ChatGPT may free time needed by the interventional radiologist for administration / documentation: A study on the RSNA PICC line reporting template.

Jan F. Senge[1,2], Matthew T. McMurray[3], Fabian Haupt[3], Philipe S. Breiding[4], Claus Beisbart[5,6], Keivan Daneshvar[3], Alois Komarek[3], Gerd Nöldge[3], Frank Mosler[3], Wolfram A. Bosbach[3]*

Affiliation:
[1] Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany
[2] Max-Planck Dioscuri Centre for Topological Data Analysis, Warsaw, Poland
[3] Department of Diagnostic, Interventional and Pediatric Radiology (DIPR), Inselspital, Bern University Hospital, University of Bern, Switzerland
[4] University Institute of Diagnostic and Interventional Neuroradiology, Inselspital, Bern University Hospital, University of Bern, Switzerland
[5] Institute of Philosophy, University of Bern, Bern, Switzerland
[6] Center for Artificial Intelligence in Medicine, University of Bern, Bern, Switzerland
*Correspondence: WolframAndreas.Bosbach@Insel.CH
Received: date; Accepted: date; Published: date

Abstract:

Motive: Documentation and administration, unpleasant necessities, take a substantial part of the working time in the subspecialty of interventional radiology. With increasing future demand for clinical radiology predicted, time savings from use of text drafting technologies could be a valuable contribution towards our field.

Method: Three cases of peripherally inserted central catheter (PICC) line insertion were defined for the present study. The current version of ChatGPT was tasked with drafting reports, following the Radiological Society of North America (RSNA) template.

Key results: Score card evaluation by human radiologists indicates that time savings in documentation / administration can be expected without loss of quality from using ChatGPT. Further, automatically generated texts were not assessed to be clearly identifiable as AI-produced.

Conclusions: Patients, doctors, and hospital administrators would welcome a reduction of the time that interventional radiologists need for documentation and administration these days. If AI-tools as tested in the present study are brought into clinical application, questions about trust into those systems eg with regard to medical complications will have to be addressed.

Introduction

In radiology, interventional radiology (IR) is the subspecialty which uses imaging for guiding minimally invasive surgical procedures. Imaging modalities applied today include fluoroscopy, ultrasound (US), computed tomography (CT), or magnetic resonance imaging (MRI) [1]. Our study investigated how IR could benefit from automated text drafting tools. We tested for the template of the Radiological Society of North America (RSNA) for peripherally inserted central catheter (PICC) lines [2] whether reports can be drafted by artificial intelligence (AI) based natural language processing models, ie ChatGPT [3].

Today, the interventional radiologist spends a substantial amount of time on administration / documentation work. As in other medical fields [4], [5], this activity is seen as an unpleasant necessity. It does not serve the immediate patient outcome. At the same time, demand for clinical radiology services is predicted to continue to grow in the future to a level that might not be able to be met by the workforce in its size today [6], [7]. This is why time savings through the use of AI text drafting would be a valuable and welcome contribution to the future of IR, from the viewpoint of patients, doctors, and hospital administrators alike.

Initial steps towards the computing technology required for AI in IR and elsewhere can be found in the work of Konrad Zuse [8]. The development of AI using digital computers was first proposed in eg [9], [10]. As other subfields of AI, natural language processing in combination with reinforcement learning has recently seen some remarkable advances [3], [11]. This is a broader development, not limited to the tool applied in the present study [12]. Regarding the application in radiology and elsewhere, the strengths of properly trained

Keywords: AI, natural language processing, documentation, overview, time, text drafting, machine learning

language processing lie in the huge knowledge base that is made available [13], and in the ability to communicate in different styles of language [14]. So far, rather few studies on AI based language processing have been published in IR. One relevant study has demonstrated limitations concerning accuracy of recommendations for IR procedures [15]. This result is similar to what we found in a previous study about the handling of technical and medical information in report drafting for distal radius fracture [16], [17].
For evaluating the ability of ChatGPT to handle the RSNA PICC line template [2], we defined 3 distinct cases and iterated those for a parameter study (n = 5). Output texts were evaluated for content similarity and rated by 8 human radiologists. The main focus of the study was to determine if automation of text drafting seems feasible and will save time of the interventional radiologist.

Method and Materials

The methodology of the presented study follows the concept of the previous work [16]: cases were defined within the framework of a current RSNA template. ChatGPT was tasked with report drafting. The output texts were evaluated for similarity by comparisons in python. The quality of output texts was assessed by human radiologists using a score card.

RSNA template

The RSNA PICC insertion template can be found in [2]. Template items are listed in Table a. In the present study, three distinct cases were defined varying regarding eg anatomy, clinical information, and occurring complications. The impression had to be generated by the AI tool. “Patient ID”, and “Study ID” were added as parameters for the present study, in addition to the template items contained in [2].

ChatGPT parameter study

The defined cases were given as command file to ChatGPT [3] on 04 May 2023 and iterated (n = 5), producing 15 output cases in total. The command was set to

> “*Write a radiology report which contains this exact information:*”.

No instruction was given on text structure, unlike before in [16]. The returned outputs were saved as txt-files. The previous study on distal radius fracture report drafting [16] relied on an earlier version of ChatGPT [18].

Similarity analysis in Python

An analysis of similarity between text output files was performed following a method used before relying on bag of words in python: cosine similarity [0, 1] of vectors given by key word occurrence in command files defining the indicator vector space [16], [19].

Score card assessment

Table b contains the structure of the score card given to radiologists participating in this study as raters. In total, 5 questions had to be answered for each of the 15 output texts. For this, raters had to grade on an ordinal scale [+2, +1, 0, -1, -2] how much they agree / disagree with the following statements:

1. The report contains all relevant information.
2. I agree with the report’s structure.
3. It is apparent that the text was written by an AI text drafting tool.
4. I would send this text unchanged as report to the referring physician.
5. In this case, the AI tool would have saved me time in my documentation / administration work.

Agreement regarding Questions 1, 2, 4, and 5 expresses a positive view on the ability of ChatGPT. As part of study’s design, Question 3 was deliberately worded to require disagreement from the rater for expressing a positive view on the ability of ChatGPT. Raters were blinded to the results of the other raters.
In total, 8 raters participated, 6 board certified radiologists, 2 residents. The total work experience averaged 22.5 years (min 6, max 49) for the board certified radiologists, with an average of 14.2 years within IR (min 1, max 34). Both residents were in their second year of residency training with 0.5 years in IR.

Interrater agreement and reliability.

For analysing the agreement and reliability between raters, a set of variables was calculated from the score card results, Table d. Each variable took values in the interval [-1, 1]. The approach followed the methodology used before in [16]. Three agreement measures were calculated: exact agreement, one-apart agreement, and weighted agreement with weights for ordinal scales defined in [20]. Chance-corrected

interrater reliability variables for the present study included: Gwet's AC1/AC2 (unweighted/weighted), the Brennan-Prediger coefficient, Conger's kappa (generalization of Cohen's kappa for multiple raters), Fleiss' kappa, and Krippendorff's Alpha. These coefficients can be defined via $1 - \frac{1-P_O}{1-P_e}$, where of P_O and P_e are measures of observed and chance agreement, respectively. The different variables only differ in the definition of P_O and P_e, for detailed formulas see [20]. Imbalance in the occurrences of certain (pairs of) scores in the overall crosstabulation matrix makes traditionally used kappa variables as well as Krippendorff's Alpha prone to low reliability values. This paradoxon is further explained in [20]. Gwet's AC and the Brennan-Prediger coefficient are less influenced by this imbalance effect. Computations were made using the package provided in [21].

Results

Sample output text

Table c contains the output example for case 1, iteration 1, defined in Table a. It can be seen that in principle ChatGPT can draft a PICC line report following the required input from the command file. Throughout the present study, output text structure varied compared to the example of Table c. ChatGPT repeatedly changed the contained section headings. A variation of output text structure was not seen before in [16] where text structure had been an explicit part of the command file.

Text similarity throughout the parameter study

Fig. 1 lists the headings of sections produced by ChatGPT and extracted from the 15 output files. With the exception of "Patient ID" and "Study ID", no section heading appears in all 15 iterations. The average values included for Question 2 (I agree with the report's structure) demonstrate some substantial variation between the 15 cases. Performance was particularly rated as poor whenever no section on complications was included. Within the set of 5 iterations for each of the three cases, score of Question 2 drops / increases whenever the section on complications is omitted / included by ChatGPT.

Fig. 2 provides a similarity comparison on a finer level and shows the cosine similarity calculated using bags of words. The comparison between the command files shows a [3, 3] matrix with the main diagonal taking the max value of 1, comparing the command files with themselves. Pairwise similarity between different command files lies between 0.75 and 0.80.

The comparison between command files and output files is plotted as a [3, 15] grid. As before in [16], similarity exhibits plateaus of grid size [1, 5] along the main diagonal, resulting from comparison between each of the command files to the 5 corresponding output files. Outside these three plateaus, similarity drops substantially. This pattern was seen before in the previous study. It demonstrates again that ChatGPT has the ability to adjust its output to minor changes in the command file. Remarkably, the similarity of the ChatGPT output from one case to a command file of a different case is on average not much lower than the similarity between the respective command files. Accordingly, not much similarity is lost when we move from a command file to the output. While the [1, 5] similarity plateaus were highly homogeneous in the previous study, now on-plateau similarity varies markedly between values from 0.89 to 0.97. This, equally as before the change in text structure in Fig. 1, is new compared to [16] and can be attributed to the omission of prescribed text structure in the command file in the present study.

Output text quality in scorecard assessment

Fig. 3 plots the distribution of the rater responses per question. Fig. 4 shows the average rater response with one standard deviation as error bar. Table d contains in its first panel the mode, median, range, mean and standard deviation.

Overall, raters agreed with the statements offered in Questions 1, 2, 4, and 5; while disagreeing regarding Question 3. This can be interpreted as a clear positive statement about the quality of the AI generated PICC insertion reports.

Strong agreement was the most frequently given answer (strong disagreement in case of question 3 which was deliberately worded to require disagreement for a positive statement about the ChatGPT capabilities). By overall rater opinion, all relevant information was included (Question 1) in an agreeable text structure (Question 2). Raters overall disagreed with the statement that the output texts shown to them had been apparently written by an AI tool; accordingly, raters would not identify them as written by AI rather than by a human radiologist (Question 3). Question 4, whether the text draft could be sent out unchanged, saw a minor drop in mean agreement, as compared to the three previous questions. This indicates that raters would have considered editing the text draft manually before sending it. Question 5 received stronger

agreement again, which affirms that, under the view of the participating human radiologists, AI-based automated text drafting will save time required in IR for administration / documentation.

Essential points raised by raters in their comments concerned text structure and handling of medical complications. Note that complications already influenced results of Question 2 and Fig. 1. Raters missed medical treatment suggestions which should have been included by ChatGPT in the PICC insertion report for the referring doctor by their opinion.

Rater agreement and interrater reliability

An observation already made in [16] is confirmed by this study in Fig. 5 (standard deviation among raters, plotted over absolute rater mean; only negative means obtained under question 3 which required disagreement for a positive statement). Whenever texts are assessed to be of greater quality (abs(mean) $\longrightarrow 2.0$), variation between raters drops (standard deviation $\longrightarrow 0.0$). However, once quality is imperfect (abs (mean) $\longrightarrow 0$), there is an increasing variation between the raters' expression of lack of agreement (standard deviation $\longrightarrow 1.8$). The point scatter in Fig. 5 can be interpolated linearly by regression analysis, $R^2 = 0.830$. On the view of the authors, the pattern in Fig. 5 reflects real life situations in eg case presentations where proportion of disagreement between radiologists may increase with greater need for discussion.

Section 2 of Table d contains the calculated rater agreement. Question 4 which received the lowest absolute mean also shows the lowest agreement between raters for all three agreement variables. This reflects the pattern observed before in Fig. 5 and discussed above. By definition, the agreement variables increase in most cases for wider defined range when calculated per question: exact match < one-apart match < weighted match.

Section 3 of Table d contains the calculated interrater reliability. Fair reliability was calculated for AC1 (unweighted / identity) and AC2 (weighted); as well as for weighted Conger's kappa, Fleiss' kappa, and Krippendorff's Alpha. The remaining measures led to only slight reliability. This range of values is more consistent than what was obtained before in [16] for the evaluation of distal radius fracture reports. Most remarkable is the drop of AC1/2 from (identity: substantial, weighted: almost Perfect) in [16] to (identity: fair, weighted: fair) in the present study. Brennan-Prediger also saw a drop compared to the levels obtained in [16].

The interrater reliability between individual raters is shown as pairwise heatmap in Fig. 6 for weighted AC2. Raters are sorted for decreasing AC2 when calculating it for k raters. It can be seen that pairwise reliability reaches values of up to 0.95. Weighted AC2 decreases to 0.87 for the first four raters. When the remaining four raters of the total 8 are added, AC2 decreases to 0.41. Independently of the rates given, this finding too corresponds to real life experience according to which agreement between individual radiologists might well vary.

Fig. 7 plots the reliability variables for each question. It demonstrates that AC2 and the Brennan-Prediger coefficient (except for Question 4) reached also in the present study greater values than the remaining variables, as before in [16]. The overall drop of AC2 and the Brennan-Prediger coefficient was effectively caused by question 4.

Conclusions and future work

In the present study, we tested ChatGPT [3] for its ability to draft IR reports after PICC line insertion. Reports had to follow the current RSNA template [2] for three predefined study cases (Table a). Evaluation of the report drafts by human radiologists led to an overall positive assessment. One main result is: time savings in clinical administration / documentation of IR procedures can be expected from using ChatGPT (question 5). Future work will have to assess further the expectable magnitude of time savings when compared to today's form of report writing which does typically not use AI generated drafts.

Overall, raters did not identify the output texts as written by an AI tool; this indicates that reports written by AI are for the raters indistinguishable from reports written by human radiologists (Question 3).

Due to the non-deterministic behaviour of ChatGPT, a parameter study was performed for each of the defined study cases (revisions n = 5). Unlike our previous study [16] in which text structure had been part of the input command, no required output text structure was given as part of the command file. As a result, the variation in text structure was stronger than in [16] (see Fig. 1). This drop in text similarity compared to [16] was also seen when calculating cosine similarity (see Fig. 2). Lack of reporting of complications as a separate report section by ChatGPT lowered scores on text structure (see Fig. 1).

In the set of scores received from the raters, a clear pattern could be identified (linear regression, $R^2 = 0.83$) that standard deviation increases for lesser absolute mean, Fig. 5. This pattern reflects real life situations

where proportion of disagreement between radiologists may increase with greater need for discussion. Pairwise analysis of interrater reliability in Fig. 6 showed that also as in real life agreement between individual raters varied (max AC2 0.95, min AC2 0.41).

Mathematics in medical diagnostics is a wide field [22] with potentially many options for optimising healthcare and hospital operations, not limited to automation of clinical documentation [23], [24]. AI tools might well find their way into application and support the interventional radiologist in his administration / documentation tasks. Time savings, as can be expected from the results of the present study, would be an important improvement [4], [5]. Patients, doctors, and hospital administrators would agree on that.

Future work in this field will have to look deeper into ethical issues that may arise due to the application of ChatGPT in IR. One issue is whether professionals (radiologists, nurses etc.) trust AI-written reports. Also, patients may lose trust when they hear that reports are drafted using AI [25]. A second issue is how responsibility is shared between humans and AI [26]: Should humans stay in the loop? And who takes responsibility if something goes wrong? Finally, the privacy of patients is an issue because reinforcement learning uses input data to further train the model. Still, with continuing exposure of users and patients to AI tools and with steady improvements of technology and its ethical use, trust can be expected to grow.

Acknowledgements and funding: The authors wish to thank for all the useful discussions leading to this manuscript.

Declaration of interests: The authors declare no competing financial interests.

Ethics approval: not required

Online supplement: study raw data deposited under doi.org/10.5281/zenodo.8140755

Bibliography

[1] UVA Radiology and Medical Imaging, Ed., “What is interventional radiology?,” *Inside View*. https://blog.radiology.virginia.edu/interventional-radiologist-definition/ (accessed Jun. 17, 2023).

[2] Medical College of Wisconsin, “PICC Insertion,” *RSNA RadReport*, 2012. https://radreport.org/home/188/2012-05-29 00:00:00 (accessed May 03, 2023).

[3] OpenAI LLC, Ed., “ChatGPT — Release Notes (May 3).” https://help.openai.com/en/articles/6825453-chatgpt-release-notes (accessed May 04, 2023).

[4] S. Woolhandler and D. U. Himmelstein, “Administrative work consumes one-sixth of u.s. physicians’ working hours and lowers their career satisfaction,” *Int. J. Heal. Serv.*, vol. 44, no. 4, pp. 635–642, 2014, doi: 10.2190/HS.44.4.a.

[5] S. M. Erickson, B. Rockwern, M. Koltov, R. M. Mclean, and M. Practice, “Putting Patients First by Reducing Administrative Tasks in Health Care: A Position Paper of the American College of Physicians Putting Patients First by Reducing Administrative Tasks in Health Care: A Position Paper of the American College of Physicians,” *Ann. Intern. Med.*, vol. 166, no. 9, pp. 659–661, 2017, doi: 10.7326/M16-2697.

[6] M. Henderson, “Radiology Facing a Global Shortage,” *RSNA News*, 2023. https://www.rsna.org/news/2022/may/global-radiologist-shortage (accessed May 16, 2023).

[7] G. Sutherland, N. Russell, R. Gibbard, and A. Dobrescu, *The Value of Radiology , Part II - The Conference Board of Canada*, no. June. Ottawa, CAN, 2019.

[8] K. Zuse, “Aus mechanischen Schaltgliedern aufgebautes Speicherwerk,” DE924107, 1937

[9] A. M. Turing, “I.-Computing machinery and intelligence,” *Mind - A Q. Rev. Psychol. Philos.*, vol. 236, pp. 433–460, 1950.

[10] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, “A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence,” 1955. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf (accessed Oct. 30, 2021).

[11] D. Glowacka, A. Howes, J. P. Jokinen, A. Oulasvirta, and Ö. Azimsek, “RL4HCI: Reinforcement Learning for Humans, Computers, and Interaction,” *Ext. Abstr. 2021 CHI Conf. Hum. Factors Comput. Syst.*, pp. 1–3, 2021, doi: 10.1145/3411763.3441323.

[12] R. Doshi, K. Amin, P. Khosla, S. Bajaj, S. Chheang, and H. P. Forman, “Utilizing Large Language Models to Simplify Radiology Reports : a comparative analysis,” *medRxiv Prepr.*, 2023, doi: 10.1101/2023.06.04.23290786.

[13] R. Bhayana, F. S. Krishna, and R. R. Bleakney, "Performance of ChatGPT on a Radiology Board-style Examination : Insights into Current Strengths and Limitations," *Radiology*, vol. 307, no. 5, p. e230582, 2023.

[14] Q. Lyu, J. Tan, M. E. Zapadka, J. Ponnatapura, C. Niu, K. J. Myers, G. Wang, and C. T. Whitlow, "Translating Radiology Reports into Plain Language using ChatGPT and GPT-4 with Prompt Learning: Promising Results, Limitations, and Potential," *Vis. Comput. Ind. Biomed. Art*, vol. 6, no. 9, pp. 1–10, 2023, doi: 10.1186/s42492-023-00136-5.

[15] M. Barat, P. Soyer, and A. Dohan, "Appropriateness of Recommendations Provided by ChatGPT to Interventional Radiologists," *Can. Assoc. Radiol. J.*, pp. 1–6, 2023, doi: 10.1177/08465371231170133.

[16] W. A. Bosbach, J. F. Senge, B. Nemeth, S. H. Omar, M. Mitrakovic, C. Beisbart, A. Horvath, J. T. Heverhagen, and K. Daneshvar, "Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier," *Curr. Probl. Diagn. Radiol.*, 2023, [Online]. Available: https://doi.org/10.1067/j.cpradiol.2023.04.001

[17] W. A. Bosbach, J. F. Senge, B. Nemeth, S. H. Omar, M. Mitrakovic, C. Beisbart, A. Horvath, J. T. Heverhagen, and K. Daneshvar, "Online supplement to manuscript: 'Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier.' Current problems in diagnostic radiology (2023).," *zenodo*, 2023, doi: 10.5281/zenodo.7908791.

[18] OpenAI LLC, Ed., "ChatGPT — Release Notes (Jan 9)." https://help.openai.com/en/articles/6825453-chatgpt-release-notes (accessed Jan. 11, 2023).

[19] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, "More than bags of words: Sentiment analysis with word embeddings," *Commun. Methods Meas.*, vol. 12, no. 2–3, pp. 140–157, 2018.

[20] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD (USA): Advanced Analytics, LLC, 2014.

[21] K. Gwet and A. Fergadis, "irrCAC - Chance-corrected Agreement Coefficients," 2023. https://irrcac.readthedocs.io/en/latest/index.html# (accessed Mar. 05, 2023).

[22] W. A. Bosbach, J. F. Senge, and P. Dlotko, Eds., "2022 Proceedings of the 4th International Conference on Trauma Surgery Technology: Mathematics in medical diagnostics," 2022, pp. 1–36. doi: 10.5281/zenodo.7191419.

[23] W. A. Bosbach, M. Heinrich, R. Kolisch, and C. Heiss, "Maximization of Open Hospital Capacity under Shortage of SARS-CoV-2 Vaccines-An Open Access, Stochastic Simulation Tool," *Vaccines*, vol. 9, no. 6, p. 546, 2021, doi: 10.3390/vaccines9060546.

[24] W. A. Bosbach, "Open-access supplement: Maximisation of open hospital capacity under shortage of SARS-CoV-2 vaccines," *zenodo*, 2021, doi: 10.5281/zenodo.4589333.

[25] J. J. Hatherley, "Limits of trust in medical AI," *J. Med. Ethics*, vol. 46, no. 7, pp. 478–481, 2020, doi: 10.1136/medethics-2019-105935.

[26] M. Verdicchio and A. Perin, "When Doctors and AI Interact: on Human Responsibility for Artificial Risks," *Philos. Technol.*, vol. 35, no. 11, pp. 1–28, 2022, doi: 10.1007/s13347-022-00506-6.

Table a: cases for defined parameter study.

	RSNA template items [2]	**Case 1**	**Case 2**	**Case 3**
Patient (additional study parameter)		Patient ID KEHW7830 Study ID 2379430	Patient ID OMSW2397247 Study ID 395370	Patient ID HBET29475 Study ID 19482047
Procedure		PICC insertion	PICC insertion	PICC insertion
Technique		Seldinger US and fluoroscopy guidance	Seldinger US and fluoroscopy guidance	Seldinger Venography and fluoroscopy guidance
Site	Right arm Left arm	Right arm	Left arm	Left arm
	Basilic vein Brachial vein Cephalic vein	Brachial vein	Basilic vein	Cephalic vein
Catheter	Single-lumen Double-lumen Triple-lumen	Triple-lumen	Single-lumen	Double -lumen
PICC placement	Peripherally placed PICC line. The arm was prepped and draped in sterile fashion. Lidocaine 1% was used for local anesthetic. Under fluoroscopic and ultrasound guidance, the vein was patent and accessed with a micropuncture needle. A guide wire was then advanced into the vein. A vascular sheath was then advanced over a guide wire, and a PICC line was trimmed. The PICC line was then advanced into the central venous system. After confirmation of the catheter position, the catheter was sutured in place at the skin entry site.	Peripherally placed PICC line. The arm was prepped and draped in sterile fashion. Lidocaine 1% was used for local anesthetic. Under fluoroscopic and ultrasound guidance, the vein was patent and accessed with a micropuncture needle. A guide wire was then advanced into the vein. A vascular sheath was then advanced over a guide wire, and a PICC line was trimmed. The PICC line was then advanced into the central venous system. After confirmation of the catheter position, the catheter was sutured in place at the skin entry site.	Peripherally placed PICC line. The arm was prepped and draped in sterile fashion. Lidocaine 1% was used for local anesthetic. Under fluoroscopic and ultrasound guidance, the vein was patent and accessed with a micropuncture needle. A guide wire was then advanced into the vein. A vascular sheath was then advanced over a guide wire, and a PICC line was trimmed. The PICC line was then advanced into the central venous system. After confirmation of the catheter position, the catheter was sutured in place at the skin entry site.	Peripherally placed PICC line. The arm was prepped and draped in sterile fashion. Lidocaine 1% was used for local anesthetic. Under fluoroscopic and ultrasound guidance, the vein was patent and accessed with a micropuncture needle. A guide wire was then advanced into the vein. A vascular sheath was then advanced over a guide wire, and a PICC line was trimmed. The PICC line was then advanced into the central venous system. After confirmation of the catheter position, the catheter was sutured in place at the skin entry site.
Clinical information		68 years, male, indication for chemotherapy	82 years, female, coagulopathy preventing more permanent access	75 years, female, total parenteral nutrition
Comparison	None	Chest x-ray from 22 Nov 2022	Chest x-ray from 14 Sept 2021	Chest x-ray from 24 Aug 2022
Findings	Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium.	Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium.	Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium.	Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium.
	Complications: No immediate complication.	Complications: No immediate complication.	Complications: Arterial puncture with bleeding and haematoma	Complications: air embolus 2 ml
Fluoroscopy time:	[min]	Fluoroscopy time: 2:36 min	Fluoroscopy time: 4:16 min	Fluoroscopy time: 3:23 min
Impression	AI auto generated			

Table b: radiology scoring card.

Case / iteration	**Question 1:** The report contains all relevant information.	**Question 2:** I agree with the report's structure.	**Questions 3:** It is apparent that the text was written by an AI text drafting tool.	**Question 4:** I would send this text unchanged as report to the referring physician.	**Question 5:** In this case, the AI tool would have saved me time in my documentation / administration work.
n = 1					
n = 2					
n = ..					

strongly agree	2
agree	1
undetermined	0
disagree	-1
strongly disagree	-2

Table c: output example case 1, iteration 1, generated by ChatGPT [3] *for values defined in Table a for the RSNA PICC line template* [2].

ChatGPT output text, case 1, iteration 1:

Radiology Report:

Patient ID: KEHW7830
Study ID: 2379430

Procedure: Peripherally inserted central catheter (PICC) insertion under ultrasound and fluoroscopy guidance

Indication: Chemotherapy

Technique: The right arm was prepped and draped in sterile fashion. Lidocaine 1% was used for local anesthesia. Under fluoroscopic and ultrasound guidance, the brachial vein was accessed with a micropuncture needle. A guide wire was advanced into the vein, followed by a vascular sheath. A triple-lumen PICC line was trimmed and advanced into the central venous system. The catheter position was confirmed, and the catheter was sutured in place at the skin entry site.

Findings: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium.

Complications: No immediate complications were noted.

Fluoroscopy time: 2 minutes and 36 seconds.

Impression: Peripherally placed triple-lumen PICC line successfully inserted in the right arm brachial vein under fluoroscopic and ultrasound guidance. The catheter tip is positioned at the junction of SVC and right atrium. No immediate complications were noted.

Signed,
[Radiologist Name]

Table d: simple statistics of score card results, rater agreement, and interrater reliability.

1. simple statistics of score card results

Question	Case	mode	median	range	mean	stdev
1	1	2	2	1	1.8	0.40
	2	2	2	4	1.43	0.92
	3	2	2	4	1.35	1.06
2	1	2	2	4	0.90	1.55
	2	2	2	4	1.23	1.23
	3	2	2	4	1.63	1.07
3	1	-2	-2	3	-1.48	0.84
	2	-2	-1	4	-1.13	1.21
	3	-2	-1	4	-0.95	1.30
4	1	1, 2	1	4	0.40	1.58
	2	2	1	4	0.53	1.57
	3	2	2	4	0.60	1.66
5	1	2	2	4	1.45	1.09
	2	2	2	4	1.15	1.19
	3	2	2	4	0.98	1.33

2. rater agreement in score card results

	match	Question 1	Question 2	Question 3	Question 4	Question 5
	exact match	0.49	0.49	0.35	0.25	0.39
	one-apart match	0.88	0.73	0.74	0.49	0.64
	weighted match	0.87	0.78	0.81	0.62	0.77

3. interrater reliability in score card results over all questions and cases

Coefficient name	value	weights	P_o	P_e	confidence interval	Benchmark: Landis-Koch
AC1 (identity)AC2 (weighted)	0.27	identity	0.39	0.17	0.21 – 0.32	Fair
	0.41	weighted	0.77	0.61	0.30 – 0.52	Fair
Brennan-Prediger	0.19	weighted	0.77	0.72	0.08 – 0.30	Slight
	0.24	identity	0.39	0.20	0.19 – 0.29	Slight
Conger's kappa	0.33	weighted	0.77	0.66	0.23 – 0.44	Fair
	0.13	identity	0.39	0.30	0.08 – 0.17	Slight
Fleiss' kappa	0.33	weighted	0.77	0.66	0.22 – 0.43	Fair
	0.11	identity	0.39	0.32	0.06 – 0.16	Slight
Krippendorff's Alpha	0.33	weighted	0.77	0.66	0.22 – 0.44	Fair
	0.11	identity	0.39	0.32	0.06 – 0.16	Slight

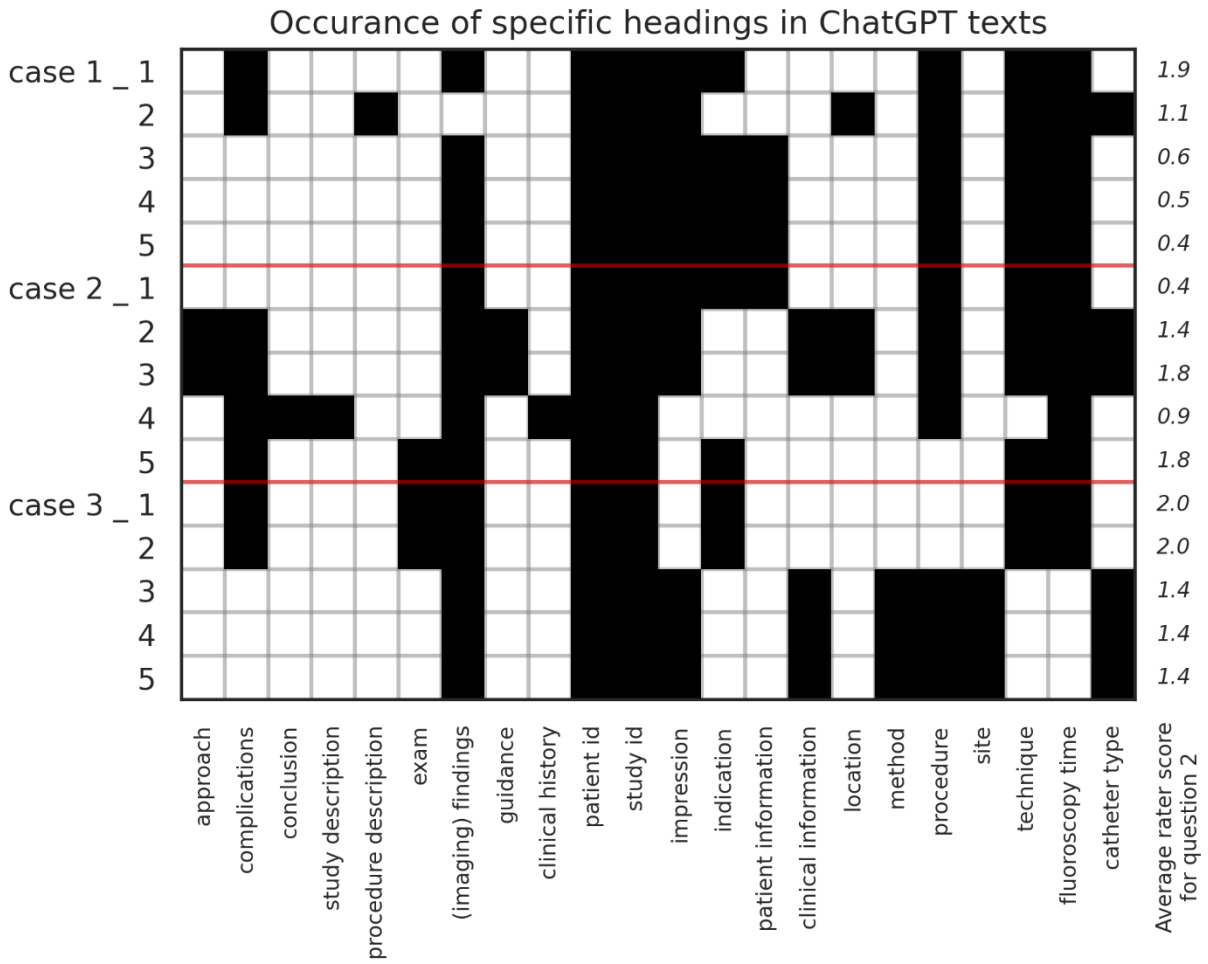

Fig. 1: section headings extracted from the 15 output files, sorted alphabetically by second word in heading, together with average value from raters for Question 2: I agree with the report's structure.

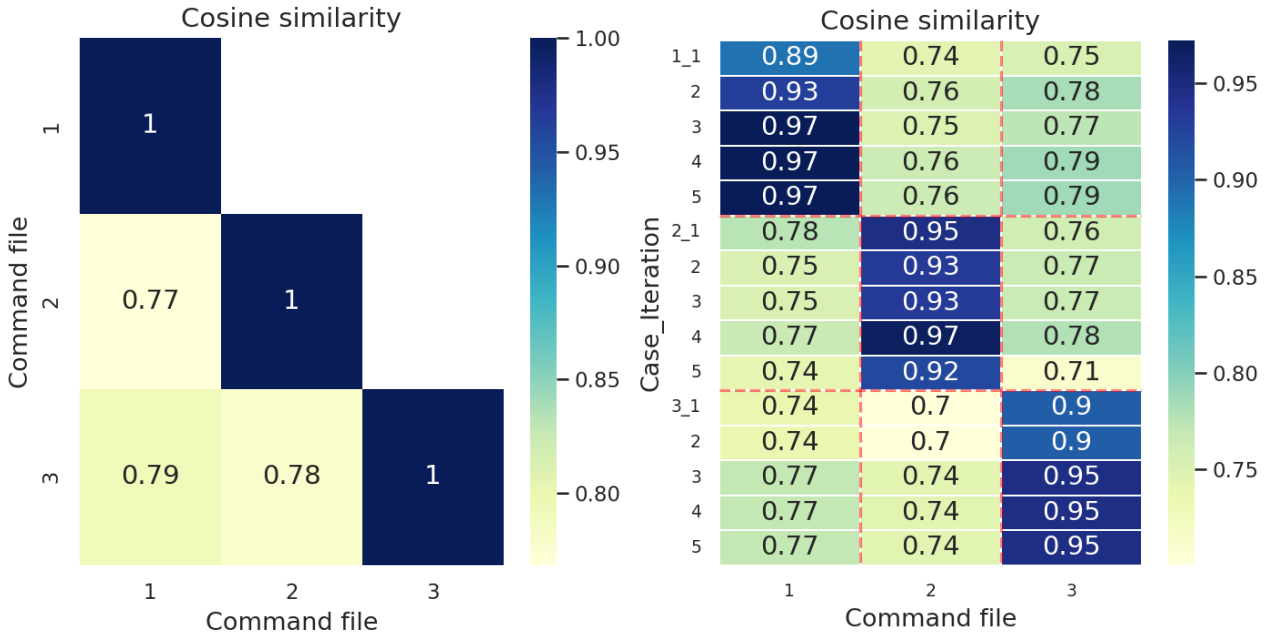

Fig. 2: cosine similarity matrix between command files, and between command files and output files, computed by bag of words in Python.

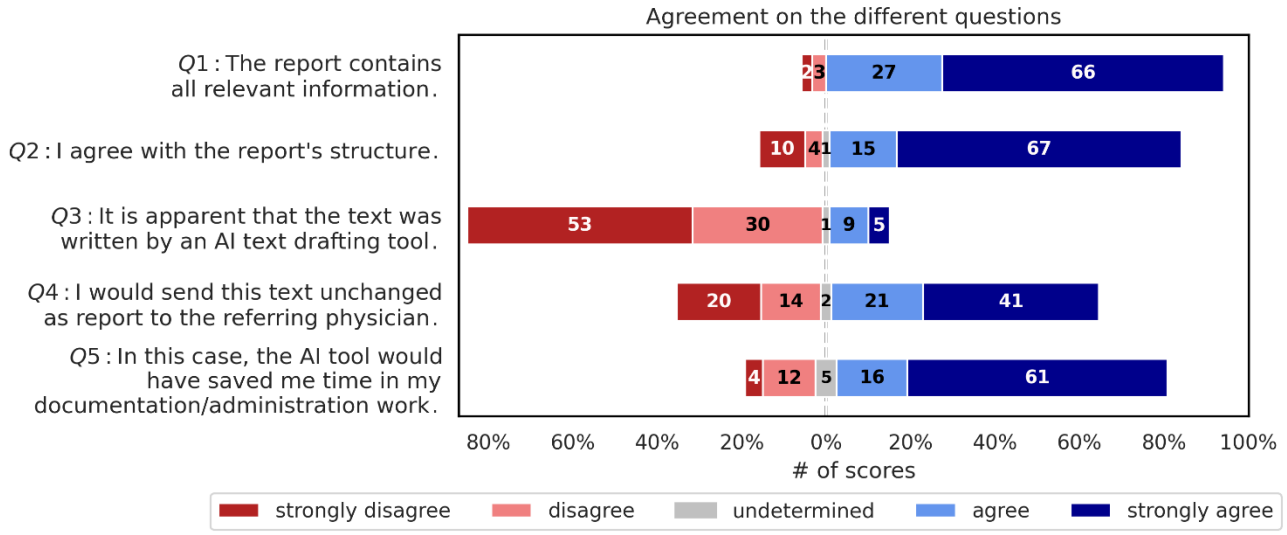

Fig. 3: score card assessment with distribution of dis / agreement by raters per question.

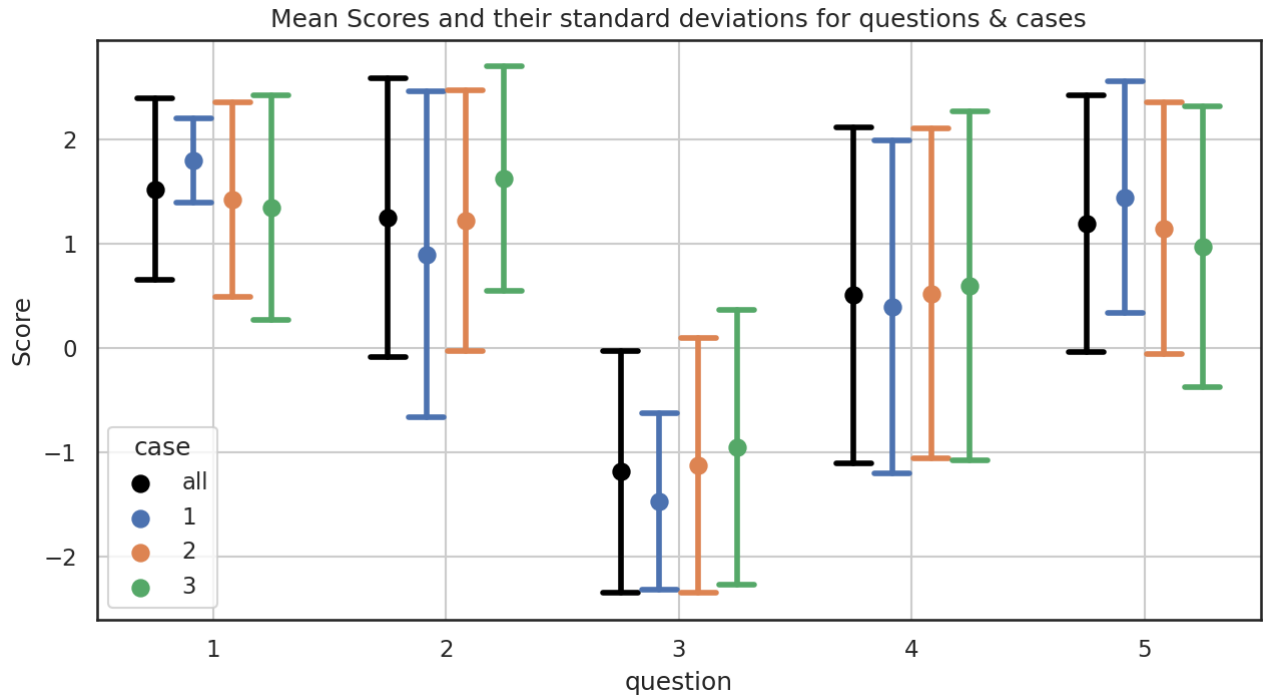

Fig. 4: mean score with error bar of ± 1 standard deviation.

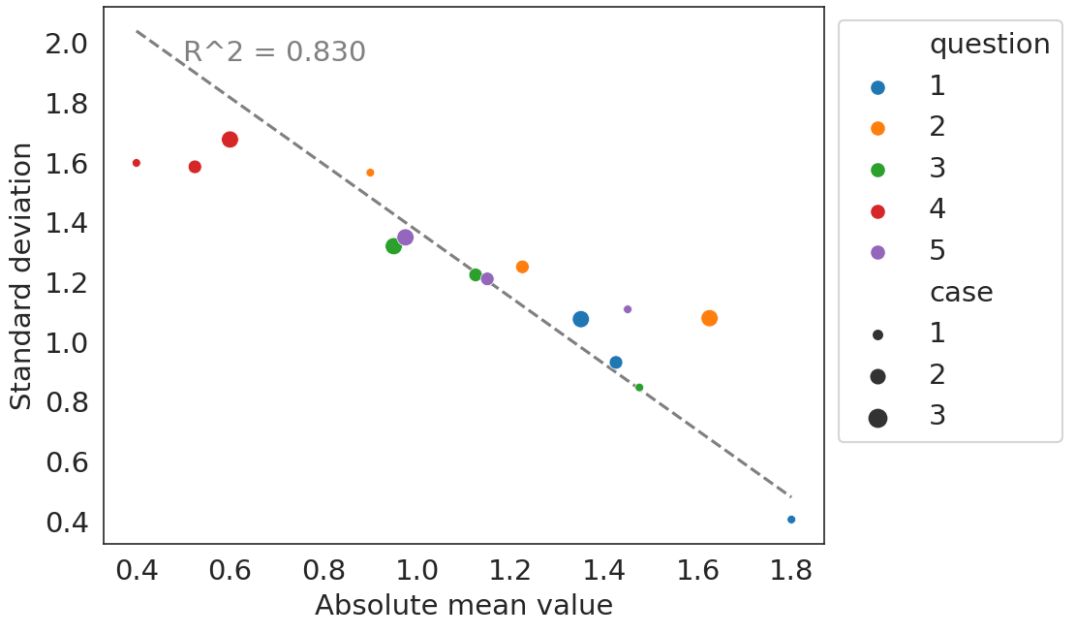

Fig. 5: standard deviation plotted over absolute mean, aggregated per question per case, 15 data points.

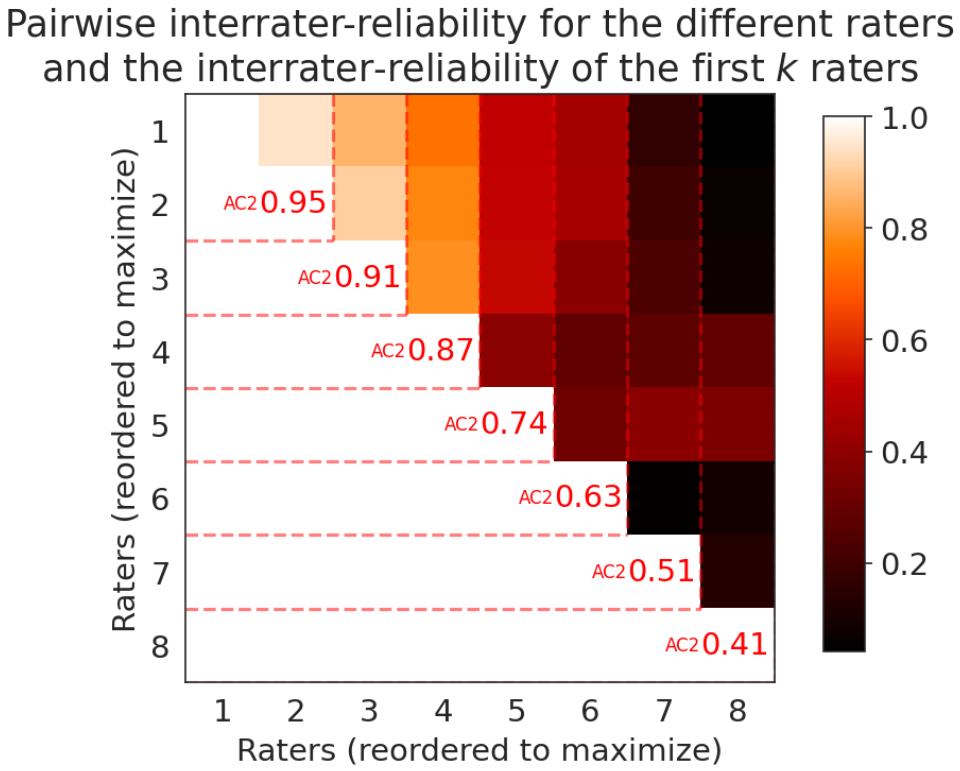

Fig. 6: pairwise interrater reliability as heat map, as well was interrater reliability for group of the first k raters (red). The raters are sorted for descending magnitude of Gwet's AC2 for greater group of raters.

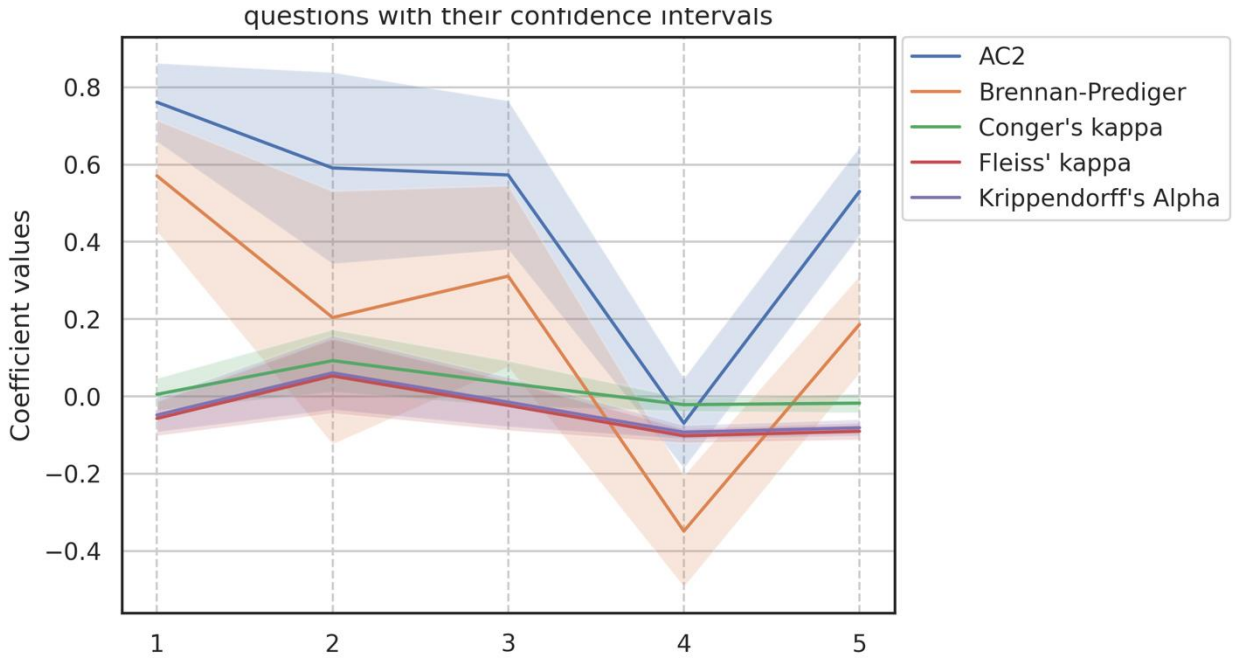

Fig. 7: weighted Interrater reliability variables per question.