Effect of AI-assisted software on inter- and intra-observer variability for preschool children X-ray bone age assessment

Kai Zhao
Peking University First Hospital
Shuai Ma
Peking University First Hospital
Zhaonan Sun
Peking University First Hospital
Xiang Liu
Peking University First Hospital
Ying Zhu
Peking University First Hospital
Yufeng Xu
Peking University First Hospital
Xiaoying Wang (wangxiaoying@bjmu.edu.cn)
Peking University First Hospital

Research Article

Keywords: Bone age, pediatric, radiographs, artificial intelligence, variability

Posted Date: June 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1737407/v1

Abstract

Purpose

The purpose of this study was to investigate the effect of AI-assisted software on residents' inter-observer agreement and intra-observer reproducibility for preschool children X-ray bone age assessment.

Methods

This prospective study was approved by Institutional Ethics Committee. Six board-certified residents interpreted 56 bone age radiographs ranging from 3 to 6 years with structured reporting by modified TW3 method. The images were interpreted on two separate occasions, once with and once without the assistant of AI. After a washout period of 4 weeks, the radiographs were reevaluated by each resident in the same way. The average bone age results of three experts were the reference bone age. Both TW3-RUS and TW3-Carpal were evaluated. The root mean squared error (RMSE), mean absolute difference (MAD) and bone age accuracy within 0.5 years & 1 year were used as metrics of accuracy. Inter-observer agreement and intra-observer reproducibility were evaluated using intraclass correlation coefficient (ICCs).

Results

With the assistance of bone age AI software, the accuracy of residents' results improved significantly. For Inter-observer agreement comparison, The ICC results with AI assistance among 6 residents were higher than results without AI assistance on the two separate occasions. For intra-observer reproducibility comparison, the ICC results with AI assistance were higher than results without AI assistance between the 1st reading and 2nd reading for each resident.

Conclusions

For preschool children X-ray bone age assessment, besides improving diagnostic accuracy, bone age AI-assisted software can also increase inter-observer agreement and intra-observer reproducibility. AI-assisted software can be an effective diagnostic tool for residents in actual clinical settings.

What Is Known

1. X-ray bone age assessment in children and adolescents is a very important tool for pediatricians in the diagnosis of endocrine and metabolic diseases.
2. AI-assisted diagnosis software has begun to be used in some hospitals in daily practice, but the impact of software on resident X-ray interpretation, especially the influence of consistency, is still

unknown.

What Is New

1. This study found with the assistance of AI software, diagnostic accuracy was significantly improved for preschool children X-ray bone age assessment.
2. Bone age AI-assisted software could also increase inter-observer agreement and intra-observer reproducibility.
3. AI-assisted software can be an effective diagnostic tool for residents in actual clinical settings.

Introduction

X-ray bone age assessment (BAA) in children and adolescents is a very important tool for pediatricians in the diagnosis of endocrine and metabolic diseases about growth and development[1]. The Greulich-Pyle (GP) and the Tanner-Whitehouse 3 (TW3) methods are the most widely used diagnostic criteria[2]. The GP method is an atlas-based method that determines bone age by comparing examiner's radiographs of the hands and wrists with the most similar standard radiographs in the GP atlas, which is a simple and easy method that can be used in clinical practice. Different from the GP method, the TW3 method, which has been modified twice, is a scoring system that measures individual bone maturity by scoring and summing multiple bones such as metacarpal, phalanx, and carpal bones, which is a quantitative method. It is more accurate than the GP method but more time consuming[2, 3].

With the rapid development of deep learning algorithms and the rapid improvement of computer hardware in the past few years, artificial intelligence AI-assisted diagnosis software has begun to be applied in the hospitals, among which the bone age AI-assisted software is one of the earliest[4–9]. AI-assisted diagnosis software for bone age has achieved good diagnostic performance[9–16]. Multiple studies have proven that the results of AI-assisted diagnosis software for bone age are as accurate as those of expert[11, 12, 14]. While many studies have shown that AI assistance improves diagnostic accuracy rate of radiologists[9, 14, 16, 17]. However, little research has been done with the impact of AI-assisted software on inter-observer (variation among different observers) and intra-observer (variation within individual observers) variability of residents in clinical practice.

The purpose of this study was to investigate the effect of AI-assisted software on residents' inter-observer agreement and intra-observer reproducibility for preschool children X-ray bone age assessment.

Methods

Patients

This study was approved by the Institutional Review Board and Ethics Committee. Informed consent was waived because the X-rays of the wrist were collected retrospectively and analyzed anonymously.

Stratified random sampling by age and gender was performed from the children with physiological age of 3-6 years old from a 12-month period in the picture archiving and communication system (PACS) system. For the 3 years old preschool children, 14 cases including 7 male and 7 female were included in the reading database, which was the same for 4-6 years old preschool children. None of the cases in this reading database participated in the training and verification of the AI software.

AI-assisted software for bone age assessment

The bone age AI-assisted diagnosis software used in the study was provided by SHENRUI Software Company. The development of the software follows the modified TW3 standard (modified for the Chinese people). The software is based on X-ray image preprocessing, deep learning network for detecting and grading wrist epiphysis to realize automatic identification and bone age assessment.

Image Interpretation

A total of 6 board-certified residents were trained with the modified TW3 standard before bone X-ray interpretation. The residents completed the images interpretation twice, with 4 weeks washout period between two interpretations. In order to reduce the influence of errors, for each interpretation, a random cross-reading method was used. The images in the database were randomly divided into two parts, one part was interpreted with AI assistance, and the other part was interpreted without AI assistance, with 2 weeks washout period between two parts (Figure 1).

Reference Bone Age

The reference bone age was determined by three pediatric radiologists with 12, 18, 23 years of clinical experience who are familiar with bone age assessment based on X-ray radiographs. The average of the independent results of the three experts was used as the gold standard for this study. In case of a discrepancy over 2 years, the image would be discussed together until a consensus was reached.

Statistical Analyses

Statistical analysis was performed by using the SPSS v19 (SPSS Inc., Chicago, Illinois, USA). For comparison of the accuracy of bone age between “without AI” and “with AI”, the root mean squared error (RMSE), mean absolute difference (MAD) and bone age accuracy within 0.5 years & 1 year of the 1st interpretation were used as metrics.

Inter-observer agreement. For the 1st interpretation, the intraclass correlation coefficients (ICCs) with 95% confidence intervals for the 6 residents (resident 1–6) were compared between results with and without AI. For the 2nd interpretation, the ICCs for the 6 residents (resident 1–6) were also compared by the same way. An ICC value greater than 0.75 is excellent, from 0.75 to 0.60 is good, from 0.59 to 0.40 is fair and below 0.40 is poor agreement.

Intra-observer reproducibility. Intra-observer agreement comparing the results of the same resident's interpretations at two different times for all of the residents was determined via intraclass correlation coefficient (ICC) with 95% confidence intervals.

Results

Characteristics of the bone age X-ray database

Among the 56 cases, 2 cases were excluded due to unqualified image quality and 54 images were enrolled in the final database. 3 cases in the " TW3-RUS" were excluded because the gold standard exceeded the lower limit of the modified TW3 standard. 51 cases in the " TW3-RUS" were finally included in the final analysis. 8 cases in the " TW3-Carpal" were excluded because the gold standard exceeded the lower limit of the modified TW3 standard. 46 cases in the " TW3-Carpal" were finally included in the final analysis. Distribution of Sex and Age for all cases is presented in Table 1.

Table 1
Sex and age distribution

	Age	Male	female	Total
TW3-RUS	3 years	5	5	10
	4 years	7	6	13
	5 years	7	7	14
	6 years	7	7	14
	Total	26	25	51
TW3-Carpal	3 years	7	1	8
	4 years	7	4	11
	5 years	7	6	13
	6 years	7	7	14
	Total	28	18	46

Model Accuracy In Baa

With the assistance of bone age AI software, the accuracy of residents' results improved significantly. The average RMSE of TW3-RUS decreased from 0.806 years to 0.501 years, while the average MAD decreased from 0.608 years to 0.379 years. The accuracy increased from 56.4–69.6% within 0.5 years. The accuracy increased from 77.6–91.3% within 1 years. TW3-RUS interpretation accuracy is presented in Table 2. The average RMSE of TW3-Carpal decreased from 0.508 years to 0.323 years, and the average

MAD decreased from 0.355 years to 0.229 years. The accuracy increased from 67.4–82.6% within 0.5 years. The accuracy increased from 93.5–100% within 1 years. TW3-Carpal interpretation accuracy is presented in Table 3.

Table 2
TW3-RUS interpretation accuracy in the 1st interpretation

	average RMSE	average MAD	accuracy within 0.5 year	accuracy within 1 year
without AI	0.806	0.608	56.4%	77.6%
with AI	0.501	0.379	69.6%	91.3%
Elevated value	0.305	0.229	13.10%	13.8%

Table 3
TW3-Carpal interpretation accuracy in the 1st interpretation

	average RMSE	average MAD	accuracy within 0.5 year	accuracy within 1 year
without AI	0.508	0.355	67.4%	93.5%
with AI	0.323	0.229	82.6%	100%
Elevated value	0.186	0.126	15.2%	6.5%

Comparison Of Inter-observer Agreement

The results of inter-observer agreement for diagnostic consistency are presented in Table 4. For the inter-observer agreement comparison of TW3-RUS, the ICC results among 6 residents were elevated from 0.833 to 0.977 with the assist of AI in 1st interpretation while from 0.833 to 0.977 in 2nd interpretation. For the inter-observer agreement comparison of TW3-Carpal, the ICC results among 6 residents were elevated from 0.902 to 0.977 with the assist of AI in 1st interpretation while from 0.896 to 0.948 in 2nd interpretation.

Table 4
Inter-observer agreement of residents

	TW3-RUS		TW3-Carpal	
	without AI	with AI	without AI	with AI
	ICC (%95 CI (min–max))	ICC (%95 CI (min–max))	ICC (%95 CI (min–max))	ICC (%95 CI (min–max))
1st interpretation	0.833 (0.767–0.890)	0.977 (0.965–0.985)	0.902 (0.851–0.942)	0.977 (0.963–0.987)
2nd interpretation	0.897 (0.828–0.921)	0.975 (0.963–0.984)	0.896 (0.842–0.938)	0.948 (0.920–0.970)

Comparison Of Intra-observer Reproducibility

The results of intra-observer reproducibility are presented in Table 5. For intra-observer reproducibility of TW3-RUS between the 1st reading and 2nd reading, the ICC results with AI assistance were higher than results without AI assistance for each resident. The results were similar for TW3-Carpal.

Table 5
Intra-observer reproducibility of residents

	TW3-RUS		TW3-Carpal	
	without AI between 1st and 2nd interpretation	with AI between 1st and 2nd interpretation	without AI between 1st and 2nd interpretation	with AI between 1st and 2nd interpretation
	ICC (%95 CI (min–max))	ICC (%95 CI (min–max))	ICC (%95 CI (min–max))	ICC (%95 CI (min–max))
Resident 1	0.793 (0.663–0.876)	0.986 (0.976–0.992)	0.888 (0.793–0.941)	0.976 (0.955–0.988)
Resident 2	0.870 (0.783–0.924)	0.971 (0.950–0.984)	0.930 (0.869–0.964)	0.975 (0.953–0.987)
Resident 3	0.898 (0.828–0.941)	0.959 (0.930–0.977)	0.860 (0.744–0.925)	0.891 (0.799–0.943)
Resident 4	0.857 (0.762–0.916)	0.986 (0.975–0.992)	0.906 (0.826–0.951)	0.977 (0.956–0.988)
Resident 5	0.951 (0.916–0.972)	0.991 (0.985–0.995)	0.936 (0.880–0.967)	0.969 (0.940–0.984)
Resident 6	0.802 (0.678–0.882)	0.976 (0.958–0.986)	0.922 (0.853–0.959)	0.967 (0.936–0.983)

Discussion

In this study, changes of diagnostic accuracy, inter-observer agreement and Intra-observer reproducibility between with and without AI assistance were investigated. The results showed that AI-assisted software can eliminating both inter- and intra-rater variability. Furthermore, with the assistance of bone age AI software, the diagnostic accuracy of bone age assessment can be improved for less experienced radiologists.

With the use of AI and machine learning, especially the most known machine learning method deep learning, new possibilities for automated BAA have emerged[5–7]. The most popular deep learning is convolutional neural networks (CNNs), which has tremendous progress in recent years, and there are numerous publications about the use of CNNs in BAA[4, 8–11, 14, 15]. Radiological Society of North America (RSNA) launched a BAA challenge in 2017 and many machine learning methods achieved good results[18, 19]. The AI tool used in this study is based on CNNs method.

The emergence of fully automatic AI software help us overcome complexity and time consumption in the interpretation process. Most publications discuss the data between AI and radiologists with convincible good results about improved accuracy or reduced complexity and time. But it is not yet the reality to send the AI results directly to the pediatrician without confirmation of radiologist. In clinical practice, the purpose of AI-assisted software is to assist the radiologist but not to use it independently. Only by validating the results of AI-assisted software in in daily routine can it truly prove its value. So two images interpretation scenarios “without AI” and ”with AI” were included in our research. Our results demonstrated that with the assistance of AI, accuracy of residents’ results improved significantly, which were same as most similar publications.

One of the challenges in BAA is the variability in radiologist clinical interpretation of bone age radiographs, both for inter- and intra- observer. Will automated bone age tools eliminate enhance inter-observer diagnostic consistency or intra-observer diagnostic reproducibility? There is only a few papers focused on it. A study by Tajmir et al[16] revealed that AI BAA improved the radiologist performance while decreased the variation (ICC without AI was 0.9914, with AI was 0.9951). Only three radiologists participated in image interpretation. Lee et al[12] developed a deep learning-based hybrid (GP and modified TW) method for BAA and the ICC of the two radiologists slightly increased with AI model assistance (from 0.945 to 0.990). In another study by Koc et al[20], the ICC were 0.980 for with AI and 0.980 with AI (BoneXpert). The inter-observer variability was not eliminate in their research. Our study demonstrated that, AI bone age tools can eliminate both inter-observer variability and intra- observer variability. 6 observers were analyzed and the intra- observer variability was also compared.

It is well known that the GP and TW methods are most commonly used clinical approaches for BAA. GP is the most popular method among pediatricians and radiologists, as BAA by GP is relatively quick and easy to learn. But GP method itself has significant inter-observer and intra-observer variability[21]. The TW method is considered to be more accurate and objective than the GP method and lower variability than GP[22, 23]. So we chose the TW method and the TW-based AI software in our study. Skeletal maturity varies by ethnicity, geographic location, and socioeconomic status. Caucasian reference standards cannot be expected to be used for comparison in China. So a modified TW3 standard modified for Chinese people was applied in our research. The bone age reference standards modified for Chinese was approved by the national official standards certification center. The AI software used in the research was also designed for Chinese by modified TW3 standard.

TW3-Carpal was less evaluated than TW3-RUS as the epiphysis of ulna and all carpal bones are less reliable as indicators of bone age for female from 2 to 7 years old and males from 3 to 9 years old. But the inter- and intra- observer variability can be evaluated as there is evaluation criterion for TW3-RUS. Thus, we designed our study to investigate the inter- and intra- observer variability for TW3-RUS and also for TW3-Carpal. And the study population was preschool children between 3 years to 6 years.

There were several limitations in this study. First, this was a single-center study with a small and single-ethnicity sample size, and only preschool children were enrolled. In the future, prospective multicenter

studies with more cases will be performed. Second, the interpretation time was not recorded. Time consumption should be compared though many research already demonstrated that AI-assisted software can obviously reduce the diagnostic time[9, 10, 14].

For preschool children X-ray bone age assessment, besides improving diagnostic accuracy, bone age AI-assisted software can also increase inter-observer agreement and intra-observer reproducibility. AI-assisted software can be an effective diagnostic tool for residents during BAA.

Abbreviations

AI Artificial Intelligence

BAA Bone age assessment

GP Greulich-Pyle

ICC Intraclass correlation coefficient

MAD Mean absolute difference

PACS Picture archiving and communication system

RMSE Root mean squared error

TW3 Tanner-Whitehouse 3

Declarations

Funding This study was supported by Youth clinical research project of Peking University First Hospital under project number 2019CR29.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Kai Zhao, Shuai Ma, Zhaonan Sun and Xiang Liu. The first draft of the manuscript was written by Kai Zhao and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval This study was approved by the Institutional Review Board and Ethics Committee of Peking University First Hospital (IRB No. 2017[1382]).

Consent to participate Informed consent was waived because the X-rays of the wrist were collected retrospectively and analyzed anonymously.

Consent to publish Not applicable.

References

1. Alshamrani, K., & Offiah, A. C. (2020). Applicability of two commonly used bone age assessment methods to twenty-first century UK children. Eur Radiol, 30(1), 504–513. doi:10.1007/s00330-019-06300-x
2. Berst, M. J., Dolan, L., Bogdanowicz, M. M., Stevens, M. A., Chow, S., & Brandser, E. A. (2001). Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. AJR Am J Roentgenol, 176(2), 507–510. doi:10.2214/ajr.176.2.1760507
3. Booz, C., Yel, I., Wichmann, J. L., Boettger, S., Al Kamali, A., Albrecht, M. H.,.. . Bodelle, B. (2020). Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated algorithm compared to the Greulich-Pyle method. Eur Radiol Exp, 4(1), 6. doi:10.1186/s41747-019-0139-9
4. Bull, R. K., Edwards, P. D., Kemp, P. M., Fry, S., & Hughes, I. A. (1999). Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. Arch Dis Child, 81(2), 172–173. doi:10.1136/adc.81.2.172
5. Creo, A. L., & Schwenk, W. F., 2nd. (2017). Bone Age: A Handy Tool for Pediatric Providers. Pediatrics, 140(6). doi:10.1542/peds.2017-1486
6. Dallora, A. L., Anderberg, P., Kvist, O., Mendes, E., Diaz Ruiz, S., & Sanmartin Berglund, J. (2019). Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. PLoS One, 14(7), e0220242. doi:10.1371/journal.pone.0220242
7. Gyftopoulos, S., Lin, D., Knoll, F., Doshi, A. M., Rodrigues, T. C., & Recht, M. P. (2019). Artificial Intelligence in Musculoskeletal Imaging: Current Status and Future Directions. AJR Am J Roentgenol, 213(3), 506–513. doi:10.2214/AJR.19.21117
8. Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., Cicero, M.,.. . Flanders, A. E. (2019). The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology, 290(2), 498–503. doi:10.1148/radiol.2018180736
9. Kim, J. R., Shim, W. H., Yoon, H. M., Hong, S. H., Lee, J. S., Cho, Y. A., & Kim, S. (2017). Computerized Bone Age Estimation Using Deep Learning Based Program: Evaluation of the Accuracy and Efficiency. AJR Am J Roentgenol, 209(6), 1374–1380. doi:10.2214/AJR.17.18224
10. King, D. G., Steventon, D. M., O'Sullivan, M. P., Cook, A. M., Hornsby, V. P., Jefferson, I. G., & King, P. R. (1994). Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods. Br J Radiol, 67(801), 848–851. doi:10.1259/0007-1285-67-801-848
11. Koc, U., Taydas, O., Bolu, S., Elhan, A. H., & Karakas, S. P. (2021). The Greulich-Pyle and Gilsanz-Ratib atlas method versus automated estimation tool for bone age: a multi-observer agreement study. Jpn J Radiol, 39(3), 267–272. doi:10.1007/s11604-020-01055-8

12. Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V., & Langlotz, C. P. (2018). Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology, 287(1), 313–322. doi:10.1148/radiol.2017170236
13. Lea, W. W., Hong, S. J., Nam, H. K., Kang, W. Y., Yang, Z. P., & Noh, E. J. (2022). External validation of deep learning-based bone-age software: a preliminary study with real world data. Sci Rep, 12(1), 1232. doi:10.1038/s41598-022-05282-z
14. Lee, K. C., Lee, K. H., Kang, C. H., Ahn, K. S., Chung, L. Y., Lee, J. J.,.. . Shim, E. (2021). Clinical Validation of a Deep Learning-Based Hybrid (Greulich-Pyle and Modified Tanner-Whitehouse) Method for Bone Age Assessment. Korean J Radiol, 22(12), 2017–2025. doi:10.3348/kjr.2020.1468
15. Nadeem, M. W., Goh, H. G., Ali, A., Hussain, M., Khan, M. A., & Ponnusamy, V. A. (2020). Bone Age Assessment Empowered with Deep Learning: A Survey, Open Research Challenges and Future Directions. Diagnostics (Basel), 10(10). doi:10.3390/diagnostics10100781
16. Ren, X., Li, T., Yang, X., Wang, S., Ahmad, S., Xiang, L.,.. . Wang, Q. (2019). Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment From Hand Radiograph. IEEE J Biomed Health Inform, 23(5), 2030–2038. doi:10.1109/JBHI.2018.2876916
17. Satoh, M. (2015). Bone age: assessment methods and clinical applications. Clin Pediatr Endocrinol, 24(4), 143–152. doi:10.1297/cpe.24.143
18. Siegel, E. L. (2019). What Can We Learn from the RSNA Pediatric Bone Age Machine Learning Challenge? Radiology, 290(2), 504–505. doi:10.1148/radiol.2018182657
19. Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., & Leonardi, R. (2017). Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal, 36, 41–51. doi:10.1016/j.media.2016.10.010
20. Tajmir, S. H., Lee, H., Shailam, R., Gale, H. I., Nguyen, J. C., Westra, S. J.,.. . Do, S. (2019). Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skeletal Radiol, 48(2), 275–283. doi:10.1007/s00256-018-3033-2
21. Wang, F., Cidan, W., Gu, X., Chen, S., Yin, W., Liu, Y.,.. . Jin, Z. (2021). Performance of an artificial intelligence system for bone age assessment in Tibet. Br J Radiol, 94(1120), 20201119. doi:10.1259/bjr.20201119
22. Wang, F., Gu, X., Chen, S., Liu, Y., Shen, Q., Pan, H.,.. . Jin, Z. (2020). Artificial intelligence system can achieve comparable results to experts for bone age assessment of Chinese children with abnormal growth and development. PeerJ, 8, e8854. doi:10.7717/peerj.8854
23. Zhou, X. L., Wang, E. G., Lin, Q., Dong, G. P., Wu, W., Huang, K.,.. . Fu, J. F. (2020). Diagnostic performance of convolutional neural network-based Tanner-Whitehouse 3 bone age assessment system. Quant Imaging Med Surg, 10(3), 657–667. doi:10.21037/qims.2020.02.20

Figures

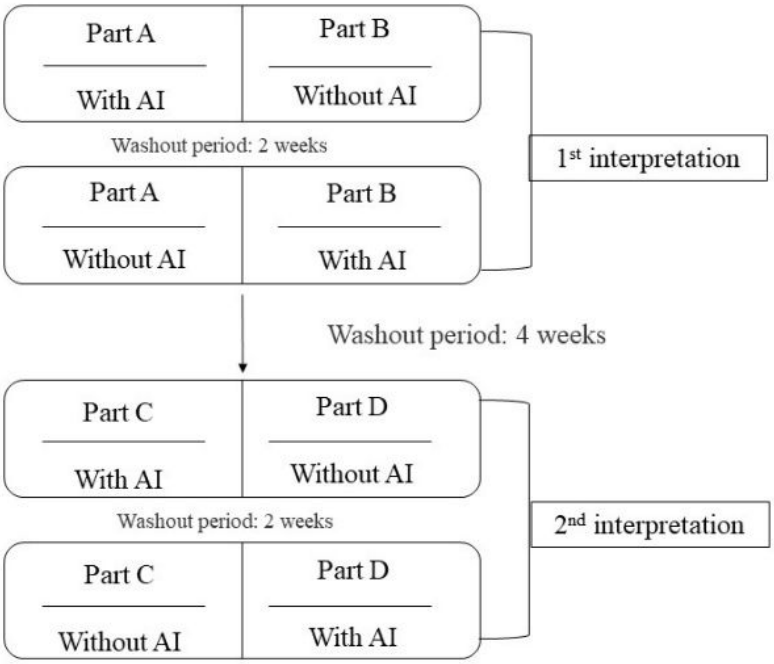

Figure 1

Flow chart of image interpretation for 6 residents. (The database were divided randomly and equally into A & B for the first interpretation and C & D for the second interpretation. Part A, B, C, D were not the same for each resident)