The New England Journal of Statistics in Data Science Volume 0, 1–8 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS111

The Framework That Survives Bad Models: Human-AI Collaboration for Clinical Trials

Yao CHEN[1], David OHLSSEN[1], Aimee READIE[1], Gregory LIGOZIO[1], Ruvie MARTIN[1], Alain MARTI[2], and Thibaud COROLLER[1,*]

Abstract

Artificial intelligence (AI) holds great promise for supporting clinical trials, from patient recruitment and endpoint assessment to treatment response prediction. However, deploying AI without safeguards poses significant risks, particularly when evaluating patient endpoints that directly impact trial conclusions. We compared two AI frameworks against human-only assessment for medical image-based disease evaluation, measuring cost, accuracy, robustness, and generalization ability. To stress-test these frameworks, we injected bad models—ranging from random guesses to naive predictions—to ensure that observed treatment effects remain valid even under severe model degradation. We evaluated the frameworks using two randomized controlled trials with endpoints derived from spinal X-ray images. Our findings indicate that using AI as a supporting reader (AI-SR) is the most suitable approach for clinical trials, as it meets all criteria across various model types, even with bad models. This method consistently provides reliable disease estimation, preserves clinical trial treatment effect estimates and conclusions, and retains these advantages when applied to different populations.

keywords and phrases: AI, Drug development, Framework, Bad models.

1. INTRODUCTION

Today, AI has become an integrated part of our life, from large language models (LLMs) [17, 7], car navigation, to accurate weather forecasting. To become an integrated part of our life, those models have shown exceptional performance (i.e., models' ability to successfully complete a task), and the risk of any failure is greatly outweighed by the value they bring. However, in comparison, AI has had little impact on healthcare and drug development despite multiple areas where it can help. First, AI has been suggested for data analysis of vast amounts of data generated by clinical trials [6]. It can help researchers identify potential drug candidates, predict patient responses, and optimize trial designs. This can lead to personalized medicine, leading to more effective and targeted therapies. Secondly, AI has the potential to enhance the recruitment of suitable patients [8], one of the biggest challenges in clinical trials. AI can help identify and recruit eligible patients more efficiently [23] by analyzing electronic health records, omics information, and other relevant non-traditional data sources. And finally, disease assessment [22] using AI to evaluate disease progression (e.g., grading medical images) can greatly reduce clinical trial costs and time, as well as reducing error (reader variability, human entry error).

To bring AI to clinical trials, researchers can either try to improve model performance to match human performance and place all risk on AI model alone, or mitigate AI risks in clinical trials, accepting the model may perform worse than human readers. Recent regulatory guidance has similarly emphasized that AI use in drug development should be evaluated according to its context of use, risk, robustness, human oversight, and impact on regulatory decision-making [18, 19, 5]. In this study we will focus on the latter, understanding how a trained model (as high performance as possible) will perform if used for clinical trials. We also explored the cases where the AI model is dangerous (e.g. random prediction) in order to stress test the limit of our framework and ensure strong safeguards. Inspired by the study from Ng *et al.* [13] which looked at using AI to classify cancer images (human vs AI-independent vs. AI-supported), we modified the concept to be applied in immunology to classify spinal X-ray images. We further extended their original scope by quantifying the treatment effect (key endpoint in clinical trial). In summary, in this study we aim to quantify several key aspects of using AI in clinical trials, namely:

1. **Efficiency**: Can AI reduce clinical trial time and cost?
2. **Robustness**: Is using a moderately performing AI detrimental to the disease evaluation? To what extent is the model's performance acceptable for safe use?
3. **Generalization**: Can this framework be applied to more than one study?

*Corresponding author.
[1]Novartis Pharmaceuticals Corp., East Hanover, NJ.
[2]Novartis Pharma AG, Basel, Switzerland.

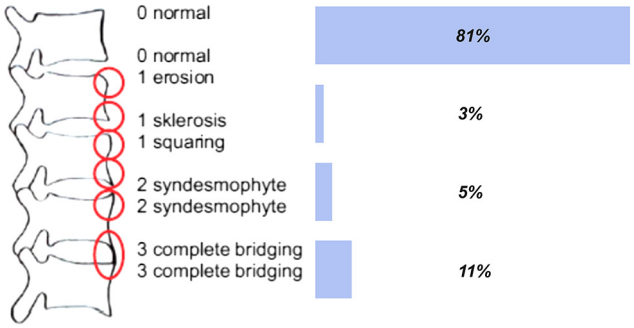


Figure 1: mSASSS definition [16, 20] and distribution in the dataset. mSASSS scoring criteria (0–3) based on structural changes at vertebral corners, alongside their frequency in the dataset. Most corners are normal (score 0, 81%), while erosion/sclerosis/squaring (score 1), syndesmophytes (score 2), and bridging (score 3) are less common (3%, 5%, and 11%, respectively).

4. **Accuracy**: Does use of AI lead to the same clinical finding at the clinical trial level (e.g. similar treatment effect estimation, same study conclusion)?

2. METHODS

2.1 Data

In this work, we used X-ray data from two phase III clinical studies, MEASURE I and PREVENT to test the effectiveness and generalization of the frameworks. Our AI grading model described later in Section 2.3 was trained on MEASURE I, predictions from the validation sets of MEASURE I and PREVENT, which guarantee the independence with the model were used for the test of the framework.

MEASURE I [1] is a completed and anonymized phase III trial for secukinumab (a fully human anti-IL-17A monoclonal antibody) in patients with axSpA (axial spondyloarthritis). MEASURE I included 361 patients randomized to the treatment and control groups in a 2:1 ratio. Patients on the control arm started with placebo and then switched to active treatment after 16 or 24 weeks, depending on their responder status. Two (or three X-ray) sessions were scheduled for each patient at baseline, 2 years (Week 104) and 4 years (Week 208). Lateral cervical and lumbar X-rays were taken during those sessions. The modified Stoke Ankylosing Spondylitis Spinal Scores (mSASSS) [20] in Figure 1 were evaluated by at least two radiologists, and exploratory endpoints were established to assess the worsening of mSASSS from baseline.

In this study, the mSASSS worsening after two years (between baseline and week 108 X-rays) was defined as an exploratory endpoint to determine the treatment effect between the treatment and placebo groups. Two reading sessions were conducted by two radiologists. During each session, the mSASSS assessment was conducted at each vertebral corner. The total score was calculated by summing the mSASSS values across 24 corners. Two reading sessions were conducted that aligned with the collection of the Week 104 and Week 208 images. During each reading session, the mSASSS was evaluated by two radiologists, each independently scoring the vertebral corners for structural changes. The total score was calculated by summing the mSASSS values across the 24 locations. To assess for inter-rater variability, the top 5% most discrepant total score cases between the two readers were submitted for adjudication. The distribution of mSASSS in each vertebral corner is shown in Figure 1, indicating a noticeable imbalance in healthy vertebrae versus those with structural changes due to disease progression. Similarly to MEASURE I, PREVENT [4] is another trial of secukinumab in axSpA but with a different patient population, that of non-radiographic axSpA. PREVENT collected the same spinal X-rays and shared the same reading schemes as MEASURE I. However, PREVENT is very different from MEASURE I in terms of patient population. The population in PREVENT is much less severe. In fact, more than 90% of the patients have zero mSASSS in PREVENT, which makes it a perfect trial to test the generalization ability of the candidate frameworks.

2.2 Framework

Our framework (Figure 2) adapted the Ng *et al.* [13] framework to grade spinal images. For each patient, two (or more) images were graded and a composite score was computed (mSASSS) using all vertebral corners. To mimic the human process as closely as possible, we only compared total spine mSASSS as opposed to individual vertebra levels. As with human graders, this approach focuses on the consistency of the total score rather than mSASSS at each individual vertebral corner. In such a setting, the Human Double reader in Figure 2 is the current clinical application, while the two others (AI-independent and AI-supported) are the experimental methods we are scoping in this work. The AI model utilized in this test is detailed in Section 2.3. If the AI model fails to predict the mSASSS, this missingness will automatically trigger a disagreement, necessitating the involvement of at least two human readers. Additionally, we explored the use of unreliable AI models within these frameworks to evaluate the robustness of the outcomes in the following sections.

2.3 Model

A classification pipeline that incorporates vertebral unit segmentation [2] followed by mSASSS classification [12] was developed as our AI reader (Figure 3). An ensemble model of nn-UNet [15, 11] and mask-RCNN [10], VertXNet [2], was used for vertebrae segmentation and labeling. All vertebral units were extracted on the basis of the segmentation results. Each vertebral unit would then be passed to a hierarchical grading system that consists of two ResNet 152 [9] for the prediction of mSASSS.

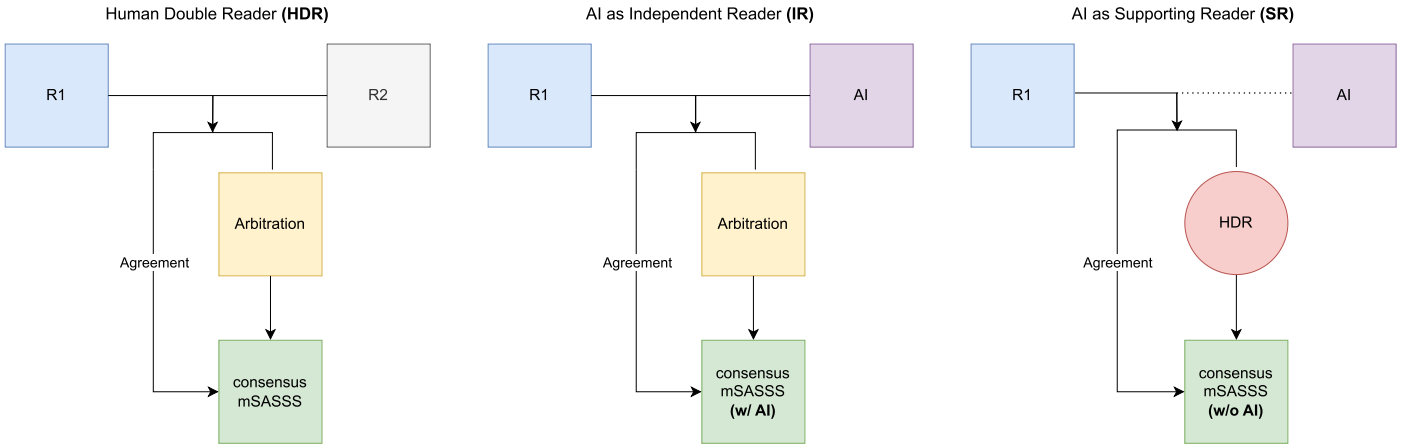


Figure 2: Introducing the framework implementation for the experiment, where we compare the gold standard, human double reader (**HDR**) to two AI approaches, namely AI as independent (**IR**) or supporting reader (**SR**) to evaluate spinal disease status (modified Stoke Ankylosing Spondylitis Spinal Scores - mSASSS) from X-ray images. At the end of each framework, the scores from each reader are pooled to obtain the consensus mSASSS. One key difference between the two AI frameworks, is that AI-IR includes the AI score for the consensus mSASSS, while AI-SR does not.

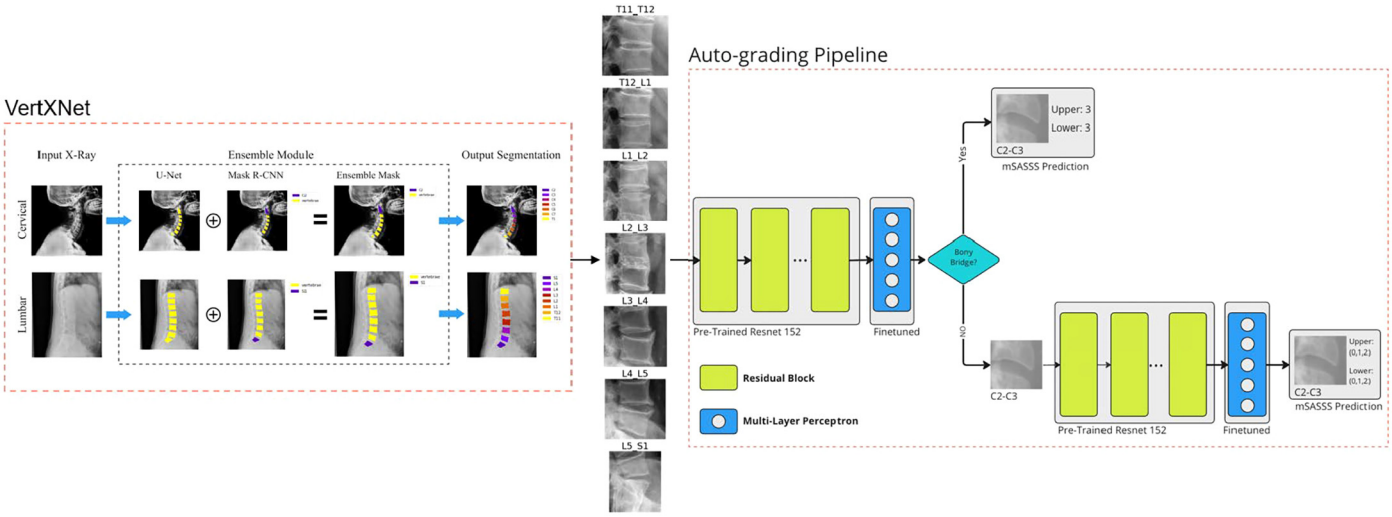


Figure 3: End to end pipeline to extract vertebral units (VUs) and mSASSS classification.

VertXNet has successfully extracted vertebral units with labels from 282 patients in MEASURE 1 and mSASSS were predicted by the auto-grading pipeline in Figure 3. mSASSS that could not be predicted from the pipeline were treated as missing as discussed in Section 2.2. All mSASSS were predicted by models that were not trained with its original score. The grading model has achieved moderate performance of 65% balanced accuracy which is the average of recall obtained on each class.

3. RESULTS

In this section, we compare the three frameworks mentioned in Figure 2 from various perspectives, including cost, robustness, and study results. Our AI model, described in Section 2.3, serves as the basis for this comparison. To evaluate the robustness of these frameworks under less effective models, we also tested them using random and naïve models. The random model predicts individual mSASSS scores from 0 to 3 with equal probability, while the naïve model assigns a score of 0 to all mSASSS scores. These models simulate scenarios where a reasonable AI model is unavailable, allowing us to assess whether the frameworks can still produce robust results despite the bias introduced by the AI model.

3.1 Cost

In this section, we evaluated and compared the cost of using different frameworks that mentioned in Figure 2. When relying solely on human readers, the results indicate that a

Table 1. Percentage of patients requiring a second human reader and arbitration grading for the trained models.

Method	Second human (%)	Arbitration (%)
HDR	100	48.92
AI-IR	**0.0**	59.83
AI-SR	59.83	**39.17**

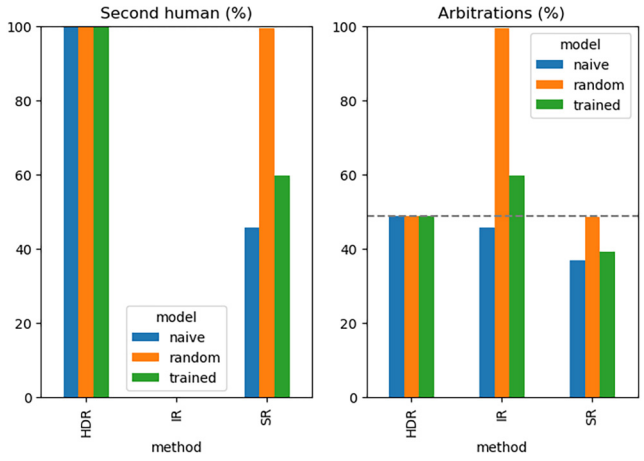


Figure 4: Cost difference between HDR, AI-IR and AI-SR when using extreme cases in prediction (naive and random model vs trained).

second human reader and arbitration are required in 100% and 48.92% of cases, respectively, as shown in Table 1. For the AI model, it was 0.0% and 59.83% when AI is used as independent reader, and 59.83% and 39.17% when used as supportive reader. When comparing cost across the frameworks with the trained models from Section 2.3, both AI approaches (AI-IR, AI-SR) are more efficient than human alone (HDR) as they require fewer patient datasets to be read by a human reader.

When stress testing our framework using a random model (each vertebral corner score is a random value between 0 and 3, uniformly distributed), we observed that AI-SR is matching the same cost as human alone – which is expected as safe-fail from the framework.

In practice, arbitration can be more costly than initial readings because it often requires the involvement of an additional reader, increased experience levels or the initial readers to repeat their assessments. We investigated the scenarios that arbitration costs are significantly higher than those of a second human reader, with costs ranging from one to five times higher. Our findings revealed that both AI methods consistently outperformed human readers. Notably, when the arbitration cost reached three times that of the second human reader, AI-SR surpassed AI-IR and continued to maintain its superiority as the cost ratio increased.

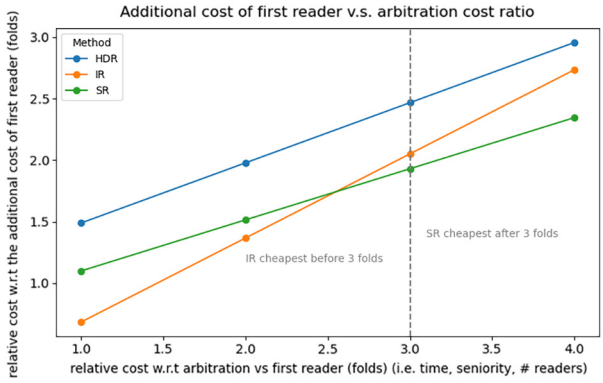


Figure 5: Impact of relative cost of arbitration w.r.t. the first human reader.

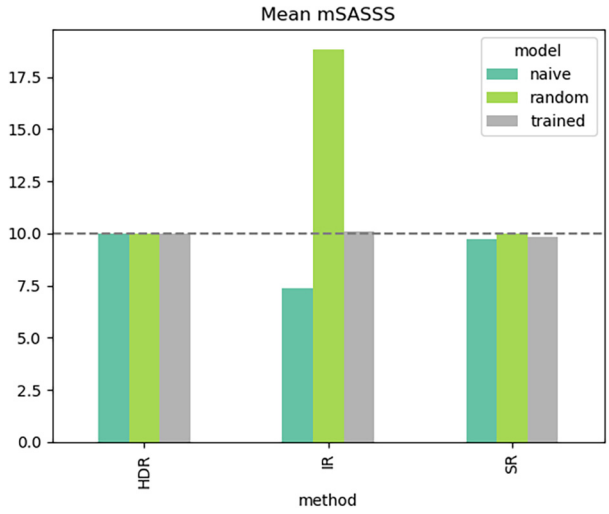


Figure 6: Mean mSASSS per patient accross all frameworks and models.

3.2 Model Robustness

The first and foremost quality of the framework is to be robust, which means that using AI should not change the disease evaluation. In this section, we compared the estimated mean mSASSS, across different models and methods in Figure 6. For the trained model, all methods yielded similar estimates (10.01, 10.09 & 9.81) for the disease burden. But we found that AI-IR wasn't robust for extreme cases (i.e. random model), yielding 18.90, almost double the true estimate. This is because the mSASSS consensus also uses the AI score, leading to a bias in the estimated value. For example, if you use a random model predicting either 0 or 1, the average value per sample is 0.5. Here for mSASSS, the average value is 1.5, because we randomly sample between 0 and 3 for each VU. On the other hand, when using a naive model (predicting only the majority class, for example here 0), the framework was much closer to the human, mostly due to the fact that the dataset predominantly in-

Table 2. Trial results for MEASURE I. Human double reader has demonstrated very close reproduction of the trial results; Independent reader has over-estimation bias in both arms and time steps; Supportive reader has very close to the trial results on all estimation consider the large standard deviation.

Method	Arm	Baseline	Week 104	Worsening
TRIAL	Treatment	10.79(17.48)	10.98(17.49)	0.54(4.73)
	Control	9.49(16.26)	11.87(18.20)	0.91(4.45)
Human double reader (HDR)	Treatment	10.85(18.09)	11.04(17.39)	0.56(3.08)
	Control	9.37(16.24)	12.19(18.61)	0.83(3.62)
Independent reader (IR)	Treatment	11.15(17.84)	11.26(17.38)	0.49(2.95)
	Control	9.64(16.14)	12.36(18.47)	0.77(3.68)
Supportive reader (SR)	Treatment	10.75(18.13)	10.99(17.40)	0.51(3.11)
	Control	9.36(16.25)	12.11(18.64)	0.78(3.63)

cludes patients with mSASSS vertebral corners of zero (approximately 80%).

3.3 Treatment Effect

In addition to the cost of the mSASSS reading and robustness of the frameworks, it is essential to verify that the conclusions drawn using these frameworks would align with those of human readers. Our simulation facilitated the comparison of results across real studies and different frameworks. In this section, we analyzed the clinical trial outcome, including summary statistics, endpoint distribution, and progression curves, across different frameworks, and compared them with the original trial results between treatment and placebo arms.

The table of summary statistics and the conclusion of the hypothesis test on the endpoint are among the most crucial results of a trial. In our studies, worsening of mSASSS from baseline to 2 years was used as the endpoint to estimate the treatment effect. From Table 2, it is evident that the conclusions on worsening are consistent across study outcomes and different frameworks. Although the control arm shows slightly worse worsening than the treatment arm, the difference is not statistically significant due to the large standard error. For other summary statistics, both HDR and AI-SR have demonstrated similar estimation with the trial results. However, AI-IR exhibits an overestimation bias at both baseline and week 104 mSASSS.

We examined the distribution of mSASSS scores and the probability of progression in Figure 7 and Figure 8. The overall distribution of mSASSS indicates that all frameworks are generating similar structural progression results as the trial results. Figure 8 depicts the proportion of patients experiencing worsening below the levels indicated on the y-axis. From this figure, we can derive critical estimates, such as the percentage of patients with mSASSS progression, where the overlapping curves confirm alignment of the conclusions across all frameworks.

All frameworks can produce the same statistical conclusions as the trial results regarding the hypothesis test. When

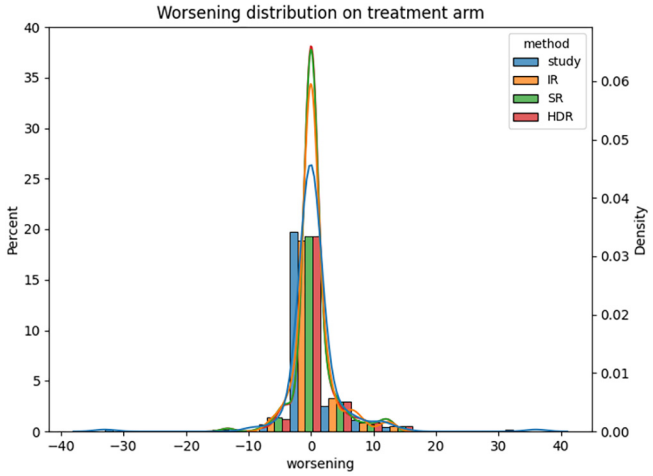


Figure 7: mSASSS worsening distribution in MEASURE I. mSASSS worsening distribution is plotted side by side for each bar; similar distributions of worsening were observed across all frameworks.

AI acts as an independent reader, it has shown a systematic bias in its estimates on both arms. However, when AI serves as a supportive reader, it aligns quite well with the trial outcomes.

3.4 Generalization

In reality, we seldom run two identical X-ray studies on design, disease area, doses or population. Even in the case of identical design and inclusion/exclusion criteria, the enrolled patient populations of the completed studies will be slightly different. To evaluate the potential benefits and robustness of AI-assisted frameworks in a new trial, we utilized the PREVENT trial. Given the differences of the patient population compared to MEASURE I and the more pronounced imbalance in patients' mSASSS readings, the AI grading model trained on MEASURE I would perform poorly on the PREVENT trial if used alone. However, our investigation revealed that all frameworks consistently

Table 3. Trial and model results for PREVENT. AI-IR performs poorly in the PREVENT population, leading to less accurate estimations. However, when used as a supportive reader, the AI provides estimations that are closely aligned with the trial results, despite the large standard deviation.

Method	Arm	Baseline	Week 104	Worsening
TRIAL	Treatment	0.74(2.49)	0.72(2.49)	0.03(0.48)
	Control	0.84(2.43)	0.88(2.60)	0.03(0.63)
Human double reader (HDR)	Treatment	0.78(2.64)	0.76(2.65)	0.03(0.53)
	Control	0.84(2.49)	0.97(2.86)	0.09(0.46)
Independent reader (AI-IR)	Treatment	1.18(2.64)	1.26(2.75)	0.09(1.51)
	Control	1.31(2.68)	1.29(3.00)	−0.10(1.41)
Supportive reader (AI-SR)	Treatment	0.73(2.59)	0.66(2.47)	0.01(0.72)
	Control	0.77(2.48)	0.88(2.85)	0.07(0.48)

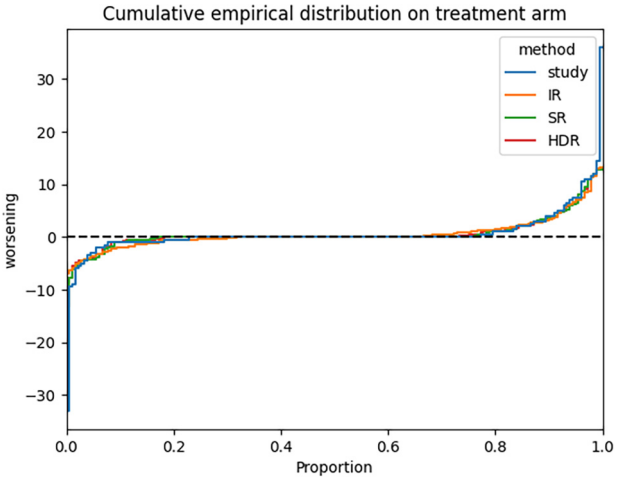


Figure 8: Empirical distribution of progression in MEASURE I. This figure shows the percentage of patients with worsening or improvement by a different extent displayed on the y-axis. All frameworks overlapping with each other pretty well.

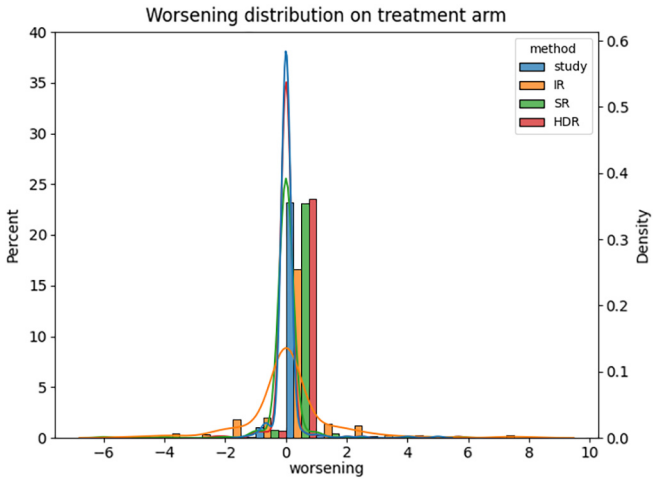


Figure 9: mSASSS worsening distribution in PREVENT. mSASSS worsening distribution is plotted side by side for each bar; similar distributions of worsening were observed across all frameworks.

reached the same conclusion as the trial results: there is no significant difference in worsening between the two arms. However, under close evaluation of the estimation and figures, AI-IR fails to generate convincing outcomes in this case. We displayed the same Table 3 and Figure 9 / 10 on the PREVENT trial as follows.

The table and figures demonstrate the generalization ability of the framework where AI is used as a supportive reader (AI-SR) for other populations in the same domain. In contrast, AI-IR did not generalize well to the new trial, showing an overestimation bias in the table and a differing progression probability curve in Figure 10. On the other hand, AI-SR closely aligned with all perspectives of the HDR and the trial results. This provides evidence that the AI-assisted grading framework, particularly AI-SR, is valid and reliable in a new trial, even with a different population.

4. DISCUSSION

In this study, we investigated the potential application of AI in clinical trials. We characterized that an effective AI framework should meet three key criteria: accuracy, efficiency, and robustness. First, the AI framework should produce the same trial results as those obtained by human-only graders. Second, the AI framework should save time and money compared to a human-only grading system. Lastly, the AI framework should deliver consistent evaluations even when its model performance is suboptimal. To test these criteria, we experimented with three different methods under varying model performances: trained, random, and naive. This approach allowed us to stress test the limits of each method within the framework.

Our research found that the AI as supportive reader (AI-SR) approach is the most suitable for clinical trials, as it meets all criteria across all model types. This framework excels in accuracy because it places AI under human super-

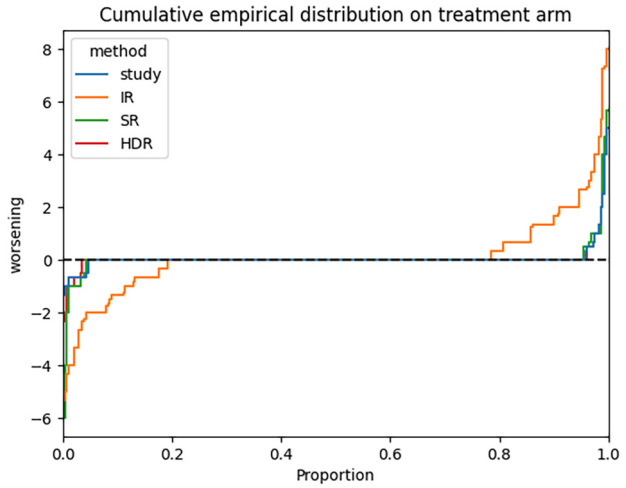


Figure 10: Empirical distribution of progression in PREVENT. This figure shows the percentage of patients with worsening or improvement by a different extent displayed on the y-axis. All frameworks overlapping with each other pretty well except AI-IR.

Table 4. Overall evaluation of all methods. AI as supporting reader (AI-SR) has proven to be the most robust, efficient and accurate method with all models. ✓ indicates success, ✗ indicates failure, ✓✗ indicates a partial success.

Framework	HDR	AI-IR	AI-SR
Cost	✗	✓	✓
Robustness	✓	✗	✓
Generalization	✓	✗	✓
Accuracy	✓	✓✗	✓

vision, ensuring safety and efficiency by reducing the number of human readers needed. It also provides accurate estimations of disease status and treatment effects. One finding is that we don't need a high performing model to see any gain in the efficiency while keeping accurate results. Like ensemble method, it does not require a strong individual grader, AI or human, but still performs consistently on the final output. In practice, when the sample size is small or the outcome is imbalanced, only moderately performing models are achievable. In such a case, it is recommended that AI serves as a supportive reader that has less weight on the final grading. This framework would still reduce trial cost even when the arbitrator is much more expensive than the first/second reader. From an implementation standpoint, AI-SR integrates naturally into existing central imaging workflows: the AI serves as a supporting reader alongside the first human reader, with human adjudication triggered only upon disagreement, thereby mirroring the established double-read-plus-adjudication paradigm while preserving human oversight. Several practical considerations are important for prospective deployment. To ensure reproducibility, the AI model weights, image preprocessing steps, and post-processing rules should be locked before prospective inference or trial image reads begin. To preserve study integrity, the AI-SR decision logic should be prespecified in the Statistical Analysis Plan or imaging charter, the AI should remain blinded to treatment assignment and visit order as required by the study design, and the handling of ungradable images should be documented prospectively. Within this framework, AI-SR can be positioned within a decision-support context of use, as a qualified human reader remains the ultimate arbitrator. This approach is consistent with key principles emphasized in the FDA's draft guidance on AI to support regulatory decision-making [18], the FDA Good Machine Learning Practice guiding principles [19], and the EMA reflection paper on AI in the medicinal product lifecycle [5]. At the same time, AI graders could also be updated by pooling more studies. As their performance enhances, the weight of the AI graders could also be elevated. Consequently, AI serving as independent reader could be more reliable.

To the best of our knowledge, this is the first study to examine the impact of an AI-assisted grading framework on the evaluation of treatment effects and trial conclusions in a clinical setting. This finding is particularly significant as it highlights the importance of the choice of the right AI framework for assessing outcomes of clinical trials.

Accepted 19 October 2025

REFERENCES

[1] Baeten, D., Sieper, J., Braun, J., Baraliakos, X., Dougados, M., Emery, P., Deodhar, A., Porter, B., Martin, R., Andersson, M., Mpofu, S. and Richards, H. B. (2015). Secukinumab, an Interleukin-17A Inhibitor, in Ankylosing Spondylitis. *New England Journal of Medicine* **373**(26) 2534–2548. https://doi.org/10.1056/NEJMoa1505066. https://www.nejm.org/doi/pdf/10.1056/NEJMoa1505066.

[2] Chen, Y., Mo, Y., Readie, A., Ligozio, G., Mandal, I., Jabbar, F., Coroller, T. and Papież, B. W. (2024). VertXNet: an ensemble method for vertebral body segmentation and identification from cervical and lumbar spinal X-rays. *Scientific Reports* **14**(1) 3341.

[3] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M. et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging* **26** 1045–1057.

[4] Deodhar, A., Blanco, R., Dokoupilová, E., Hall, S., Kameda, H., Kivitz, A. J., Poddubnyy, D., van de Sande, M., Wiksten, A. S., Porter, B. O. et al. (2021). Improvement of signs and symptoms of nonradiographic axial spondyloarthritis in patients treated with secukinumab: primary results of a randomized, placebo-controlled phase III study. *Arthritis & Rheumatology* **73**(1) 110–120.

[5] European Medicines Agency (2024). *Reflection Paper on the Use of Artificial Intelligence in the Medicinal Product Lifecycle*. EMA/CHMP/CVMP/83833/2023.

[6] Fournier, L., Costaridou, L., Bidaut, L., Michoux, N., Lecouvet, F. E., de Geus-Oei, L. -F., Boellaard, R., Oprea-Lager, D. E., Obuchowski, N. A., Caroli, A. et al. (2021).

Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers. *European radiology* **31** 6001–6012.

[7] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

[8] Harrer, S., Shah, P., Antony, B. and Hu, J. (2019). Artificial intelligence for clinical trial design. *Trends in pharmacological sciences* **40**(8) 577–591.

[9] He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778.

[10] He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* 2961–2969.

[11] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2) 203–211.

[12] Mo, Y., Chen, Y., Readie, A., Ligozio, G., Coroller, T. and Papież, B. W. (2023). Towards automatic scoring of spinal X-ray for ankylosing spondylitis. *arXiv preprint arXiv:2308.05123*.

[13] Ng, A. Y., Glocker, B., Oberije, C., Fox, G., Sharma, N., James, J. J., Ambrózay, É., Nash, J., Karpati, E., Kerruish, S. et al. (2023). Artificial intelligence as supporting reader in breast screening: a novel workflow to preserve quality and reduce workload. *Journal of Breast Imaging* **5**(3) 267–276.

[14] Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T. L., Bernatz, S., Hosny, A., Mak, R. H., Birkbak, N. J. et al. (2024). Foundation model for cancer imaging biomarkers. *Nature machine intelligence* **6**(3) 354–367.

[15] Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* 234–241. Springer.

[16] Taylor, H. G., Wardle, T., Beswick, E. J. and Dawes, P. T. (1991). The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. *British Journal of Rheumatology* **30**(5) 330–335. https://doi.org/10.1093/rheumatology/30.5.330. https://academic.oup.com/rheumatology/article-pdf/30/5/330/5041584/30-5-330.pdf.

[17] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. -A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[18] U. S. Food and Drug Administration (2025). *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products: Draft Guidance for Industry*. Draft guidance.

[19] U. S. Food and Drug Administration, Health Canada and Medicines and Healthcare products Regulatory Agency (2021). *Good Machine Learning Practice for Medical Device Development: Guiding Principles*.

[20] Van Der Heijde, D., Braun, J., Deodhar, A., Baraliakos, X., Landewé, R., Richards, H. B., Porter, B. and Readie, A. (2019). Modified stoke ankylosing spondylitis spinal score as an outcome measure to assess the impact of treatment on structural progression in ankylosing spondylitis. *Rheumatology* **58**(3) 388–400.

[21] Vera, P. and Makhlouf, M. (2023). *BIOptimus: Pre-training an Optimal Biomedical Language Model with Curriculum Learning for Named Entity Recognition*. arXiv:2308.08625. https://arxiv.org/abs/2308.08625.

[22] Yang, Y., Yuan, Y., Zhang, G., Wang, H., Chen, Y. -C., Liu, Y., Tarolli, C. G., Crepeau, D., Bukartyk, J., Junna, M. R. et al. (2022). Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature medicine* **28**(10) 2207–2215.

[23] Zhang, B., Zhang, L., Chen, Q., Jin, Z., Liu, S. and Zhang, S. (2023). Harnessing artificial intelligence to improve clinical trial design. *Communications Medicine* **3**(1) 191.

Yao Chen. 1 Health Plaza, East Hanover, NJ, Novartis Pharmaceuticals Corp., US.
E-mail address: yao.chen@novartis.com

David Ohlssen. 1 Health Plaza, East Hanover, NJ, Novartis Pharmaceuticals Corp., US.
E-mail address: david.ohlssen@novartis.com

Aimee Readie. 1 Health Plaza, East Hanover, NJ, Novartis Pharmaceuticals Corp., US.
E-mail address: aimee.readie@novartis.com

Gregory Ligozio. 1 Health Plaza, East Hanover, NJ, Novartis Pharmaceuticals Corp., US.
E-mail address: gregory.ligozio@novartis.com

Ruvie Martin. 1 Health Plaza, East Hanover, NJ, Novartis Pharmaceuticals Corp., US.
E-mail address: ruvie.martin@novartis.com

Alain Marti. Lichtstrasse 35, Basel, Novartis Pharma AG, CH.
E-mail address: alain.marti@novartis.com

Thibaud Coroller. 1 Health Plaza, East Hanover, NJ, Novartis Pharmaceuticals Corp., US.
E-mail address: thibaud.coroller@novartis.com