Title

The MI-CLAIM-GEN checklist for generative artificial intelligence in health

Authors

Miao, Brenda Y
Chen, Irene Y
Williams, Christopher YK
et al.

Publication Date

2025-02-06

DOI

10.1038/s41591-024-03470-0

Peer reviewed

https://doi.org/10.1038/s41591-024-03470-0

The MI-CLAIM-GEN checklist for generative artificial intelligence in health

Brenda Y. Miao, Irene Y. Chen, Christopher Y. K. Williams, Jaysón Davidson, Augusto Garcia-Agundez, Shenghuan Sun, Travis Zack, Suchi Saria, Rima Arnaout, Giorgio Quer, Hossein J. Sadaei, Ali Torkamani, Brett Beaulieu-Jones, Bin Yu, Milena Gianfrancesco, Atul J. Butte, Beau Norgeot & Madhumita Sushil

The MI-CLAIM checklist has been revised to take account of new capabilities of large language models and other generative artificial intelligence tools.

The 'Minimum information about clinical artificial intelligence modeling' (MI-CLAIM) checklist[1], released in 2020, provided a set of transparent, reproducible reporting guidelines for artificial intelligence (AI) modeling studies in medicine. Since then, there have been many advances in generative models for clinical AI research[2,3], including new capabilities of large language models (LLMs), diffusion models, vision language models, and other multimodal models.

In response to gaps in standards and best practices for the reporting of clinical generative AI research identified by US Executive Order 14110[4] and several emerging national networks for clinical AI evaluation[5], we began to formalize some of these guidelines by building on the original MI-CLAIM checklist. The new checklist, MI-CLAIM-GEN (Table 1), aims to address differences in training, evaluation, interpretability, and reproducibility of new generative models compared with non-generative (predictive) models.

Part 1: study design

All elements of the study design from the original MI-CLAIM checklist, including clear descriptions of the research question and cohort selection, also apply to generative AI studies. For cohort selection that involves unstructured or multimodal data, keyword search terms, regular expressions, or other selection criteria should be made available. If qualitative factors, such as manual chart review, are used to identify patient cohorts, these should be detailed and the qualifications of the reviewer (such as years of practice or specialty) should be reported. We discourage the use of ambiguous language, such as 'patients diagnosed with diabetes', in favor of more reproducible terms and standard ontologies, such as 'patients with at least two of the following ICD-10 codes: E11.*, E13.*'.

Datasets used should be representative of the clinical settings presented by the research question and any deviations should be described in detail. This may include data deidentification, including date-shifting for privacy protection, text redaction, which Electronic Health Record (EHR) vendor the data was derived from, if the data were limited to specific department(s), or other limitations compared with real-world settings and patient populations. Sensitivity analyses should be performed where appropriate to justify any selection criteria that deviate from established guidelines. Specifications for handling missing data should also be provided, if applicable.

Similar to traditional machine learning, the evaluation of generative AI studies often involves the use of categorical or continuous data labels. The source of the labels, including annotation guidelines and inter-annotator agreement for human labels, should be clearly documented. For unstructured outputs, such as summaries, that do not readily map to simple labels, we discuss both automated and human evaluation strategies in part 4.

Part 2: data and resource assessment

In addition to new model architectures, reporting on generative clinical models must also include additional information about external datasets or tools that a model may interact with through approaches such as retrieval augmented generation. We develop checklist items to reflect the inclusion of these different components of 'compound AI systems'[3] (Fig. 1).

Researchers should be careful of training data memorization, known as data leakage or contamination[6]. Training data may include values used for pretraining, fine-tuning, reinforcement learning, or other training schemes that update model weights, as well as any external datasets or tools for in-context learning.

Prompt engineering or other in-context learning should be performed using a development dataset that is also kept separate from the final test dataset. Previous studies have used 5% of the data or around 100 samples in prompt validation datasets[7]. Data splits should be performed at the patient level, with all data from each patient only included in one of the splits to maintain independence. Although the same validation dataset should be used for prompt engineering between different models, the best prompt selected for each model may vary. As prompt engineering is a rapidly evolving field, readers should follow best practice guidelines laid out by model developers and researchers.

Part 3: baseline model selection

Baseline model comparisons are important to provide controls for evaluating model performance. Generative model performance should be compared with rigorously selected baseline models, which may include other generative models but also non-generative approaches. Baseline models may also include previous versions of models and open-source model baselines should be included when available.

Any post-processing of model and baseline outputs should be described in the methods, including how errors or unexpected outputs are handled. If large training datasets are used for baselines models compared to zero- or few-shot approaches for generative models, researchers should report their performance across various volumes of data. Discussion of tradeoffs between compute and cost requirements is encouraged to improve understanding of the scalability and efficiency of these non-generative models.

Table 1 | MI-CLAIM-GEN checklist for generative AI clinical studies

Section	Checklist	Changes
Part 1: Study design	The clinical problem in which the model will be employed is clearly detailed in the paper	None
	The research question is clearly stated	None
	All cohort selection criteria and study design are detailed in such a way that they can be reproduced by an external researcher	Modified
	Identify whether the output data type categorical, continuous, or unstructured	Modified
	The characteristics of the cohorts are detailed in the text and are shown to be representative of real-world clinical settings	
Part 2: Resources and optimization	Model/application components are clearly detailed including: base model(s) used, embedding model(s), retrieval model(s), and other auxiliary models or tools	Modified
	The origin of all data sources for model training, finetuning, or inference is described and the original format is detailed in the paper	Modified
	All data preprocessing for model training, finetuning, or inference is described, including appropriate randomization and other transformations	Modified
	The independence between training, validation (including for prompt engineering), and test sets has been described, and data are split at the patient level	Modified
Parts 3–4: Model performance and evaluation	The state-of-the-art solution used as a baseline for comparison has been identified and detailed	None
	The performance comparison between the baseline and the proposed model is presented with the appropriate statistical significance	None
	Identify what evaluation(s) were performed, and provide clear justifications for the primary metrics used for each evaluation; describe whether overlap accuracy, semantic accuracy, and/or clinical utility were assessed	Modified
	If applicable, details on human evaluation are described, including any evaluation guidelines, level of experience of evaluators, inter-reviewer scores, etc	New
Part 5: Model examination	Relevant interpretability techniques, error analysis, and/or other approaches are applied to demonstrate an absence of unreasonable risk and brittleness, including a low risk of catastrophic and especially undetected failure	Modified
	A discussion of the risk revealed by the examination results is presented with respect to model/algorithm performance	None
	Describe step(s) taken to discuss, identify, and/or mitigate model biases, privacy and security concerns, and other potential harms	New
	A discussion and/or assessment of relevant distribution shifts and their impact on the model's performance has been provided	None
	Recommendations or discussion of post-deployment evaluation have been provided	New
Part 6: Reproducibility; data and model transparency	Choose appropriate tier: Tier 1: complete sharing of the code and data, including all prompts tested, hyperparameters used, software dependencies, model versions, and compute requirements Tier 2A: complete sharing of the code with synthetic data provided Tier 2B: complete sharing of the code Tier 3: no sharing of code or data	Modified
	A clinical model card is included summarizing the model capabilities, intended use, descriptions of any dataset or other integrations, limitations, potential biases, and risks	New
	If applicable, model weights are released to a secure repository with appropriate use agreements	New

Part 4A: automated model evaluation

Evaluation metrics for generative models should distinguish between metrics that measure 'overlap accuracy', which measure proportions of overlapping subunits (such as tokens or pixels), 'semantic accuracy', which compare the meanings of outputs and labels, and 'clinical utility', which measure how models affect clinical workflows or downstream patient outcomes[8]. We identify best practices for both automated and clinical expert evaluations, with a focus on metrics developed to handle the complex, unstructured outputs from generative models. Continuous evaluation and monitoring of these models will help to identify changes in model behavior, including dataset shifts or changes to model versions[2,9].

Similar to traditional machine learning classification set-ups, accuracy, F1 scores (for imbalanced datasets), or other suitable metrics should be reported, along with class distribution, for categorical labels. For continuous outputs, such as time saved or changes to patient activity scores, best practice statistical approaches and reporting should be applied, including appropriate estimators for causal effects and multiple hypothesis testing.

For unstructured text outputs, automated overlap scoring methods such as the BLEU and ROUGE scores are commonly used, but these do not assess whether the answers are clinically accurate and are often poorly correlated with human evaluation on biomedical tasks[10,11]. They also may not be suitable for tasks that lack standard reference documents, such as document summarization. Semantic scoring methods, such as BERT-based scoring methods, panels of similar metrics, or even using other generative models for evaluation, may provide more evidence of clinical similarities. However, rigorous evaluation is required

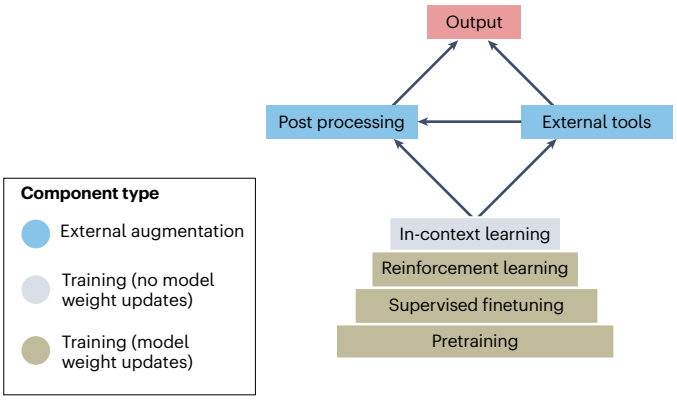

Fig. 1 | Components of model training and inference to report for end to end replication. Independent datasets and data splits (validation, test) used during any stage of model training should all be reported. This includes any data used for in-context learning, such as databases used for retrieval augmented generation or any prompt engineering performed. Any post-processing, including external tool usage, should also be reported. Models merging multiple, existing models should provide components for each model.

before applying these approaches at scale on new, clinical tasks and their credibility for the given study must be articulated if used.

Part 4B: human model evaluation

Human model evaluation remains the gold-standard for assessing semantic accuracy and clinical utility of generative models. As much as possible, evaluation should be conducted in a blinded fashion, with Turing-like assessments against ground truth values or across multiple metrics to gauge the accuracy, appropriateness, bias, and other aspects of model performance[10,11]. For complex outputs or simulated scenarios, objective structured clinical examination (OSCE) type evaluations can be considered to assess model performance across multiple axes that better reflect real-world clinical encounters or workflows[9,11]. Inter-reviewer variability and any evaluation guidelines should be reported. If models are deployed to interactive settings, user-provided prompts and any resulting variability should also be evaluated and discussed.

Part 5A: explainability and feature importance

Transparency research for generative models remains an active field of investigation, and therefore suggestions from the original MI-CLAIM checklist to apply best-practice methods, such as local explainability approaches (for example, LIME and SHAP), gradient and attention analyses, probing methods, or counterfactual analyses[12], are kept. However, these methods should be rigorously evaluated when applied to new clinical tasks, particularly because previous methods were often developed for models with shorter context lengths or less complex tasks. In addition, caution should be exercised when interpreting model-generated explanations, which do not always align with final outputs, and should not be used as a method of explainability[13].

Error analysis and sensitivity analysis (ablation tests), including prompt sensitivity tests, are also strongly encouraged to provide a better understanding of model behavior, particularly if evaluation datasets or models are not made publicly available. It is becoming increasingly important to understand how generative models may fail in clinical settings, which can provide insights into their capabilities and limitations beyond accuracy metrics.

Part 5B: bias, privacy and harm assessments

Identifying generative model biases and other potential harms pose new challenges for clinical research[2,11]. The MI-CLAIM-GEN introduces new checklist items to report whether studies discuss, identify and/or mitigate these and other potential harms.

Models trained on biased data can perpetuate biases in generated content, impacting downstream patient care and decision-making. Analysis of model performance across diverse patient subgroups and data subtypes to identify biases is strongly encouraged. All available details about data distribution of any training and evaluation datasets should also be reported, including patient sociodemographic information, any data imbalance, the timeframe when the data were collected, and any changes to best practice medical guidelines during this timeframe. To assess cultural and social biases, researchers should consider engaging with a diverse set of clinical evaluators. External validation to assess model fairness and robustness across different data distributions should also be performed if possible.

Models that may be deployed to real-world clinical care settings should be assessed with patient-centered approaches that are inclusive of diverse cultural and social communities[9,11]. These models should also be scrutinized for up-to-date cybersecurity vulnerabilities, such as adversarial prompt injections[2], and other potential model harms.

Part 6: end-to-end pipeline replication

Reproducible methods for generative modeling research should enable the community to replicate data collection and cohort selection, model development and inference, and end-to-end evaluation. New checklist items were added to identify the level of transparency presented, with separate tiers for reproducible data processing and model training or usage.

For data and analytic transparency, all code, dependencies and data should be provided in secure, accessible repositories. If full real-world datasets cannot be shared, a sample of the raw data, synthetic data or data structures and dictionaries should be provided. Model weights should be released and treated with the same care as clinical datasets. Use of any synthetic data and strategies for generation should follow individual journal guidelines on data reporting. Along with datasets used and code used for analysis, all prompts tested should be released, along with corresponding results.

Infrastructure, cost and compute requirements to run or develop the model should be included as part of the methods. These may include the type and quantity of hardware used, the operating system and the training time if applicable. For reproducible model development or usage, any random seeds used and other hyperparameters should also be reported, along with detailed descriptions of model inputs, versions and implementation frameworks, especially if code and/or data are not provided. External datasets for retrieval augmented generation, base model(s) used, embedding model(s), retrieval model(s), and other auxiliary models or tools, should also be disclosed (Fig. 1).

Drawing from best practices set out for all model development, the checklist also includes a section to report clinical model cards[14] or labels[15] that summarize the model capabilities, intended use, training data and limitations, potential biases, and model risks (Fig. 2). While the MI-CLAIM-GEN checklist summarizes whether the current clinical generative AI study has been conducted and reported using best-practice recommendations, model cards provide additional

Model Summary

Model name: Sepsis-GPT **Developer:** MI-CLAIM Health

FDA Clearance: N/A **Last updated:** June 20, 2024 **Version:** 1.0

Intended usage

- **Indication for use:** Assisting emergency physicians in diagnosing and managing patients with suspected sepsis.
- **Out-of-scope uses (contraindications):** Not intended for use in patients under 18 years old, pregnant women, or those with immunocompromised conditions.

Development

- **Pretrained model:** Clinical-T5 (derived from T5-Base)
- **Pretraining dataset description:** Further pretrained on MIMIC-III and -IV clinical notes
- **Fine-tuning**
  - **Method:** Supervised fine-tuning using labeled clinical notes, vital signs, and laboratory data
  - **Dataset:** 50,000 emergency department visits with confirmed sepsis diagnoses and severity labels derived from electronic health records
  - **Target:** Binary sepsis diagnosis and multiclass severity assessment
- **Prompt engineering**
  - **Method:** Few-shot learning with 3 random examples of clinical notes and corresponding diagnoses/severity assessments
  - **Dataset:** Curated set of representative clinical note snippets from 100 patients with annotations
  - **Target:** Accuracy of diagnosis and severity assessment, minimizing false negatives
- **External tools**
  - Sepsis-3 diagnostic criteria, SOFA score calculator, antibiotic recommendation engine

Validation and performance

Validation type	AUC (Diagnosis)	F1 (Management)	Cohort size	Dataset	Citation
Internal (retrospective)	0.83	0.56	1,000	Link 1	doi:####
Internal (prospective)	NA	NA	NA	NA	NA
External (retrospective)	0.68	0.63	2,500	Link 2	doi:####

- **Primary clinical metric:** Accuracy of sepsis diagnosis and appropriateness of management recommendations
- **Continuous monitoring recommendations:** Weekly review of a random sample of 50 model outputs by a clinical expert to assess the quality, appropriateness, and bias

Warnings

- **Risks resulting from bias findings:** Potential underdiagnosis in patients from underrepresented racial/ethnic groups.
- **Risks resulting from clinical findings:** False negative diagnoses could lead to delayed treatment; false positives could lead to overtreatment.
- **Other known or suspected risks within the intended domain:** Model may underperform on cases with incomplete data or atypical presentations.

Other information

- **Citation:** Placeholder et al. Sepsis-GPT model for sepsis diagnosis using real-world clinical data. 2024.
- **License:** MIT license

Fig. 2 | Components of a clinical model card. An example model card, formatted as a clinical 'model facts' label[15], for a fictional model created to assist in clinical decision support around sepsis diagnosis and management. The clinical model card should provide a summary of how a model was developed, intended use, out-of-scope uses, performance, limitations, and recommendations for safe deployment. SOFA, sequential organ failure assessment.

transparency around model development, intended uses, and known limitations to support the appropriate use of these models in future research or deployment.

The MI-CLAIM-GEN checklist can be found on Github at the following link: https://github.com/BMiao10/MI-CLAIM-GEN. We welcome continuous community feedback as the generative modeling landscape evolves, and provide this space as a community forum for readers to identify and engage with best-practice approaches within each section of the MI-CLAIM-GEN checklist.

Brenda Y. Miao[1] ✉, Irene Y. Chen[2,3,4], Christopher Y. K. Williams[1], Jaysón Davidson[1], Augusto Garcia-Agundez[5], Shenghuan Sun[1], Travis Zack[1,6], Suchi Saria[7,8,9,10], Rima Arnaout[1,2,11], Giorgio Quer[12], Hossein J. Sadaei[12,13], Ali Torkamani[12,13], Brett Beaulieu-Jones[14], Bin Yu[3,15,16], Milena Gianfrancesco[5], Atul J. Butte[1,17], Beau Norgeot[18,19] & Madhumita Sushil[1,19]

[1]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA. [2]UCSF-UC Berkeley Joint Program in Computational Precision Health, University of California, Berkeley and University of California, San Francisco, Berkeley, CA, USA. [3]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA. [4]Berkeley AI Research, University of California, Berkeley, Berkeley, CA, USA. [5]Department of Medicine, Division of Rheumatology, University of California, San Francisco, San Francisco, California, USA. [6]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA. [7]Bayesian Health, New York, NY, USA. [8]Department of Computer Science, Johns Hopkins University Whiting School of Engineering, Baltimore, MD, USA. [9]Department of Health Policy & Management, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA. [10]Department of Medicine, Johns Hopkins Medicine, Baltimore, MD, USA. [11]Departments of Medicine, Radiology, and Pediatrics, University of California, San Francisco, San Francisco, CA, USA. [12]Scripps Research Translational Institute, La Jolla, CA, USA. [13]Department of Integrative Structural and Computational Biology, Scripps Research, La Jolla, CA, USA. [14]Department of Medicine, University of Chicago, Chicago, IL, USA. [15]Department of Statistics, University of California, Berkeley, Berkeley, CA, USA. [16]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. [17]Center for Data-driven Insights and Innovation, University of California, Office of the President, Oakland, CA, USA. [18]Qualified Health PBC, Palo Alto, CA, USA. [19]These authors contributed equally: Beau Norgeot, Madhumita Sushil. ✉e-mail: miao.brenda1@gmail.com

Published online: 06 February 2025

References

1. Norgeot, B. et al. *Nat. Med.* **26**, 1320–1324 (2020).
2. Moor, M. et al. *Nature* **616**, 259–265 (2023).
3. Gupta, R. et al. *The Berkeley Artificial Intelligence Research Blog* http://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/
4. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. (2023).
5. Shah, N. H. et al. *JAMA* **331**, 245–249 (2024).
6. Balloccu, S., Schmidtová, P., Lango, M. & Dušek, O. Leak, cheat, repeat: data contamination and evaluation malpractices in closed-source LLMs. In *Proc. 18th Conference of the European Chapter of the Association for Computational Linguistics* (EACL, 2024).
7. Miao, B. Y. et al. Preprint at https://doi.org/10.48550/arXiv.2402.03597 (2024).
8. Ayers, J. W., Desai, N. & Smith, D. M. *JAMA* **331**, 639–640 (2024).
9. Mehandru, N. et al. *npj Dig. Med.* **7**, 84 (2024).
10. Van Veen, D. et al. *Nat. Med.* **30**, 1134–1142 (2024).
11. Tu, T. et al. Preprint at https://doi.org/10.48550/arXiv.2401.05654 (2024).
12. Zhao, H. et al. *ACM Trans. Intell. Syst. Technol.* **15.2**, 1–38 (2024).
13. Turpin, M., Michael, J., Perez, E. & Bowman, S. R. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proc. 37th Conference on Neural Information Processing Systems* (NeurIPS, 2023).
14. Mitchell, M. et al. Model cards for model reporting. In *Proc. Conference on Fairness, Accountability, and Transparency* 220–229 (2019).
15. Sendak, M. P., Gao, M., Brajer, N. & Balu, S. *npj Digit. Med.* **3**, 1–4 (2020).

Competing interests

B.Y.M. is an employee at SandboxAQ. I.Y.C. is a minority shareholder in Apple, Amazon, Alphabet and Microsoft. S. Sun is an employee at Ruby Robotics. T.Z. is a medical consultant for Xyla Health. M.G. is an employee of Pfizer. A.J.B. is a co-founder and consultant to Personalis and NuMedii; a consultant to Samsung, Mango Tree Corporation, and in the recent past, 10x Genomics, Helix, Pathway Genomics and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, and Merck and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Facebook, Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, CVS, Nuna Health, Assay Depot, Vet24seven, Regeneron, Sanofi, Royalty Pharma, AstraZeneca, Moderna, Biogen, Paraxel and Sutro, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson & Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. A.J.B. receives royalty payments from Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis; has received research funding from the NIH, Peraton (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal and Progenity. None of these organizations or companies had any influence or involvement in the development of this manuscript. B.N. is a co-founder at Qualified Health PBC. All other authors have no conflicts of interest to disclose.