Proceedings of ICMMGH 2027 Symposium: Clinical OMICS for Drug Discovery and Personalized Medicine
DOI: 10.54254/2753-8818/2027.MB39152

Artificial Intelligence in Biosensing and Wearable Health Monitoring

Jinxin Gao

Epsom College Malaysia, Guangzhou, China
amy.gao28@epsomcollege.edu.my

Abstract. Artificial Intelligence (AI) can facilitate the interpretation of data obtained from biosensors and wearable devices. However, the full utility of AI in the measurement and clinical follow-up process is not yet fully realized. This narrative review will explore the applications of machine learning in conjunction with personalized monitoring and multimodal sensing in healthcare. It will cover already published studies in this field and provide examples of signal classification, detection and other opportunities while addressing blind spots in data and coverages, referencing testing and population representation. The review will cover diagnostic accuracy and the need for evidence of improved patient outcomes. The review will also cover a hypothetical example of a screening test to illustrate that in cases of sensitivity and specificity being close to 100%, large numbers of false alerts can still occur. The review aims to communicate that, for clinical utility of such systems, the follow-up and control of model updates will be necessary.

Keywords: Biosensors, Wearable Sensors, Machine Learning, Clinical Validation

1. Introduction

A measurement is clinically useful if it leads to appropriate action. Glucose measurement aids in the management of diabetes. Measurement of an irregular pulse may be suggestive of atrial fibrillation and warrant an electrocardiogram. But between a measurement and action, several opportunities exist for error. The surface used for sensing may wear out. The algorithm may misclassify a signal and an alert may reach an unresponsive service. The V3 framework attempts to address this by separating verification of the sensing technology, analytical validation of the measurements and clinical validation in a defined context of use [1]. In this review, we use this structure to organise our arguments.

This review encompasses a broad interpretation of biosensing to incorporate biochemical measurement and associated physiological monitoring, which rely on different measurement principles. An example of a chemical measurement is an enzymatic glucose sensor. An example of an optical measurement principle is photoplethysmography (PPG), where a change in blood volume is measured, and an example of an electrical measurement principle is electrocardiography (ECG). A device being wearable does not necessarily mean that machine learning is being used within the device. Some systems learn from data to create a model and/or pattern, while others use rules and/or

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2027.MB39152

a combination of conventional signal processing and fixed algorithms. Being able to measure multiple parameters does not mean the device uses AI and/or machine learning.

This review focuses on primary sources relating to machine learning, personalization, and multimodal sensing. It includes research methods, guidance, and regulatory sources, with source checking completed on 10 October 2026. This review provides illustrative examples of study designs to show research gaps and limitations, and draws conclusions from evidence rather than pooling studies. This is not a systematic review. The devices and applications illustrated do not represent the totality in this field. This review does not describe any original clinical data. The examples presented in this review illustrate hypothetical scenarios and, as such, are not clinical data or research findings. The equation representing a scenario where sensitivity and specificity are fixed, and the only variable is the prevalence of a condition, and what the expected outcome would be, is not an empirical finding of a research study.

2. Future directions grounded in current evidence

2.1. Personalised monitoring

Repeated measures may highlight deviations from a subject's typical physiology. Mishra and colleagues, for example, studied smartwatch data from 32 people with COVID-19, and found that 26 of them showed changes in heart rate, activity, or sleep [2]. This study supports investigating within-person changes as an early warning signal. It does not, however, establish that an increased heart rate identifies a particular infection. Other factors (such as strenuous activity, sleep disruption or other illnesses) can result in a similar change, and therefore an alert should be investigated further. Personalised thresholds should allow for ordinary variation, while avoiding the gradual acceptance of clinically important deterioration as the person's new baseline.

There are similar issues around how the performance of a system should be assessed. For example, a model may appear accurate if evaluation admits information from the same person or a later period that would have been unavailable at prediction time. A prospective evaluation needs to specify the baseline learning period, when model updates are allowed, and how alerts will be judged. It should also include people who are unable to provide a long history of reliable measures. Without these safeguards, reported improvements may largely apply to established users. The system may work less well for new or intermittent users, or for people whose health conditions cause rapid physiological changes.

2.2. Multimodal and continuous sensing

Experimental devices are increasingly combining measurements of different substances. Gao and colleagues demonstrated a wearable array that measured sweat glucose, lactate, sodium and potassium as well as temperature [3]. Sempionatto and colleagues combined ultrasound monitoring of blood pressure and heart rate with electrochemical measurements in an epidermal patch [4]. Tehrani and colleagues demonstrated a wearable microneedle array for continuous measurement of glucose plus either lactate or alcohol in interstitial fluid [5]. These devices involve different biological compartments and measurement principles. Their results should not be combined into a claim that one established device can continuously measure all relevant biomarkers.

An additional sensor is useful if it improves interpretation of another measurement. Movement, for example, might help determine when an optical pulse signal is unreliable. Data fusion studies should compare the combined model with each channel in isolation and with a simple baseline,

DOI: 10.54254/2753-8818/2027.MB39152

using the same participants and reference labels. They should also test what happens when a channel fails. Measurements in sweat, interstitial fluid and blood may have different time responses; measurements in each fluid cannot be assumed to reflect the same concentration at a single time point. Synchronisation, calibration and sampling conditions therefore have to be part of the model specification.

2.3. What a wearable atrial fibrillation study actually shows

Van Vliet and colleagues analysed data from 150 participants monitored with a wrist device integrating PPG and single-lead ECG for 28 days [6]. Compared to 24-hour Holter monitoring, the PPG classifier demonstrated a sensitivity and specificity of 98.8% and 98.2%, respectively. However, only 54.9% of the 90-second PPG windows were analyzable. Over the 28-day period, PPG indicated atrial fibrillation (AF) in 26.7% of participants, and single-lead ECG confirmed AF in 19.3% of the participants. The 24-hour Holter monitoring indicated AF in 14.7% of the participants. These percentages refer to different endpoints and observation periods. Therefore, improvements to the algorithm of the PPG classifier cannot completely account for the increased yield. A PPG alert should not be reported as an ECG-confirmed diagnosis. The study was conducted at a single centre and did not include participants with Fitzpatrick skin types V and VI. The authors of the study suggest further validation for extended monitoring.

2.4. Biochemical detection and clinical outcomes

Machine learning can spot patterns in various fields, and biochemical sensor arrays are one example. Kim and colleagues used machine learning to determine patterns in the fluorescent response of modified carbon nanotubes (CNTs) in 269 serum samples from symptomatic individuals. Their classifiers identified high-grade serous ovarian carcinoma with a sensitivity of 87% and a specificity of 98% [7]. While the study demonstrated the diagnostic potential for sampled patients, it did not evaluate the performance of the assay for the screening of asymptomatic patients. The disease prevalence, stage, and competing diagnoses in an asymptomatic population may be different. Nor does it demonstrate a reduction in mortality. Different studies will need to address all of these.

Sensor-guided diabetes care offers a useful methodological comparison. In a six-month randomized study of 168 persons with type 1 diabetes, Brown et al. reported an adjusted improvement of 11 percentage points in time within target glucose range for closed-loop control versus sensor-augmented pump therapy [8]. The study evaluated a complete care system and a clinically meaningful glycaemic endpoint. The trial should not be taken as proof that machine learning always improves biosensors. Closed-loop control and machine learning are not interchangeable. Its relevance here is methodological: sensor systems need evidence about the care they support as well as evidence of measurement accuracy.

3. Data quality and machine learning evaluation

3.1. Reliable measurements and defensible labels

Recognizing biosensor signals requires expertise in the interpretation of values. Raw signals need consistent units, timestamps and calibration data. Reference labels must be linked to an appropriate clinical test at a relevant timescale. A normal laboratory results obtained hours after a wearable alert may not establish whether a brief event occurred earlier. A sample may be labelled only by a

DOI: 10.54254/2753-8818/2027.MB39152

patient's eventual diagnosis, which may not be useful for training an early warning classifier. The reference method, timing and adjudication process must be specified before evaluating the classifier.

Many important decisions in quality control may influence what information will be reported. Filtering may remove noise, but it can also alter clinically important signal characteristics. If many windows of uncertainty are excluded, it may improve the classifier's accuracy, but leave many time periods with no signal. Reports should state how much usable data remained, the reasons for discarding data, and how missing data were distributed across participants and treated. Missing values should not be treated as normal readings. Stating that no reliable reading is available can be more useful than providing a falsely confident estimate.

Machine learning models may mitigate reproducible measurement drift or detect common artefacts when trained and tested against an appropriate reference. This gives a more precise account of their role than claiming either that AI can fix all measurement errors or that it cannot improve a noisy signal. Correction models have a limited range of validity and need testing across sensor age, operating conditions and manufacturing batches. If the sensor no longer responds meaningfully to the target, the model cannot recover the missing clinical information simply by producing a plausible number. It is the responsibility of a monitoring system to distinguish a measurement failure from valid, albeit unusual, physiology.

3.2. Model choice and independent evaluation

Model choice is dependent on the prediction task and the evidence. Regularised regression provides a reference model for structured features. Tree ensembles and support vector machines are candidates when nonlinear relationships seem plausible. Considering waveforms or other longitudinal data, neural networks may be an appropriate choice. Models of any architecture may be prone to overfitting. A more complex model should show a useful improvement under the same evaluation conditions, with attention to uncertainty, computational demands and failure modes. Reporting only the highest score from several attempted models conceals the degree of model selection.

Partitions of training and test data should occur at the level relevant to clinical deployment. Data from a few patients, even thousands of windows from each patient, should not be considered the same as data from thousands of independent patients. To assess performance in new users, all data from an individual should be in the same split. For a forecasting task, the test data should follow the training data, and the preprocessing should avoid using future data. Feature selection, scaling and hyperparameter tuning belong inside the training process. External testing at another institution or on another device should assess whether the full pipeline works beyond its original development environment.

The TRIPOD + AI statement offers guidance for the reporting of clinical prediction models based on regression and machine learning [9]. A transparent report for a biosensor study should enable tracing of the target outcome, population, input, processing, model development and evaluation. Exclusions and how unavailable measurements were handled should be reported. Making code and permitted data available allows other researchers to scrutinise the analysis. When open sharing is not possible, controlled access may provide an alternative. Reporting guidelines cannot ensure model quality. A well-reported study can still have weak reference labels, a narrow sample and/or an unsuitable comparator.

DOI: 10.54254/2753-8818/2027.MB39152

3.3. Metrics that reflect clinical consequences

Accuracy alone can be misleading when dealing with rare conditions. With a prevalence of 1%, an always-negative classifier will have an accuracy of 99% but will not detect a single case. Sensitivity is the proportion of affected cases detected, while specificity is the proportion of unaffected cases correctly classified. Positive predictive value (PPV) asks how many positive results are true positives. These measures need to be related to the unit of analysis. Accuracy for a signal window, an episode and a patient describe different aspects of performance.

Assume a hypothetical test has 95% sensitivity and 98% specificity. If 10,000 individuals are tested once at 1% prevalence, scientists would expect 95 true positives and 198 false positives. This would result in a PPV of 32.4% (95 divided by 293). Using the same sensitivity and specificity at 10% prevalence would give a PPV of 84.1%. Table 1 shows these calculations. They demonstrate the impact of prevalence assuming constant sensitivity and specificity and should not be seen as the performance measure of any of the devices presented here. A practical assessment would also quantify the burden of repeated alerts, confirmatory testing and missed clinically important events.

Table 1. Hypothetical screening results per 10,000 people

Prevalence	True positives	False positives	PPV
1%	95	198	32.4%
5%	475	190	71.4%
10%	950	180	84.1%

Illustrative calculation assuming 95% sensitivity, 98% specificity, one independent test per person and complete reference classification. No clinical dataset was used.

Risk prediction models also need calibration. When predicted risk is 20%, the observed event rate should be close to 20% among comparable predictions in the relevant population. Van Calster and colleagues explain why good discrimination does not necessarily mean that predicted probabilities are accurate [10]. Models intended for practical use should be evaluated for both discrimination and calibration, with confidence intervals and performance reported at the threshold actually used. This threshold needs a clinical justification: the consequences of a missed event differ from those of an unnecessary repeat test or an invasive follow-up investigation.

4. Clinical and ethical considerations

4.1. Bias across the measurement pathway

Bias can also exist before an algorithm receives any input from a measurement device. In a retrospective study, Sjoding and colleagues found occult hypoxaemia more often in Black than White patients whose pulse oximeter readings were between 92% and 96% [11]. The paper discusses pulse oximetry and recorded racial categories. It does not establish that other optical wearables share the same error pattern, nor that race is a direct measure of skin pigmentation. Yet, it does highlight the need to evaluate measurement performance in the intended population. Relevant factors can include directly measured skin pigmentation, age, perfusion, clinical condition, and device placement.

Fairness also depends on the relationship between the data and the model. Elgendi et al. evaluated ten classifiers on three medical datasets with controlled sex-ratio sampling scenarios [12]. Performance disparities varied greatly among datasets, and balanced sampling did not address all of

DOI: 10.54254/2753-8818/2027.MB39152

them. Demographic balancing alone may therefore be inadequate. The numerical results are specific to the datasets, models and fairness metric investigated. When applying biosensors to a new application, the study should analyse the model errors for subgroups, as well as the amount of usable data, with uncertainty reporting when a subgroup contains few people.

4.2. Explanations and responsibility for alerts

Explanations should enable clinicians to comprehend the evidence informing an alert, for example, the relevant waveform, its quality, the time of change, and the values that led to the output. A feature-importance visualisation does not mean a model has learned a causal relationship or that its prediction is accurate. Clinical evaluation should examine how people use the information, including whether explanations encourage inappropriate confidence. DECIDE-AI offers instructions for reporting early live assessments of AI decision support, emphasising the interaction between the system, its users and the clinical setting [13].

Including a clinician in the workflow does not automatically establish responsibility for an alert. The service must identify who reviews it, how quickly they do so and what happens if the patient cannot be reached. Users need a clear understanding of the product's role: does it provide general wellness information, support diagnosis or control treatment for a defined indication? Automated treatment changes need their own validation and safeguards. A general sensor alert should not become unsupervised medication advice. Evaluation should cover response times, unnecessary contacts, treatment errors and patients who stop using the system.

4.3. Privacy and regulatory oversight

Continuous health monitoring can generate detailed information about personal health and behavior. European Commission guidance identifies health data as special category data requiring additional conditions for processing under Article 9 of the General Data Protection Regulation [14]. A biosensor service needs to communicate what personal data is collected, the purpose of processing, who has access to the data, and the retention period. A range of controls include processing locally, encryption, and access control. Use of audit trails can help detect, document and address issues in real time. Local processing can help address some data transfer concerns; however, the device and any software updates also need to be protected.

Regulatory requirements vary based on intended use and jurisdiction. For example, in August 2025, the US Food and Drug (FDA) Administration issued guidance describing predetermined change control plans for Artificial Intelligence/Machine Learning (AI/ML) -based software for medical devices [15]. These plans describe anticipated changes to the software, how they are developed and validated and how their effects are evaluated. The FDA reviews change control plans to assess the manufacturer's ability to ensure ongoing clinical reliability, and approves plans accordingly. For a biosensor, a software update may change the frequency of an alert, the threshold for an alert, or how the sensor performs in relation to a subpopulation, and may be considered a change in clinical reliability.

In Europe, the European Commission amended the AI Act in May 2023, and extended the application date for most of its requirements to 2 August 2028. From 10 October 2026, the European Commission's guidance states that the timetable and prioritization of the conformity assessment of medical devices incorporating AI systems included in the scope of the Act, will be the same as for high-risk AI systems [16]. There are earlier deadlines for other requirements. These dates do not change current obligations for conformity of medical devices. Developers must determine which

DOI: 10.54254/2753-8818/2027.MB39152

provisions apply to their intended device and monitor updates to the implementation timeline. A single deadline does not cover all AI applications.

5. From laboratory research to routine care

5.1. Validation matched to intended use

The path of clinical use should start with a specific problem, such as whether the system can identify the attacks that need to be confirmed in the target outpatient population. Bench testing should confirm that the sensor performs in its intended environment. Analytical studies compare the derived measurement with an appropriate reference; clinical studies establish whether that measurement relates meaningfully to the intended condition or outcome [1]. With the idea of improving nursing, prospective comparative research should determine what happens when clinicians respond to the results. A good classifier alone does not justify a decrease in enrollment, an increase in survival or a decrease in overall costs.

Usability also needs to be demonstrated. The V3+ framework extends the original V3 components to include usability validation [17]. For wearable systems, relevant tasks include applying devices, maintaining contact, charging them, and understanding their information. The test should involve people who actually use the system, including those with limited numerical experience or related physical and sensory disabilities. If patients can't get reliable readings at home, the elegant interface is useless. The study should record technical failures and exits while successfully using them, not just report participants who complete each measurement.

5.2. Reliability, cost and implementation

Manufacturing and maintenance can affect performance after a laboratory prototype has been validated. Concerns include variation between sensor batches, ageing reagents and contamination of contact surfaces. Engineering evaluation should therefore consider storage conditions, repeated use where relevant, expected wear duration and replacement procedures. One practical design approach is to separate reusable electronics from a disposable microneedle array, as demonstrated by Tehrani and colleagues [5]. The cost of specific services still depends on the quality of production, consumables, training and support. Promising materials or manufacturing processes alone do not prove that clinical products are affordable.

Cost analysis should include the whole care pathway and the work generated by the sensor. An inexpensive wearable may create substantial follow-up costs if its alerts are poorly targeted. A more expensive system, on the other hand, may be justified if it consistently improves management and reduces other burdens. The study should take into account the efforts of confirmation tests, staff review time, replacement equipment and patients. Remote monitoring also needs to make arrangements for people who are disconnected and do not have compatible mobile phones. These factors help researchers determine whether technically effective equipment can support consistent care beyond a well-supported research environment.

6. Conclusion

Biosensors and wearable systems can perform repeated physiological measurements, and selected studies show that algorithms can extract useful diagnostic information from these measurements. The evidence of different applications is still unbalanced. Multimodal prototypes explore technical

DOI: 10.54254/2753-8818/2027.MB39152

possibilities; diagnostic research to check the performance of specific groups of people; compare what happens when clinical trial tests use a complete system. Each type of evidence answers different questions. The next stage of development needs to link reliable sensing with independent assessment and feasible clinical responses. An effective artificial intelligence biosensor will produce measurements and predictions that, as well as the actions they support, will remain reliable for the people and the environment who use them.

References

[1] Goldsack, J. C., Coravos, A., Bakker, J. P., et al. (2020). Verification, analytical validation, and clinical validation (V3): The foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *npj Digital Medicine, 3*, 55. https://doi.org/10.1038/s41746-020-0260-4

[2] Mishra, T., Wang, M., Metwally, A. A., et al. (2020). Pre-symptomatic detection of COVID-19 from smartwatch data. *Nature Biomedical Engineering, 4*, 1208–1220. https://doi.org/10.1038/s41551-020-00640-6

[3] Gao, W., Emaminejad, S., Nyein, H. Y. Y., et al. (2016). Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature, 529*, 509–514. https://doi.org/10.1038/nature16521

[4] Sempionatto, J. R., Lin, M., Yin, L., et al. (2021). An epidermal patch for the simultaneous monitoring of haemodynamic and metabolic biomarkers. *Nature Biomedical Engineering, 5*, 737–748. https://doi.org/10.1038/s41551-021-00685-1

[5] Tehrani, F., Teymourian, H., Wuerstle, B., et al. (2022). An integrated wearable microneedle array for the continuous monitoring of multiple biomarkers in interstitial fluid. *Nature Biomedical Engineering, 6*, 1214–1224. https://doi.org/10.1038/s41551-022-00887-1

[6] van Vliet, M., Aalberts, J. J. J., Hamelinck, C., et al. (2025). Ambulatory atrial fibrillation detection and quantification by wristworn AI device compared to standard holter monitoring. *npj Digital Medicine, 8*, 177. https://doi.org/10.1038/s41746-025-01555-9

[7] Kim, M., Chen, C., Wang, P., et al. (2022). Detection of ovarian cancer via the spectral fingerprinting of quantum-defect-modified carbon nanotubes in serum by machine learning. *Nature Biomedical Engineering, 6*, 267–275. https://doi.org/10.1038/s41551-022-00860-y

[8] Brown, S. A., Kovatchev, B. P., Raghinaru, D., et al. (2019). Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes. *New England Journal of Medicine, 381*, 1707–1717. https://doi.org/10.1056/NEJMoa1907863

[9] Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378

[10] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, 230. https://doi.org/10.1186/s12916-019-1466-7

[11] Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial bias in pulse oximetry measurement. *New England Journal of Medicine, 383*, 2477–2478. https://doi.org/10.1056/NEJMc2029240

[12] Elgendi, M., Elkhalifa, A., Khandoker, A., et al. (2026). Dataset effects outweigh algorithmic effects in determining fairness of healthcare machine learning. *npj Digital Medicine, 9*, 571. https://doi.org/10.1038/s41746-026-02723-1

[13] Vasey, B., Nagendran, M., Campbell, B., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine, 28*, 924–933. https://doi.org/10.1038/s41591-022-01772-9

[14] European Commission. (2026, October 10). Information for individuals. *Data protection guidance*. https://commission.europa.eu/law/law-topic/data-protection/information-individuals_en

[15] US Food and Drug Administration. (2025, August). Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. *Final guidance*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence

[16] European Commission. (2026, October 10). Navigating the AI Act. *Implementation guidance*. https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act

[17] Bakker, J. P., Barge, R., Centra, J., et al. (2025). V3+ extends the V3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. *npj Digital Medicine, 8*, 51. https://doi.org/10.1038/s41746-024-01322-2