OPEN ACCESS

Digital Medicine LETTER

Human-AI collaboration versus expert intuition: a head-to-head comparison in patient classification

Maria Laura Manca,[1] Paola Migliorini,[2,*] Diana Boraschi,[3,4,5,6,*] on behalf of the BioCog Consortium[7]

[1]Department of Clinical and Experimental Medicine and Department of Mathematics, University of Pisa, and Hospital-University of Pisa, Pisa 56100, Italy

[2]Department of Internal Medicine, University of Pisa, and Hospital-University of Pisa, Pisa 56100, Italy

[3]Shenzhen University of Advanced Technology (SUAT), Shenzhen 518055, China

[4]Laboratory of Inflammation and Vaccines, Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science (CAS), and China-Italy Joint Laboratory for Pharmacobiotechnology for Medical Immunomodulation, Shenzhen 518055, China

[5]National Research Council, Napoli 80131, Italy

[6]Stazione Zoologica Anton Dorhn, Napoli 80121, Italy

[7]In addition to Diana Boraschi, the BioCog consortium ("Biomarker Development for Postoperative Cognitive Impairment in the Elderly" grant n. 602461/HEALTH-F2-2014-60246) also includes Friedrich Borchers, Insa Feinkohl, Gunnar Lachmann, Rudolf Mörgeli, Kwaku Ofosu, Maria Olbert, Tobias Pischon, Claudia Spies, Georg Winterer, Alissa Wolf, Fatima Yürek, Norman Zacharias (all Charité–Universitätsmedizin, Berlin), Jürgen Gallinat, Simone Kühn (all University Medical Center, Hamburg), Jeroen de Bresser, Edwin van Dellen, Jeroen Hendrikse, Ilse Kant, Simone van Montfort, Arjen Slooter (all University Medical Center, Utrecht), David Menon, Laura Moreno-López, Jacobus Preller, Emmanuel Stamatakis, Stefan Winzeck (all University of Cambridge), Giacomo Della Camera, Paola Italiani, Daniela Melillo (all National Research Council, Napoli), Roland Krause, Reinhard Schneider (all University of Luxembourg), Karsten Heidtke, Peter Nürnberg (all ATLAS Biolabs GmbH, Berlin), Franz Paul Armbruster, Axel Böcher, Ina Diehl, Thomas Bernd Dschietzig, Bettina Hafen, Katarina Hartmann, Anja Helmschrodt, Marion Kronabel, Jana Ruppert, Marius Weyer (all Immundiagnostik AG, Bensheim), Malte Pietzsch, Simon Weber (all Cellogic GmbH, Berlin), Ariane Fillmer and Bernd Ittermann (all Physikalisch-Technische Bundesanstalt, Berlin)

*Correspondence: paola.migliorini@unipi.it (P.M.); diana.boraschi@gmail.com (D.B.)

Received: September 29, 2025; Accepted: December 19, 2025; Published Online: December 22, 2025; https://doi.org/10.59717/j.xinn-med.2026.100182

Citation: Manca M.L., Migliorini P., Boraschi D (2026). Human-AI collaboration versus expert intuition: a head-to-head comparison in patient classification. *The Innovation Medicine* **4**:100182.

INTRODUCTION

Recent literature emphasizes that successful AI implementation in medicine requires effective human-AI collaboration rather than replacement strategies.[1,2] However, despite the widespread adoption of machine learning in healthcare,[1,3] empirical evidence comparing collaborative versus traditional approaches remains scarce, highlighting the need for more rigorous evaluation.[4] This gap is critical, as clinicians increasingly view AI as "consulting an experienced colleague"[2] rather than an autonomous decision-maker. We present the first documented head-to-head comparison of classification of the same immunological dataset by human-AI collaboration versus expert domain knowledge. This provides empirical evidence for ongoing debates about optimal human-AI integration in clinical practice.

We analyzed the index of immune response (the ratios between inflammatory IL-1β and anti-inflammatory IL-1Ra) by blood leukocytes of 102 elderly surgical patients, measured at baseline (S0) and under increasing inflammatory stimulation with bacterial lipopolysaccharide (S1-S3), both pre-surgery (T1) and three months post-surgery (T3). More in detail, anticoagulated blood was exposed for 24 h to culture medium alone (S0) or to increasing LPS concentrations (0.01, 0.1, 1 ng/mL, i.e., S1, S2, S2). Cell-free supernatants were assessed by ELISA for IL-1β, sIL-1R2 and IL-1Ra. The concentration of free IL-1β (not blocked by sIL-1R2) was calculated by applying the law of mass action, considering a molecular mass of 17 kDa for IL-1β and 47 kDa for sIL-1R2, and a 1:1 interaction with a Kd of 2.7 nM.[5] Free active IL-1β, i.e., the ratio between free IL-1β and IL-1Ra (the IL-1 receptor antagonist) multiplied by 1000, was taken as index of immune response. Multiplication by 1000 was included because blockade of IL-1β activity by IL-1Ra, despite their similar affinity for the the receptor IL-1R1, notably needs about 1000-fold IL-1Ra excess for efficient functional inhibition.

This dataset underwent two independent classification processes:

AI-Assisted Approach: Developed through documented interactions between a biomathematician and Claude (Anthropic's AI assistant, Claude Sonnet 4). The AI-assisted method required a biomathematician capable of iterative dialogue with AI tools, translating statistical outputs into biologically meaningful hypotheses. The iterative process involved statistical clustering suggestions by AI, human interpretation of biological relevance, and refinement based on clinical outcomes.

Expert-Driven Approach: Independent classification by two experienced immunologists (one basic and one clinical) acting in concert, using functional immunological understanding, physiological plausibility, and clinical relevance as primary criteria. Thus, the expert-driven approach demanded deep immunological domain knowledge to immediately recognize functionally relevant patterns.

This comparison intentionally highlights a critical challenge in medical AI implementation: without domain-specific expertise, AI risks applying statistically valid methods to biologically inappropriate contexts. The biomathematician-AI partnership required continuous biological validation, while the immunological expertise enabled direct recognition of clinically meaningful classifications.

The AI-assisted approach yielded five clusters (named based on statistical concepts): Optimal (55 patients, 53.9%), Dysregulated (11, 10.8%), Quiescent (15, 14.7%), High-Baseline (14, 13.7%), and Others (7, 6.9%). The expert-driven approach produced four clusters (named based on immunological/clinical concepts): Normal (50, 49.0%), Anergic (21, 20.6%), Hyperreactive (18, 17.6%), and Inflamed (13, 12.7%) (Figure 1).

Key differences emerged in both philosophy and execution. The AI approach captured statistical patterns and created categories such as "Dysregulated" (bell-shaped response curves that do not perfectly align with the incremental dose-response curves considered as Optimal) and "Others" (statistical outliers that do not fit in any of the other clusters). The expert approach eliminated these ambiguous categories, redistributing patients based on immunological function: "Anergic" (immune/inflammatory hyporesponsiveness) and "Hyperreactive" (excessive immune/inflammatory reactivity).

Notably, the experts applied stricter biological criteria. For "Normal" patients, the AI accepted any S3 response ≥20, while the expert required S3 values between 20 and 80, reducing this group from 55 to 50 patients but achieving greater immunological coherence.

Both classifications demonstrated prognostic value for post-surgical immune recovery (T1→T3 transitions, with 59/102 patients classified as Normal/Optimal at T3), but with important differences:

1. Normal/Optimal patients showed similar retention rates (72.0% vs. 65.5%), but expert-defined categories demonstrated clearer biological trajectories.

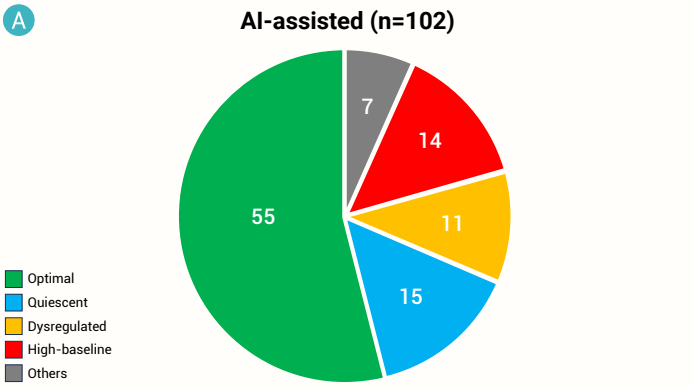
2. The expert "Anergic→Normal" transition (48% recovery) provided more mechanistic insight than the AI "Quiescent→Optimal" (53%), reflecting understood immune reconstitution processes.

Both approaches similarly identified patients with high baseline inflammation (High baseline/Inflamed) as having limited recovery potential (recovery was 3/14 for AI-assisted, 1/14 for expert driven evaluations).

THE HUMAN-AI DIALOGUE

The AI-assisted approach implemented a ***human-in-the-loop process***

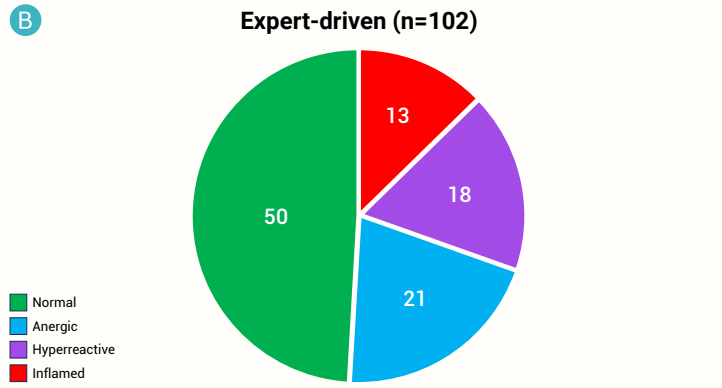
A



Cluster	N	S0	S3	Rationale
Optimal	55	≤5	≥20	Statistical pattern
Quiescent	15	≤5	<20	Low response
Dysregulated	11	≤5	variable	Bell-shaped curve
High-baseline	14	>5	variable	Chronic inflammation
Others	7	variable	variable	Statistical outliers

Stability: 65.5% remain Optimal at T3
Ambiguous catergories: 17.7% (Dysregulated, Others)

B



Cluster	N	S0	S3	Rationale
Normal	50	≤5	20-80	Physiological response
Anergic	21	≤5	≤20	Insufficient reactivity
Hyperreactive	18	≤5	>90	Excessive reactivity
Inflamed	13	>10	variable	Chronic inflammation

Stability: 72.0% remain Normal at T3
Clear biological catergories: 100%

C

Clinical Outcome: T1 to T3 transitions

Optimal (55)	Remain Optimal: 36 (65.5%)
Quiescent (15)	Become Optimal: 8 (53.3%)
Dysregulated (11)	Become Optimal: 6 (54.5%)
High-baseline (14)	Become Optimal: 3 (21.4%)
Others (7)	Become Optimal: 6 (85.7%)
Total achieving Optimal at T3: 59/102 (57.8%)	

Normal (50)	Remain Normal: 36 (72.0%)
Anergic (21)	Become Normal: 10 (47.6%)
Hyperreactive (18)	Become Normal: 12 (66.7%)
Inflamed (13)	Become Normal: 1 (7.7%)
Total achieving Normal at T3: 59/102 (57.8%)	

Figure 1. Comparative analysis of AI-assisted versus expert-driven patient classification Analysis of immune/inflammatory responsiveness was run on blood samples from 102 elderly surgical patients (A) AI-assisted classification yielded five clusters, defined on the basis of statistical pattern recognition. (B) Expert-driven classification produced four clusters, defined with stricter immunological criteria. (C) Clinical outcome showing T1 to T3 transitions demonstrates achieving identical overall success rates (57.8%), but expert classification shows superior stability in optimal/normal patients (72.0% vs. 65.5%). Note the elimination of the ambiguous categories in the expert approach, with patients redistributed based on functional immune status rather than statistical patterns.

characterized by iterative cycles of computational analysis and human interpretation. Documentation of the human-Claude interactions revealed the following workflow:

1) AI Pattern Proposal: Claude generated initial statistical clustering based on numerical patterns of the immune response index data across the 102-patient dataset.

2) Human Biological Evaluation: The biomathematician assessed each proposed cluster for biological plausibility, considering known immunological mechanisms and physiological constraints.

3) Iterative Refinement: Based on biological considerations, the biomathematician reformulated queries to Claude, specifying constraints or requesting alternative clustering approaches. A critical example involved hierarchical organization of clusters to avoid patient overlap - initial "free" clustering by Claude showed classification ambiguities that required human guidance to establish clear decision boundaries.

4) Quality Control: As a consistency test, two patients were intentionally evaluated twice through the AI process; the biomathematician verified that both assessments assigned them to identical clusters, confirming the reproducibility of the human-AI workflow.

5) Clinical Validation: Proposed classifications were evaluated against clinical outcomes (T1→T3 transitions) to assess prognostic value.

While the specific clustering algorithm used internally by Claude Sonnet 4 is not disclosed by Anthropic, the human-in-the-loop process ensured that biological validity had precedence over algorithmic output. The five clusters were defined upon iterative refinement guided by biological plausibility rather than purely statistical optimization. This human-in-the-loop framework proved essential for multiple reasons. First, the large dataset (102 patients, each with 8 measurements) made Claude's automated analysis crucial - manual processing would have been time-consuming and error-prone. Second, while Claude excelled at identifying subtle statistical patterns and maintaining systematic consistency, it occasionally suggested groupings that, while mathematically coherent, lacked biological interpretability (e.g., the "Dysregulated" category based solely on bell-shaped curves without immunological rationale). Third, the biomathematician observed that Claude's

performance improved notably through iterative interaction, suggesting learning from human feedback - though at this stage AI cannot replace human expertise.

Conversely, the immunology expert approach immediately eliminated statistically driven but immunologically unclear categories, demonstrating the irreplaceable value of domain expertise in creating clinically meaningful classifications. However, while the expert approach was still possible with a dataset of 816 data (4 conditions per 2 timepoints for 102 patients), this cannot be easily applied on larger datasets.

Our comparison suggests that neither pure AI nor pure expert approaches represent optimal strategies. Instead, a hybrid model is proposed:

1. AI for Pattern Detection: Computational approaches excel at identifying subtle statistical relationships in complex datasets.

2. Expert Interpretation: Domain knowledge remains essential for biological plausibility and clinical relevance.

3. Iterative Refinement: The combination amplifies strengths while mitigating individual limitations.

This finding has immediate implications for the growing use of AI in precision medicine, biomarker discovery, and patient stratification. As AI tools become more sophisticated, the question shifts from "human or machine?" to "how can human expertise and AI capabilities best collaborate?"

Our results align with emerging evidence that successful AI implementation in medicine requires maintaining human supervision rather than replacing expertise.[2,6] Recent physician surveys emphasize AI's role as a supporting eye and collaborative partner.[2] Previous implementation studies have identified shared decision-making and human supervision as critical factors for successful AI deployment in clinical practice,[7] yet empirical evidence supporting these theoretical positions has been limited.

The documented human-Claude collaboration provides a replicable model for human-AI partnerships in research. Such transparency in AI-assisted scientific processes becomes crucial as recent studies identify physician acceptance[2] and interpretability as primary barriers to clinical AI adoption.

CONCLUSIONS

While AI-assisted classification effectively identified statistical patterns, expert-driven refinement provided superior biological coherence and clinical interpretability. The optimal approach appears to involve AI for comprehensive pattern detection followed by expert refinement based on domain knowledge and clinical relevance. Without domain-specific expertise, AI risks to apply statistically valid methods to biologically inappropriate contexts.

The documented process suggests that optimal results can emerge when AI, biostatistician, and domain experts collaborate as an integrated team. In this configuration, AI amplifies human analytical capacity rather than replacing it. This empirical evidence supports recent theoretical calls for collaborative rather than replacement models in medical AI.[2]

Our findings contribute evidence to ongoing debates about medical AI implementation and demonstrate a replicable methodology for future comparative studies of human-AI collaborative approaches in clinical research.

REFERENCES

1. Basubrin M. (2025). Current status and future of Artificial Intelligence in medicine. *Cureus* **17**:e77561. DOI:10.7759/cureus.77561
2. Heinrichs H., Kies A., Nagel S.K., et al. (2025). Physicians' attitudes toward Artificial Intelligence in medicine: mixed methods survey and interview study. *J. Med. Internet Res.* **27**:e74187. DOI:10.2196/74187
3. Beam A.L. and Kohane I.S. (2018). Big data and machine learning in health care. *JAMA* **319**:1317–1318. DOI:10.1001/jama.2017.18391
4. Rajpurkar P., Chen E., Banerjee O., et al. (2022). AI in health and medicine. *Nat. Med.* **28**:31–38. DOI:10.1038/s41591-021-01614-0
5. Symons J.A., Young P.R. and Duff G.W. (1995). Soluble type II interleukin 1 (IL-1) receptor binds and blocks processing of IL-1 beta precursor and loses affinity for IL-1 receptor antagonist. *Proc. Natl. Acad. Sci. U. S. A.* **92**:1714–1718. DOI:10.1073/PNAS.92.5.1714
6. Vaccaro M., Almaatouq A. and Malone T. (2024). When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat. Hum. Behav.* **8**:2293–2303. DOI:10.1038/s41562-024-02024-1
7. Labkoff S., Oladimeji B., Kannry J., et al. (2024). Toward a responsible future: recommendations for AI-enabled clinical decision support. *J. Am. Med. Inform. Assoc.* **31**:2730–2739. DOI:10.1093/jamia/ocae209

FUNDING AND ACKNOWLEDGMENTS

The original data collection was supported by the BioCog consortium project (602461/HEALTH-F2-2014-60246). DB was further supported by the Gates Foundation (BMGF INV-059115), the National Natural Science Foundation of Guangdong Province (2024A1515012522) and the National Key R&D Program of China (2024YFE0104500). The funding bodies had no role in the design of this comparative analysis, data interpretation, or manuscript preparation.

AUTHOR CONTRIBUTIONS

MLM analyzed the data by statistical methods and interacted with the anthropic's AI assistant Claude. PM analyzed the data from the clinical point of view. DB analyzed the data from the immunological point of view. MLM wrote the manuscript and PM and DB critically revised it.

DECLARATION OF INTERESTS

DB serves as co-Editor-in-Chief of The Innovation Life and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. Other authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

This study represents a retrospective analysis of anonymized data previously collected as part of the BioCog consortium study ("Biomarker Development for Postoperative Cognitive Impairment in the Elderly", grant n. 602461/HEALTH-F2-2014-60246). All original data collection procedures were performed in compliance with relevant laws and institutional guidelines and were approved by the local medical ethics committees of the study centers in Berlin, Germany (EA2/092/14) and Utrecht, The Netherlands (14–469). Informed consent was obtained from all participants for the original data collection. Patient privacy rights were strictly observed throughout the data handling process. As this current analysis involved only retrospective evaluation of fully anonymized datasets, no additional ethics approval was required for the comparative methodology study reported herein.

DATA AND CODE AVAILABILITY

The datasets used in this study are part of the BioCog study. Due to patient privacy constraints and consortium agreements, the raw datasets are not yet publicly available. Before their publication (manuscript in preparation), access to anonymized data may be possible through the BioCog consortium for qualified researchers with appropriate institutional approvals and data use agreements. Requests for data access should be directed to the BioCog consortium leadership through the corresponding authors.