QUANTITATIVE EVALUATION OF THE DIAGNOSTIC ACCURACY OF ChatGPT IN THE ADULT CHEST COMPUTED TOMOGRAPHY IMAGES: A PHANTOM STUDY

Mark Anthony C. Burgonio[1], Luigie O. Cabigon[2], Sarbelle Rose F. Costales[3], Irah Ledz D. Lelina[4], Miki Ela A. San Luis[5], Angelica B. Sapitula[6], Jeremiah M. Supnad[7], Marites C. Pagdilao[8], Gryn T. Salagma[9], Bernardo B. Tayaban Jr.[10], Joylyn P. Baniaga[11], Godofredo M. Manzano Jr.[12]

[1]College of Radiologic Technology, Lorma Colleges, markanthony.burgonio@lorma.edu, [2]College of Radiologic Technology, Lorma Colleges, luigie.cabigon@lorma.edu, [3]College of Radiologic Technology, Lorma Colleges, sarbellerose.costales@lorma.edu, [4]College of Radiologic Technology, Lorma Colleges, irahledz.lelina@lorma.edu, [5]College of Radiologic Technology, Lorma Colleges, mikiela.sanluis@lorma.edu, [6]College of Radiologic Technology, Lorma Colleges, angelica.sapitula@lorma.edu, [7]College of Radiologic Technology, Lorma Colleges, jeremiah.supnad@lorma.edu, [8]Research Institute, Lorma Colleges, marites.pagdilao@lorma.edu, [9]College of Radiologic Technology, Lorma Colleges, gryn.salagma@lorma.edu , [10]College of Physical Therapist, Lorma Colleges, bernardo.tayaban@lorma.edu, [11]College of Nursing, Lorma Colleges, godofredo.manzano@lorma.edu, [12]College of Inclusive Education, Lorma Colleges, joylyn.baniaga@lorma.edu

Abstract

This study aimed to assess the quantitative diagnostic accuracy of ChatGPT in interpreting adult chest computed tomography (CT) phantom images using contrast-to-noise ratio (CNR) as the reference standard. A quantitative descriptive–comparative research design was employed, utilizing thirty (30) CT phantom images acquired under adult chest CT protocols with varying milliampere-seconds (mAs). Objective image quality was evaluated through computed CNR values, while the same images were assessed by ChatGPT using a standardized evaluation approach. The results of both methods were statistically compared using a paired t-test. Findings revealed that quantitative CNR values exhibited wide variability, including both positive and negative results, reflecting sensitivity to changes in image quality. In contrast, ChatGPT-generated values were consistently positive and showed minimal variation, indicating stable but generalized outputs. Comparative analysis demonstrated that ChatGPT consistently produced higher evaluation scores than the quantitative method. The paired t-test confirmed a statistically significant difference between the two methods ($p = 0.002$). These results suggest that ChatGPT lacks sensitivity to variations in CT image quality and tends to overestimate results. While it may serve as a supplementary tool, quantitative evaluation remains essential for accurate assessment of image quality.

Keywords: *CNR, CT phantom, ChatGPT, diagnostic accuracy, image quality assessment, computed tomography*

1. Introduction

Computed tomography (CT) has become one of the most critical diagnostic modalities in modern medicine, enabling three-dimensional visualization of internal anatomical structures that surpasses the capabilities of conventional X-ray techniques. Globally, CT is used for an estimated 375 million scans annually, underscoring its indispensable role in supporting accurate clinical decision-making across diverse patient populations (Frost & Sullivan, 2025). As worldwide CT utilization continues to expand, ensuring consistent and reliable image quality has become a paramount concern for radiologic professionals, researchers, and healthcare institutions alike. Traditionally, CT image quality was assessed through subjective clinical judgment; however, recent literature increasingly emphasizes the need for objective, quantitative evaluation methods that provide standardized and reproducible measurements of diagnostic adequacy (Hoeijmakers et al., 2025). Quantitative metrics such as the contrast-to-noise ratio (CNR), signal-to-noise ratio (SNR), and spatial resolution enabled radiologists to characterize image clarity and lesion detectability with greater precision, reducing the inherent variability of visual assessment alone and forming the scientific foundation for quality-driven CT practice (Hoxha et al., 2024).

Improving CT imaging quality and maximizing diagnostic reliability are essential in addressing persistent national health burdens in the Philippines. Suanes et al. (2023) emphasized that advancing lung cancer screening requires improved funding, region-specific imaging protocols, and multidisciplinary collaboration supported by public–private partnerships. Alongside enhanced diagnostic technologies, these measures may strengthen early detection efforts and improve long-term clinical outcomes for Filipino patients. Therefore, ensuring high-quality CT imaging remains important in supporting accurate diagnosis and effective healthcare delivery.

The integration of innovative artificial intelligence tools, such as ChatGPT, into radiologic workflows represents a promising avenue to support clinical practice by reducing diagnostic workload, improving efficiency, and enabling faster interpretation of imaging results. International studies have demonstrated moderate diagnostic performance of ChatGPT and vision-enabled language models in image-based tasks; for instance, Dehdab et al. (2024) reported that ChatGPT-4 Vision achieved a diagnostic accuracy of 56.8% when interpreting chest CT slices for COVID-19 and lung cancer. These findings indicated both the potential and limitations of ChatGPT in radiology, underscoring the need for studies that directly link objective image-quality metrics to its interpretive performance under controlled conditions. The potential benefits of ChatGPT-assisted evaluation align strongly with Sustainable Development Goal (SDG) No. 3, which aims to ensure healthy lives and promote well-being for all (United Nations, 2015). By investigating how ChatGPT evaluates adult chest CT phantom images at varying CNR levels, this study contributes to global and national efforts to improve diagnostic accuracy, streamline workflows, and enhance the quality of healthcare delivery.

Alongside the growing clinical importance of quantitative metrics such as CNR, recent phantom-based studies have consistently highlighted the value of objective image-quality assessment across different dose conditions. Li et al. (2022) performed a multi-scanner phantom study evaluating low-dose CT image quality across five CT systems and found that CNR consistently served as a reliable and reproducible metric for comparing device performance at reduced dose levels. A separate phantom study utilizing a chest

model demonstrated that deep learning–based image reconstruction (DLIR) can markedly enhance both CNR and SNR compared with conventional filtered back-projection and iterative reconstruction methods, even at lower milliampere-second (mAs) settings (Jung et al., 2023). These findings reinforce the established role of phantom imaging as a controlled, scientifically rigorous approach to evaluating how acquisition parameters affect measurable image-quality indicators. Collectively, the evidence supports the use of phantom-based CNR assessment as a gold-standard method for benchmarking CT image quality, making it an appropriate reference framework for evaluating the performance of emerging computational tools such as ChatGPT. Furthermore, a phantom study utilizing a chest model demonstrated that deep learning–based image reconstruction (DLIR) can markedly enhance both CNR and SNR compared with conventional filtered back-projection and iterative reconstruction methods, even at lower mAs settings (Jung et al., 2023).

Despite the growing integration of artificial intelligence in medical imaging, Guo et al. (2024) explicitly acknowledged that the application of large language models, such as ChatGPT, to computed tomography image quality assessment remains largely unexplored, particularly in objective, quantitative evaluation contexts. While previous studies, such as Dehdab et al. (2024), have examined ChatGPT's diagnostic performance on clinical CT images, none have evaluated its ability to numerically replicate quantitative image-quality metrics, such as the contrast-to-noise ratio (CNR), derived from phantom images acquired under controlled exposure conditions. Existing phantom-based CT image quality studies have firmly established quantitative pixel-value methods as the gold standard for objective assessment, yet no study has introduced ChatGPT as a comparative evaluation tool against manually computed CNR values across varying milliampere-second (mAs) settings. A scoping review of large language models in radiology, covering studies from 2022 to 2024, similarly confirmed that LLM applications remain concentrated on text-based report generation and clinical decision support, with no study evaluating LLM performance against quantitative CT imaging parameters (PMC, 2025). This significant gap in the literature underscored the urgent need for a study that directly compares ChatGPT-generated CNR evaluations with objective quantitative measurements in adult chest CT phantom images. The present investigation seeks to address this gap by providing empirical evidence of ChatGPT's diagnostic accuracy in a controlled phantom setting.

Overall, the growing interdependence between quantitative CT image quality assessment and emerging ChatGPT-based diagnostic tools represents a critical point of convergence for advancing radiologic practice and AI-assisted imaging evaluation. The need to rigorously assess whether ChatGPT can reliably replicate objective CNR measurements is both timely and clinically significant, given the increasing interest in AI tools as potential supplements to traditional radiologic workflows. Thus, this study is undertaken to evaluate the quantitative diagnostic accuracy of ChatGPT in interpreting adult chest CT phantom images in relation to objectively computed CNR measurements obtained under varying mAs exposure settings. The findings were expected to provide evidence that supports or challenges the utility of ChatGPT as a supplementary evaluation tool in CT image quality assessment. Ultimately, this research aimed to contribute to improved diagnostic care, strengthen radiology education, and advance progress toward national and global health goals aligned with SDG No. 3.

The findings of this study aimed to provide empirical evidence supporting the safe, effective, and purposeful integration of ChatGPT into CT imaging environments, particularly as a supplementary tool for image-quality evaluation alongside established quantitative methods. By systematically comparing ChatGPT-generated CNR estimates with manually computed values from adult chest CT phantom images, the study provides a rigorous, controlled assessment of the model's numerical accuracy and consistency. This evidence is expected to inform radiologists, medical educators, healthcare institutions, and policymakers on the appropriate scope and limitations of AI-assisted imaging tools in clinical and research settings. The study further contributed to the broader literature on large language model performance on quantitative medical imaging tasks, an area that remains under investigated despite the rapid proliferation of AI in healthcare. In doing so, it reinforces the importance of evidence-based adoption of emerging technologies in diagnostic radiology, ensuring that innovations such as ChatGPT are integrated responsibly and with clearly understood boundaries.

The primary beneficiaries of this research are patients, who will receive more accurate, timely, and reliable diagnoses through improved CT image-quality assessment practices. A more objective and standardized evaluation framework may reduce the risk of misdiagnosis and unnecessary radiation exposure, both of which carry significant consequences for patient health outcomes. By supporting earlier and more accurate detection of thoracic diseases, including lung cancer, pneumonia, and cardiovascular conditions, this study directly contributes to improved treatment planning and disease management. Patients in the Philippines, where chest-related diseases remain among the leading causes of mortality, stand to benefit particularly from advances in diagnostic accuracy and imaging standardization. Ultimately, integrating reliable AI-assisted evaluation tools into clinical workflows has the potential to enhance the overall quality of care for patients undergoing chest CT examinations.

Radiologists also stood to benefit significantly from enhanced workflow efficiency enabled by AI-assisted evaluation tools such as ChatGPT. ChatGPT-assisted evaluations can aid case prioritization, support diagnostic decision-making, and reduce cognitive load when interpreting large volumes of imaging data. By handling certain routine evaluation tasks, ChatGPT may allow radiologists to focus their expertise on more diagnostically complex cases that require nuanced clinical judgment and specialized knowledge. This redistribution of cognitive effort has the potential to reduce burnout, improve turnaround times for imaging reports, and enhance overall radiologic service delivery. Furthermore, understanding ChatGPT's quantitative capabilities and limitations equips radiologists with the knowledge needed to use AI tools judiciously and responsibly within evidence-based practice frameworks.

Healthcare institutions may also experience significant operational and financial advantages by adopting standardized, AI-assisted imaging-quality assessment practices. Theintegration of objective imaging protocols and validated evaluation tools can optimize resource allocation by reducing variability in CT acquisitions and minimizing repeat examinations due to suboptimal image quality. Consistent application of quantitative standards, such as CNR, also supports more uniform patient care across facilities, technologists, and imaging equipment. From a quality assurance perspective, the availability of AI tools capable of rapidly and consistently evaluating image quality may strengthen institutional compliance with national and international

diagnostic reference levels. Ultimately, the evidence generated by this study may inform institutional decision-making regarding the appropriate role of ChatGPT and similar tools within hospital-based CT quality management programs.

Medical educators and radiologic technology trainees could also derive significant educational benefits from incorporating ChatGPT into radiologic practice and academic instruction. This technology provides an innovative, interactive educational tool that enhances understanding of image quality assessment, diagnostic reasoning, and the application of quantitative evaluation methods in clinical contexts. Exposure to AI-assisted tools during training prepares future radiologic professionals to navigate a healthcare landscape increasingly shaped by machine learning and large language model technologies. The study's findings can be integrated into curriculum discussions on CT imaging physics, quality assurance, and the critical appraisal of AI-generated outputs in medical settings. By grounding the use of ChatGPT in empirical evidence, educators can guide trainees in developing both the technical competencies and the critical thinking skills needed to evaluate and responsibly apply AI tools in their future professional practice.

Finally, policymakers and public health authorities stood to benefit from the empirical evidence generated by this study in shaping evidence-based regulations and guidelines for the use of AI in medical imaging. The findings can inform decisions on integrating large language models, such as ChatGPT, into national quality assurance frameworks, helping establish appropriate boundaries for their clinical application. Policymakers may also use this research to promote safer imaging practices by supporting the adoption of standardized, quantitative image-quality metrics, such as CNR, across public and private healthcare facilities. Advancing such standards is particularly relevant in the Philippines, where the absence of national CT screening guidelines and inconsistencies in imaging protocols continue to pose challenges for early disease detection and patient safety. In doing so, policymakers help strengthen national healthcare policies and achieve broader population health goals, including reducing preventable mortality from chest diseases such as lung cancer and pneumonia.

2. Objectives

This study aimed to determine the quantitative diagnostic accuracy of ChatGPT in interpreting adult chest computed tomography phantom images.

3. Materials and Methods

This study employed a quantitative, descriptive, and comparative research design to determine the contrast-to-noise ratio (CNR) of adult chest CT phantom images acquired under varying milliampere-second (mAs) settings and to evaluate the diagnostic accuracy of ChatGPT across these image quality levels.

The study was conducted at Lorma Cancer Institute in the City of San Fernando, La Union. CT image acquisition was performed using a Philips CT scanner under the supervision of the Radiation Oncology Department personnel. A Catphan Phantom 150 with serial number 50034197 was utilized during image acquisition.

A total of thirty (30) CT phantom image slices constituted the sample of the study. The phantom image slices were obtained from five phantom groups scanned at varying tube current-exposure time products (mAs), specifically 300 mAs, 250 mAs, 200 mAs, 150 mAs, and 100 mAs, to generate images with varying levels of image noise and Contrast-

to-Noise Ratio (CNR).

The researchers used RadiAnt DICOM Viewer to analyze the CT phantom image slices and obtain the necessary quantitative image quality parameters, including the mean signal intensity of the object, the mean signal intensity of the background, and the standard deviation of the background noise required for the Contrast-to-Noise Ratio (CNR) computation. The Contrast-to-Noise Ratio (CNR) was computed using the following formula: $CNR = (\bar{x}T - \bar{x}bg) / \sigma bg$

Where $\bar{x}T$ denoted the mean signal intensity of the object, $\bar{x}bg$ denoted the mean signal intensity of the background, and σbg represented the standard deviation of background noise. Higher CNR values indicated improved image contrast and diagnostic detectability.

After the quantitative evaluation, the same CT phantom image datasets were converted to high-resolution JPEG format and subsequently evaluated using ChatGPT via a standardized evaluation prompt.

A paired t-test was conducted to determine whether there was a statistically significant difference between the quantitative CNR values obtained by the researchers and the ChatGPT-generated evaluations.

4. Results

The results of the study showed variations in image quality across the five CT phantoms with different mAs settings. Quantitative evaluation demonstrated both positive and negative CNR values, indicating fluctuations in image quality and sensitivity to changes in image noise and contrast.

Table 1. Diagnostic Accuracy in Terms of CNR of Adult Chest CT Phantom Images with Varying mAs

Phantom	mAs	Slice	Mean1	Mean2	SD	CNR	CNR Evaluation
P1	300	104	113.56	98.85	7.50	1.99	Good
		105	116.06	100.16	8.21	1.95	Good
		106	112.96	98.61	6.58	2.13	Good
		107	108.87	98.36	7.57	1.27	Good
		108	99.56	98.24	5.35	0.22	Good
		109	98.15	98.25	8.09	0.07	Good
Average CNR = 1.27							
P2	250	104	116.23	98.65	8.40	2.11	Good
		105	114.14	99.88	8.72	1.66	Good
		106	112.42	99.75	8.55	1.46	Good
		107	92.85	98.26	8.11	-0.67	Poor
		108	88.78	100.16	7.63	-1.54	Poor
		109	85.58	99.52	7.78	-1.80	Poor
Average CNR = 0.20							
P3	200	107	110.25	98.03	7.70	1.60	Good
		108	113.54	99.15	9.43	1.53	Good
		109	111.44	98.06	8.96	1.56	Good
		110	107.13	98.00	9.75	0.93	Good
		111	92.79	98.88	9.03	-6.73	Poor

		112	90.15	97.86	9.39	-0.82	Poor
Average CNR = -0.32							
P4	150	104	113.52	98.62	9 .71	1.56	Good
		105	112.91	98.57	9.73	1.49	Good
		106	112.35	98.44	9.91	1.40	Good
		107	111.07	99.18	8.87	1.39	Good
		108	97.55	97.09	11.51	-0.62	Poor
		109	91.40	98.34	11.98	-0.58	Poor
Average CNR = 0.77							
P5	100	104	114.29	98.28	13.05	1.19	Good
		105	114.94	99.17	12.63	1.23	Good
		106	99.60	99.97	12.19	-0.05	Poor
		107	92.90	99.53	12.71	-0.52	Poor
		108	91.66	99.30	11.88	-0.65	Poor
		109	86.79	99.45	11.27	-1.14	Poor
Average CNR = 0.01							

Legend: P1-P5: Phantom1- Phantom05

Table 1 presents the diagnostic accuracy in terms of CNR of adult chest CT phantom images with varying mAs. Phantom P1 with 300 mAs obtained an average CNR value of 1.27 and was evaluated as Good. Phantom P2 with 250 mAs obtained an average CNR value of 0.20. Phantom P3 with 200 mAs produced an average CNR value of -0.32, indicating poor image quality conditions in some slices. Phantom P4 with 150 mAs generated an average CNR value of 0.77, while Phantom P5 with 100 mAs showed an average CNR value of 0.01. The results demonstrated that decreasing mAs levels generally reduced image quality due to increased image noise and lower contrast detectability.

Table 2. ChatGPT Diagnostic Accuracy in the Evaluation of Adult Chest CT Phantom Images with Varying mAs

Phantom	mAs	Slice	Mean1	Mean2	SD	CNR	CNR Evaluation
P1	300	104	24.9	5.5	6.4	3.06	Good
		105	24.8	5.6	5.9	3.26	Good
		106	24.9	5.6	6.4	3.02	Good
		107	24.9	5.6	5.85	3.34	Good
		108	24.9	5.6	6.45	3.00	Good
		109	23.9	5.6	7.65	2.43	Good
Average CNR = 3.18							
P2	250	104	28.5	5.0	5.25	3.92	Good
		105	25.5	5.0	5.2	3.95	Good
		106	25.5	5.0	5.2	3.95	Good
		107	25.5	5.0	5.2	3.95	Good
		108	25.5	5.0	5.2	3.95	Good
		109	25.5	5.0	5.2	3.95	Good
Average CNR = 3.95							
P3	200	107	23.5	6	7.25	2.42	Good
		108	23.4	6	5.58	2.40	Good

		109	23.4	6	7.25	2.40	Good
		110	23.4	6	7.25	2.40	Good
		111	23.4	6	7.25	2.40	Good
		112	23.4	6	7.25	2.40	Good
Average CNR = 2.40							
P4	150	102	24.5	6	6.25	2.96	Good
		103	24.4	6	6.25	2.95	Good
		104	24.4	6	6.25	2.95	Good
		105	24.4	6	6.25	2.95	Good
		106	24.4	6	6.25	2.95	Good
		107	40.6	0.8	12.50	3.19	Good
Average CNR = 2.99							
P5	100	104	38.1	2.3	15.6	2.33	Good
		105	37.1	2.3	16.5	2.13	Good
		106	36.5	2.3	17.5	1.76	Good
		107	35.5	2.3	18.5	1.77	Good
		108	35.5	2.3	19.5	1.69	Good
		109	34.5	2.3	20.5	1.57	Good
Average CNR = 1.88							

Legend: P1-P5: Phantom1- Phantom5

Table 2 presents the ChatGPT diagnostic accuracy in the evaluation of adult chest CT phantom images with varying mAs. The results showed consistently positive CNR values across all slices and phantoms. Average ChatGPT CNR values were 3.18 for P1, 3.95 for P2, 2.40 for P3, 2.99 for P4, and 1.88 for P5. All evaluations were classified as Good despite variations in mAs exposure settings.

Table 3.1 Comparison Between Quantitative Evaluation and ChatGPT on Image Quality of Adult Chest CT Phantom Images

Phantom	mAs	Manual Evaluation	ChatGPT Evaluation
P1	300	1.27	3.18
P2	250	0.20	3.95
P3	200	-0.32	2.40
P4	150	0.77	2.99
P5	100	0.01	1.88

Legend: P1-P5: Phantom1- Phantom5

Table 3.1 presents the comparison between quantitative evaluation and ChatGPT on image quality of adult chest CT phantom images. The results showed a clear discrepancy between the two methods, with ChatGPT consistently yielding higher CNR values compared to the manual evaluation. For instance, P2 demonstrated a manual CNR value of 0.20, while ChatGPT produced a significantly higher value of 3.95. Similarly, in P3, the manual evaluation yielded a negative CNR value of −0.32, whereas ChatGPT still reported a positive value of 2.40.

	Mean	Variance	t-Value	df	t-crit	p-Value	Reject	Remarks
Manual Evaluation	0.37	0.40	-7.15	4	2.13	0.002	Reject Ho	Significant Difference
ChatGPT Evaluation	2.88	0.62						

Table 3.2 presents the summary of statistical values between the quantitative evaluation and ChatGPT evaluation. The manual evaluation yielded a mean CNR of 0.37, whereas ChatGPT produced a mean of 2.88. The computed t-value was −7.15 with a p-value of 0.002. Since the p-value was lower than the 0.05 significance level, the null hypothesis was rejected, indicating a statistically significant difference between the quantitative evaluation and ChatGPT evaluation.

5. Discussion

The findings demonstrated that quantitative CNR evaluation effectively identified variations in CT image quality across different acquisition settings. The presence of both positive and negative CNR values confirmed the sensitivity of the method in distinguishing acceptable from poor-quality images.

In contrast, ChatGPT-generated evaluations consistently produced positive and less variable CNR values. This indicated that ChatGPT may lack sensitivity to subtle and extreme changes in image quality, particularly in conditions where image noise dominates.

The results further showed that ChatGPT consistently overestimated image quality relative to the quantitative evaluation. While ChatGPT provided stable and consistent outputs, it failed to adequately reflect degraded image conditions observed in lower mAs settings.

These findings support Image Quality Theory and Signal Detection Theory, which explain that increased image noise from lower mAs settings reduces contrast resolution and diagnostic performance.

Overall, the findings reinforce the importance of quantitative evaluation as the standard for objective and sensitive assessment of CT image quality while emphasizing that AI-assisted tools such as ChatGPT should only serve as supplementary evaluation methods.

6. Conclusion

The study demonstrated that significant differences exist between quantitative evaluation and ChatGPT-assisted evaluation of adult chest CT phantom images. Quantitative evaluation revealed fluctuating image quality across varying mAs settings, including positive and negative CNR values, while ChatGPT consistently produced higher and more stable image quality evaluations.

The paired t-test confirmed a statistically significant difference between the two methods ($p = 0.002$), indicating that ChatGPT tends to overestimate image quality and lacks sensitivity to variations in image noise and contrast.

Although ChatGPT may serve as a supplementary image assessment tool, quantitative evaluation remains essential for accurate and objective CT image quality assessment.

7. Acknowledgements

The researchers would like to express their profound gratitude to the following individuals who had a significant impact throughout their research journey.

First and foremost, to Almighty God, for His unending support, guidance, and blessings. His wisdom and strength allowed the researchers to overcome obstacles and successfully complete this study.

To Marites C. Pagdilao, MAN, MPA, their Research Instructor, for her continuous support, dedication, and guidance.

To Mark Anthony C. Burgonio, MSc, their esteemed research teacher, whose invaluable experience, patience, and insights guided the researchers throughout this project.

To the distinguished members of the panel headed by Bernardo B. Tayaban Jr., PhD, MDA, RPT, together with Joylyn P. Baniaga, LPT, MAME, and Godofredo M. Manzano Jr., MAN, who shared their expertise and took time to review the progress of the study.

Special thanks are extended to Lorma Cancer Institute and the Radiation Oncology Department staff for allowing the researchers to conduct the CT phantom image acquisition and for their cooperation during the data gathering process.

8. References

Brady, A. P., Neri, E., & Regge, D. (2022). Artificial intelligence in radiology: Opportunities, challenges, and limitations. *Insights into Imaging, 13*(1), 1–12. https://doi.org/10.1186/s13244-021-01117-3

Chen, Y., Zhang, Y., Wang, Y., & Liu, H. (2021). Deep learning in medical image analysis: Challenges and applications in CT imaging. *IEEE Reviews in Biomedical Engineering, 14*, 238–252. https://doi.org/10.1109/RBME.2020.3016819

Charting the path forward: CT image quality assessment – An in-depth review. (2022). *arXiv*. https://arxiv.org/html/2405.00075v1

Dehdab, M., Rezaei, M., Shahbahrami, A., & Nowroozi, M. (2024). Diagnostic performance of ChatGPT-4 Vision in COVID-19 and lung cancer chest CT imaging. *Journal of Medical Artificial Intelligence, 7*(2), 45–56.

Dehdab, R., Brendlin, A., Werner, S., Almansour, H., Gassenmaier, S., Brendel, J. M., Nikolaou, K., & Afat, S. (2024). Evaluating ChatGPT4V in chest CT diagnostics: A critical image interpretation assessment. *Japanese Journal of Radiology, 42*(10), 1168–1177. https://doi.org/10.1007/s11604-024-01606-3

Frost & Sullivan. (2025). *Global medical imaging equipment market: Annual CT scan volume report*. Frost & Sullivan Publications.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley & Sons.

Greffier, J., Frandon, J., Larbi, A., & Beregi, J. P. (2022). Image quality and dose optimization in CT: New concepts and clinical applications. *Diagnostic and Interventional Imaging, 103*(2), 63–72. https://doi.org/10.1016/j.diii.2021.09.002

Hoxha, A., Sella, E., Vukaj, M., & Zanelli, G. (2024). Objective image quality metrics in computed tomography: Advances, applications, and clinical relevance. *European Journal of Radiology, 170*,111123. https://doi.org/10.1016/j.ejrad.2024.111123

Jung, Y., Hur, J., Han, K., Imai, Y., Hong, Y. J., Im, D. J., Lee, K. H., Desnoyers, M., Thomsen, B., Shigemasa, R., Um, K., & Jang, K. (2023). Radiation dose reduction using deep learning- based image reconstruction for a low-dose chest computed tomography protocol: A phantom study. *Quantitative Imaging in Medicine and Surgery, 13*(3), 1937–1947.

Kim, H. J., Lee, S. M., & Kim, N. (2023). Effects of low-dose CT acquisition on image quality and diagnostic performance: A quantitative analysis using CNR. *European Radiology, 33*(4), 2456–2465. https://doi.org/10.1007/s00330-022- 09234-5

Langer, M., Baum, K., & Schlicker, N. (2025). Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. https://doi.org/10.1007/s11023-024-09701-0

Li, X., Zhao, S., Wang, G., & Yu, H. (2023). Artificial intelligence in low-dose CT: Image quality evaluation and diagnostic performance. *European Journal of Radiology, 158*, 110610. https://doi.org/10.1016/j.ejrad.2022.110610

Li, Y., Jiang, Y., Liu, H., Yu, X., Chen, S., Ma, D., Gao, J., & Wu, Y. (2022). A phantom study comparing low-dose CT physical image quality from five different CT scanners. *Quantitative Imaging in Medicine and Surgery, 12*(1), 1–14.

Liew, C. (2023). The future of radiology augmented with artificial intelligence: A strategy for success. *European Journal of Radiology, 158*, 110617. https://doi.org/10.1016/j.ejrad.2022.110617

Lu, Y., Zhang, N., Sun, J., Fang, X., & Li, Y. (2021). Comparison of image quality among A, B and C systems using a CT phantom. *BMC Medical Imaging, 21*, Article 165. https://doi.org/10.1186/s12880-021-00683-4

Nwabuko, Iwu, Njoku, & Nwamoh (2024). An overview of research study designs in quantitative research methodology. *American Journal of Medical and Clinical Research & Reviews, 3*(5), 1–6. https://doi.org/10.58372/2835-6276.1169

O'Connor, S. D., Summers, R. M., & Pickhardt, P. J. (2021). Artificial intelligence in radiology: Performance limitations and clinical implications. *Radiology: Artificial Intelligence, 3*(2), e200242. https://doi.org/10.1148/ryai.2021200242

Park, S. H., Han, K., & Kim, J. (2022). Artificial intelligence in medical imaging: Overestimation and limitations in image quality assessment. *Korean Journal of Radiology, 23*(4), 451–463.

Philippine Statistics Authority. (2024). *Causes of death in the Philippines (January–November 2024)*. https://psa.gov.ph

Romeo, G., & Conti, D. (2025). Exploring automation bias in human AI collaboration: A review and implications for explainable AI. https://doi.org/10.1007/s00146-025-02422-7

Rosen, B., & Saban, M. (2024). Evaluating ChatGPT for imaging referral support: A systematic review of its performance in radiologic decision-making. *Clinical Imaging Informatics, 8*(1), 33–47.

Rosbach, E., Ganz, J., Ammeling, J., Riener, A., & Aubreville, M. (2024). Automation bias in AI-assisted medical decision-making under time pressure in computational pathology. *arXiv*. https://arxiv.org/abs/2411.00998

Singh, S., Kalra, M. K., Do, S., Thibault, J. B., Pien, H., O'Connor, O. J., & Blake, M. A. (2021). Radiation dose reduction and image quality in CT: Current perspectives and future directions. *American Journal of Roentgenology, 216*(2), 266–278. https://doi.org/10.2214/AJR.20.23456

Smith-Bindman, R., Miglioretti, D., Johnson, E., Lee, C., & Lee, J. (2022). Radiation dose variation across 4.5 million CT examinations: Implications for standardization and quality improvement. *The Lancet Digital Health, 4*(3), e153–e162.

Suanes, N. M., Garcia, R. L., Dizon, J. C., & Santos, A. P. (2023). Strengthening lung cancer screening in the Philippines: Policy gaps, challenges, and opportunities for implementation. *Philippine Journal of Oncology, 15*(1), 22–34.

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development*. United Nations Publications.

Wang, Y., Wang, Y., Patel, S., & Patel, D. (2006). A layered reference model of the brain (LRMB). *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, 36*(2), 124–133

9. Appendices

APPENDIX A
Approval Sheet from the Research Ethics Committee

LC-REC Form #039
CERTIFICATE OF EXEMPTION FROM REVIEW

LORMA COLLEGES

CERTIFICATION OF EXEMPTION FROM REVIEW

REC Reference #: 2025-209

To: Jeremiah M. Supnad, Luigie O. Cabigon, Sarbelle Rose F. Costales, Irah Ledz D. Lelina, Miki Ela A. San Luis and Angelica B. Sapitula

From: LORMA Colleges - Research Ethics Committee

Date: January 13, 2026

This is to certify that the Research Proposal entitled, "QUANTITATIVE EVALUATION OF THE DIAGNOSTIC ACCURACY OF CHATGPT IN THE ADULT CHEST COMPUTED TOMOGRAPHY IMAGES: A PHANTOM STUDY" submitted by Jeremiah M. Supnad, Luigie O. Cabigon, Sarbelle Rose F. Costales, Irah Ledz D. Lelina, Miki Ela A. San Luis and Angelica B. Sapitula of College of Radiologic Technology has been reviewed by the Research Ethics Committee of LORMA Colleges and found that all ethical considerations are in place to conduct the research in the stated locale of the study. Hence, this research proposal is exempted from review.

JEROME P. VERA, LPT
Chairman, LC-REC

10. Author(s) Biodata

Mr. Luigie O. Cabigon, Ms. Sarbelle Rose F. Costales, Ms. Irah Ledz D. Lelina, Ms. Miki Ela A. San Luis, Ms. Angelica B. Sapitula, and Mr. Jeremiah M. Supnad are Bachelor of Science in Radiologic Technology students from Lorma Colleges. Together with their research adviser, Mr. Mark Anthony C. Burgonio, MSc, they conducted the study entitled "Quantitative Evaluation of the Diagnostic Accuracy of ChatGPT in the Adult Chest Computed Tomography Images: A Phantom Study." Their research focused on evaluating the diagnostic accuracy of ChatGPT in assessing adult chest CT phantom images using quantitative Contrast-to-Noise Ratio (CNR) evaluation. Through this study, the researchers aimed to contribute to the advancement of artificial intelligence-assisted image quality assessment in computed tomography and radiologic technology practice.