv preprint arXiv:2511.17209 (2025)
8. Cohen, P.A., Jhingran, A., Oaknin, A., Denny, L.: Cervical cancer. The Lancet **393**(10167), 169–182 (2019). https://doi.org/10.1016/S0140-6736(18)32470-X
9. Delbrouck, J.B., Chambon, P., et al.: RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In: Findings of the Association for Computational Linguistics (ACL). pp. 12902–12915 (2024)
10. Elbakry, M., et al.: A multi-center benchmark for abdominal disease diagnosis and report generation from non-contrast CT. arXiv preprint arXiv:2606.16991 (2026)
11. Global Cancer Observatory: Global Cancer Observatory. https://gco.iarc.who.int/en, last accessed 2026/06/18
12. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated radiology report generation for 3D medical imaging. arXiv preprint arXiv:2403.06801 (2024)
13. Hamamci, I.E., Er, S., Shit, S., Reynaud, H., Yang, B., et al.: Better tokens for better 3D: Advancing vision-language modeling in 3D medical imaging. arXiv preprint arXiv:2510.20639 (2025)
14. Hamamci, I.E., Er, S., Wang, C., Almas, F., et al.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nature Biomedical Engineering **10**(8), 1610–1628 (2026). https://doi.org/10.1038/s41551-025-01599-y
15. Jing, A.B., Garg, N., Zhang, J., Brown, J.J.: AI solutions to the radiology workforce shortage. npj Health Systems **2**(1), 20 (2025)
16. Kwee, T.C., Kwee, R.M.: Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence. Insights into Imaging **12**(1), 88 (2021)
17. Li, Z., Liu, P., Yin, A., et al.: Global landscape of cervical cancer incidence and mortality in 2022 and predictions to 2030: The urgent need to address inequalities in cervical cancer. International Journal of Cancer **157**(2), 288–297 (2025)
18. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (ACL Workshop). pp. 74–81 (2004)
19. Mahlaza, Z., Zulu, E.O., Phiri, L.: Radiology report terminology to characterise reports in southern africa. In: Garoufallou, E., Sartori, F. (eds.) Metadata and Semantic Research. pp. 147–154. Springer Nature Switzerland, Cham (2024)
20. Mukuka, K.E., Mwape, F.: Benchmarking foundation models for cervical cancer CT reporting in Zambia. Undergraduate capstone project (csc4004), University of Zambia (2025)
21. Ostmeier, S., Xu, J., Chen, Z., et al.: GREEN: Generative radiology report evaluation and error notation. arXiv preprint arXiv:2405.03595 (2024)
22. Paiboonborirak, C., Abu-Rustum, N.R., Wilailak, S.: Artificial intelligence in the diagnosis and management of gynecologic cancer. International Journal of Gynecology & Obstetrics **171**(S1), 199–209 (2025)

23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proc. ACL. pp. 311–318 (2002)
24. Republic of Zambia: Data Protection Act No. 3 of 2021. Government Printer, Lusaka (2021)
25. Troupis, C.J., Knight, R.A.H., Lau, K.K.P.: What is the appropriate measure of radiology workload: Study or image numbers? Journal of Medical Imaging and Radiation Oncology **68**(5), 530–539 (2024)
26. Wu, C., Zhang, X., et al.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. arXiv preprint arXiv:2308.02463 (2023)
27. Wu, J., Jin, Q., Zhang, Y., et al.: Global burden of cervical cancer: current estimates, temporal trend and future projections based on the GLOBOCAN 2022. Journal of the National Cancer Center **5**(3), 322–329 (2025)
28. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675 (2019)
29. Zulu, E.O., Phiri, L.: Enterprise medical imaging in the global south: Challenges and opportunities. In: 2022 IST-Africa Conference (IST-Africa). pp. 1–9 (2022). https://doi.org/10.23919/IST-Africa56635.2022.9845508