Letter to the Editor

Jorge Díaz-Garzón Marco*, Deniz İlhan Topcu, Kornelia Galior, Niels Jonker, Abdurrahman Coskun, Berta Sufrate-Vergara, Isabel Moreno-Parro, Elisabet González-Lao, Anna Carobene, William A. Bartlett, Pilar Fernández-Calle, Sverre Sandberg and Aasne K. Aarsand, on behalf of the EFLM Technical Committee Biological Variation Database

Evaluating a custom GPT assistant for critical appraisal of biological variation studies by the BIVAC

https://doi.org/10.1515/cclm-2026-1049
Received June 29, 2026; accepted July 2, 2026;
published online July 14, 2026

Keywords: critical appraisal; GPT assistant; biological variation

To the Editor,

Reliable biological variation (BV) data underpin key interpretive and quality tools in laboratory medicine. However, BV studies remain heterogeneous in design, reporting, and statistical approaches, and BV data derived especially from more historical BV studies may not be fit for purpose. The EFLM Committee on Biological Variation and Technical Committee Biological Variation Database have developed and promoted structured frameworks to improve the quality of available BV data, notably the Biological Variation Data Critical Appraisal Checklist (BIVAC) and the Standard for Reporting Biological Variation Data Studies (STARBIV) [1–3].

The BIVAC consists of 14 quality indicators (QIs), which address key factors in the production and delivery of BV data. These QIs can elicit different scores from A-D, indicating decreasing compliance with the checklist. However, critical appraisal of studies with the BIVAC is time-consuming and also vulnerable to variability in assessment, especially when key methodological details are unclearly or incompletely reported. Recent studies that have tested large language models (LLMs) as assistants for structured appraisal tasks (e.g., risk-of-bias judgements and checklist-based critical appraisal), generally report partial agreement with expert judgements, however, emphasize the need for validation and human oversight [4–6].

In the EFLM Technical Committee for the Biological Variation Database, we have developed a custom GPT assistant intended as a tool to support BIVAC scoring by (i) mapping manuscript text to each QI, (ii) proposing an A–D grade based on the 14 QIs, and (iii) justifying each decision using verbatim quotations from the manuscript. Here we report a pilot validation focused on two properties relevant to decision support: accuracy against independent

***Corresponding author: Jorge Díaz-Garzón Marco**, Department of Laboratory Medicine, La Paz University Hospital, Paseo de la Castellana 261, 28046, Madrid, Spain; and IdiPaz-Hospital La Paz Institute for Health Research, Madrid, Spain, E-mail: jdgmarco@gmail.com. https://orcid.org/0000-0002-3171-1505
Deniz İlhan Topcu, Department of Medical Biochemistry, İzmir City Hospital, İzmir, Türkiye. https://orcid.org/0000-0002-1219-6368
Kornelia Galior, Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA, USA
Niels Jonker, Certe, Wilhelmina Ziekenhuis Assen, Assen, The Netherlands
Abdurrahman Coskun, Department of Medicine, Ruhr University Bochum, Knappschaft Kliniken Universitätsklinikum Bochum, Bochum, Germany; and Department of Medical Biochemistry, School of Medicine, Acıbadem Mehmet Ali Aydınlar University, Istanbul, Türkiye. https://orcid.org/0000-0002-1273-0604
Berta Sufrate-Vergara, Isabel Moreno-Parro and Pilar Fernández-Calle, Department of Laboratory Medicine, La Paz University Hospital, Madrid, Spain; and IdiPaz-Hospital La Paz Institute for Health Research, Madrid, Spain. https://orcid.org/0000-0002-8169-2066 (I. Moreno-Parro)
Elisabet González-Lao, Quality Department, Consorci Sanitari de Terrassa, Barcelona, Spain
Anna Carobene, Servizio Medicina di Laboratorio, Ospedale San Raffaele, Milan, Italy
William A. Bartlett, Undergraduate Teaching, School of Medicine, University of Dundee, Dundee, UK. https://orcid.org/0000-0002-4105-7471
Sverre Sandberg, Norwegian Organization for Quality Improvement of Laboratory Examinations (Noklus), Haraldsplass Deaconess Hospital, Bergen, Norway; and Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway
Aasne K. Aarsand, Norwegian Organization for Quality Improvement of Laboratory Examinations (Noklus), Haraldsplass Deaconess Hospital, Bergen, Norway; and Department of Medical Biochemistry and Pharmacology, Haukeland University Hospital, Bergen, Norway

scoring by experts from the Technical Committee and repeatability of the model's outputs across independent sessions [2].

Our analysis included 10 BV manuscripts not previously scored in the EFLM Biological Variation database (Supplementary file; paper IDs and full references provided). The manuscript selection was designed to include studies from different publication eras and linguistic contexts, incorporating both older and recent papers and one non-English publication.

Independent, manual assessment of each manuscript by the BIVAC was performed by two Technical Committee members, as per the Committee's standard procedures. The non-English paper was appraised by two native Spanish speakers. When the assessors disagreed, discrepancies were discussed, and a third expert assessed these items. The final scorings results were used as reference to which the custom GPT appraisal was compared. For the custom GPT assessment, two members separately used a custom GPT to evaluate each manuscript three times, with each repetition performed in a new chat session to enforce independence [2]. The custom GPT was based on OpenAI's ChatGPT (GPT-5, with the "thinking" reasoning mode explicitly selected). The standardized prompt instructed the model to: read each manuscript; cross-check QI titles and grading rules directly against BIVAC Table 1; score all 14 QIs; provide QI number and exact title for each QI; and justify every score using direct manuscript quotations (kept in the original language), as shown in the Supplementary file.

Results were computed at two levels (accuracy and repeatability) for each manuscript and at QI level. Accuracy was calculated by comparing each QI score produced in each of the six independent ChatGPT runs with the manual human reference score and expressed as the proportion of exact matches across all 14 QIs and all six runs (84 evaluations per manuscript). We also calculated the accuracy for each QI (60 evaluations; 6 evaluations for each paper). Repeatability was assessed using a criterion of perfect reproducibility (PR), defined as identical grading of a given QI across the six independent model runs. At the manuscript level, since BIVAC's overall classification is determined by the lowest score given to any of the 14 QIs, exact manuscript-level agreement required full concordance across all QIs and therefore the final BIVAC grade.

The human reviewers agreed on all QIs for 7 of the reviewed papers, and for three papers, they disagreed on 6 different QIs. (Yang 371: QI6 and QI7; Frankenstein 372: QI14; Cho 395: QI9, QI11 and QI14). Following the assessment of a

Table 1: Overview of the appraised publication with the reference BIVAC score and accuracy and perfect reproducibility of the custom GPT BIVAC scoring [1].

(Paper ID) [2]	Reference BIVAC score	Mean accuracy, %	Perfect reproducibility, %
Carobene et al., 2024, EuBIVAS insulin resistance markers (652)	B1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12, A13, A14	100	100
Dülgeroğlu & ercan, 2023, serum neopterin (641)	A1, A2, A3, A4, A5, B6, A7, A8, A9, A10, A11, A12, A13, A14	100	100
Baysoy et al., 2021, kidney function markers (557)	A1, A2, A3, A4, A5, A6, B7, B8, A9, A10, A11, A12, A13, A14	100	100
Jabor et al., 2024, PIVKA-II (645)	A1, A2, A3, A4, A5, B6, B7, B8, A9, A10, A11, A12, A13, A14	92	71
Frankenstein et al., 2011, hs-troponin T (372)	A1, A2, A3, A4, A5, A6, A7, C8, B9, C10, B11, A12, C13, A14	86	79
Yang et al., 2018, postprandial lipids (371)	A1, A2, A3, A4, A5, A6, C7, A8, A9, C10, A11, A12, C13, A14	86	64
Hill et al., 2014, hemostatic factors (399)	A1, A2, A3, A4, A5, A6, B7, C8, B9, C10, A11, A12, C13, A14	87	71
Cho et al., 2008, androgens in PCOS (395)	A1, A2, A3, A4, A5, A6, C7, C8, B9, C10, B11, C12, C13, A14	82	64
Ross et al., 1988, hematologic parameters (318)	A1, C2, A3, A4, A5, C6, B7, C8, B9, C10, C11, C12, C13, A14	82	36
Ricós & codina, 1989, intraindividual BV (only available in Spanish language) (138)	A1, B2, A3, B4, A5, A6, B7, C8, B9, C10, B11, C12, C13, B14	26	0

[a]Perfect reproducibility was defined as the proportion of BIVAC quality items (A, B or C) for which the custom GPT, assigned identical grades across all independent runs. Accuracy was defined as the proportion of exact grade matches between GPT, outputs and the human reference scoring, averaged across quality items and runs. [b]The number in parentheses refers to the publication ID, in the EFLM, Biological Variation Database.

third reviewer for the discrepant scorings, consensus was achieved for all, with the final scorings shown in Tabel 1. The evaluation by the custom GPT achieved a mean accuracy of 84 %, complete manuscript-level agreement for all 14 QIs for 3 of 10 manuscripts (30 %) (Table 1) and mean PR of 69 %. For three manuscripts reproducibility was 100 %, for five manuscripts between 60 and 79 % and for two manuscripts <50 %: Ross et al., 1988 (36 %) and the Spanish-language Ricós and Codina, 1989 manuscript (0 %).

Regarding the individual QI assessments, accuracy was lowest for QI for normally distributed data (QI 9, 55 %), steady state (QI 7, 72 %), and confidence interval reporting (QI 12, 78 %). In contrast, scale (QI 1) achieved 97 % agreement, while samples (QI 3) reached 90 % and subjects (QI 2) 85 %. PR showed a similar distribution, with lowest result for steady state (QI 7), at 40 % (4/10 manuscripts), followed by normally distributed data (QI 9) and confidence interval reporting (QI 12), both at 50 % (5/10 manuscripts). In contrast, scale reached 90 % reproducibility (9/10 manuscripts), and subjects, samples, measurand, and concentrations each reached 80 % (Supplementary Table 1).

In this study, we explored using a custom GPT assistant to assess BV studies by the BIVAC, as compared to manual scoring by members of the EFLM Technical Committee for the Biological Variation Database. Overall, our results highlight the limitations in using the GPT model for full assessment and decision of the final BIVAC grade. This is as expected given BIVAC's weakest-link rule: a single QI can determine the final BIVAC grade. From this perspective, it implies that a BIVAC-guided GPT may be most valuable used for assessing each QI and highlighting candidate text for each QI. This would thus enable a more rapid manual scoring process, rather than as a standalone classifier of overall BIVAC grade [2].

Regarding the GPT's individual QI analysis and assessment, agreement and PR were lowest for steady state, normal distribution analysis, reported confidence intervals and outlier analysis QIs. All of these QIs require inferential methodological judgement – and unlike descriptive QIs such as subjects and samples – these indicators often depend on whether assumptions were adequately demonstrated, implicitly justified, or statistically tested. Notably, papers published after the introduction of the BIVAC [2] tended to present the applied methodology in a more structured and explicitly criterion-linked BIVAC manner. In contrast, older studies often lack standardized terminology, requiring greater inferential interpretation from text. For a language model operating through pattern recognition rather than statistical reasoning, distinguishing between implicit compliance and formal methodological demonstration may be particularly challenging. Consistent with this, the three manuscripts that achieved complete agreement (100 % accuracy and 100 % reproducibility; papers 557, 641 and 652) were all recently published BV studies (2021, 2023 and 2024) whose methodology was reported in close alignment with the BIVAC criteria, reducing the need for inferential interpretation by the model.

An example of this is assessment of the confidence interval QI. In some studies, confidence intervals were not explicitly reported yet could be derived from the number of results and replicate measurements provided in the article. Under BIVAC criteria, the absence of explicit reporting does not preclude an A score if the information required for calculation is available. However, this distinction between non-reported and non-calculable may be difficult for a language model to interpret consistently, potentially contributing to lower agreement for this particular QI.

When contextualizing our findings within the broader LLM appraisal literature, they align with prior observations that LLM performance is heterogeneous across appraisal domains and that human oversight remains essential. Taneri et al. compared ChatGPT-4o with Cochrane authors' Risk of Bias 2 (RoB 2) tool judgements and reported fair-to-moderate agreement with variability across domains. Di Pumpo et al. emphasized that LLM outputs should be integrated with researcher expertise rather than used autonomously, and Shereefdeen et al. observed variable agreement when applying qualitative appraisal checklists, reinforcing the need for independent verification [4–6].

A practical extension of this work would be the development of a pre-submission checker embedded within the EFLM BV website. In this scenario, authors could upload a manuscript (or draft) and receive a quote-anchored BIVAC worksheet that flagged missing reporting elements and highlighted the few final-grade-driving QIs most likely to determine the overall class. Such a tool would be explicitly advisory and could improve reporting completeness while preserving scientific accountability with authors, reviewers, and the database team [2, 3].

Limitations of this pilot include the small number of manuscripts, reliance on a single prompt and model configuration, and the evaluation of only one LLM (ChatGPT, GPT-5), which limits the generalizability of the findings to other LLM architectures. The use of larger, pre-registered benchmark sets with multiple independent reference scorings, ordinal agreement measures, and comparative validation across different LLMs would strengthen the evidence

base and provide a greater understanding of the strengths and limitations of a custom GPT assistant for assessing BV studies.

In summary, a custom GPT assistant assessing BV studies by the BIVAC achieved acceptable QI-level agreement (accuracy 84 %) and moderate repeatability (69 %), while fully correct manuscript-level classification across all 14 QIs occurred in only 3 out of the 10 evaluated papers when compared to human expert scoring as the reference. These findings suggest that currently, the value of custom GPT tools lies in item-level assistance and transparent evidence extraction rather than autonomous overall grading by the BIVAC.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: We use ChatGPT as a tool to support BIVAC scoring.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

1. Bartlett WA, Braga F, Carobene A, Coşkun A, Prusa R, Fernandez-Calle P, et al. A checklist for critical appraisal of studies of biological variation. Clin Chem Lab Med 2015;53:879–85.
2. Aarsand AK, Røraas T, Fernandez-Calle P, Ricós C, Díaz-Garzón J, Jonker N, et al. The biological variation data critical appraisal checklist: a standard for evaluating studies on biological variation. Clin Chem 2018;64:501–14.
3. Bartlett WA, Sandberg S, Carobene A, Fernandez-Calle P, Diaz-Garzon J, Coskun A, et al. A standard to report biological variation data studies: based on an expert opinion (STARBIV). Clin Chem Lab Med 2024;63:52–9.
4. Taneri PE. Human versus artificial intelligence: comparing cochrane authors' and ChatGPT's risk of bias assessments. Cochrane Evid Synth Methods 2025;3:e70044.
5. Di Pumpo M, Riccardi MT, De Vita V, Damiani G. Evaluation of a large language model (ChatGPT) versus human researchers in assessing risk-of-bias and community engagement levels: a systematic review use-case analysis. Eur J Publ Health 2025;35:1082–6.
6. Shereefdeen H, Thaivalappil A, Young I, MacKay M. A pilot study on generative artificial intelligence's reliability in qualitative research quality appraisal using CASP and JBI checklists. Inquiry 2025;62: 469580251399374.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2026-1049).