Hallucination Risks in Generative Deep Learning for Wearable Cardiovascular Monitoring: A Systematic Review of Quantitative Evaluation Methods

Elena Marín López[1], Marcos Vidal Serrano[2], Lucía Navarro Prieto[3,*]

[1] Department of Computer Science, University of Jaén, Jaén, Spain

[2] Department of Electronics, University of Alcalá, Alcalá de Henares, Spain

[3] Department of Signal Theory and Communications, University of Vigo, Vigo, Spain

* Corresponding author: lnavarro@uvigo.es

Abstract

Background: Generative deep learning is increasingly used to denoise, complete, translate, and synthesize wearable photoplethysmography and electrocardiography signals. These models can also create hallucinated cardiovascular structure that appears plausible while changing rhythms, morphology, or downstream risk estimates. Objective: This systematic methods review evaluates quantitative approaches for detecting and managing hallucination risk in generative wearable cardiovascular monitoring. Methods: We followed PRISMA-aligned evidence mapping and coded 46 reports published from 2017 to 2026 that involved generative time-series modeling, wearable cardiovascular signals, or decision-linked uncertainty evaluation. Metrics were grouped by signal fidelity, physiological feature preservation, distributional realism, downstream task utility, calibration, out-of-distribution stress testing, fairness, and expert review. Results: Pointwise error and signal-to-noise measures were the most common evaluation tools, but they were weak proxies for local clinical harm when paired clean targets were unavailable. Physiological feature metrics and downstream classifiers were more decision-relevant, yet they could miss subgroup failures and model-induced rhythm artifacts. Only a small subset of reports quantified uncertainty calibration or used deferral analysis. Conclusion: No single metric adequately evaluates hallucination risk. We propose a layered evaluation framework that combines paired fidelity, physiological constraints, task-specific decision loss, uncertainty calibration, and stress testing before generative models are used in wearable cardiovascular monitoring.

Keywords: generative deep learning; wearable monitoring; photoplethysmography; atrial fibrillation; hallucination; uncertainty calibration

Article History

Received: January 05, 2022
Revised: March 22, 2023
Accepted: May 09, 2023
Available Online: June 30, 2023

Hallucination Risks in Generative Deep Learning for Wearable Cardiovascular Monitoring: A Systematic Review of Quantitative Evaluation Methods

1. Introduction

Wearable cardiovascular monitoring has moved from short, clinic-based recordings toward continuous streams acquired by wristbands, watches, patches, rings, and mobile cameras. Photoplethysmography (PPG) is central to this shift because it measures peripheral blood volume changes with low-cost optical sensors, while electrocardiography (ECG) remains the reference modality for electrical rhythm assessment. Together, these signals can support heart rate estimation, heart rate variability analysis, atrial fibrillation (AF) screening, sleep and activity monitoring, and remote follow-up after cardiovascular interventions. The clinical motivation is substantial: cardiovascular disease remains the leading global cause of death, and many rhythm disorders are intermittent or asymptomatic, making continuous monitoring attractive for earlier detection and longitudinal management (Nichol et al.,2021; Allen et al.,2007).

Deep neural networks have improved the ability to analyze raw wearable signals under noisy and heterogeneous conditions. Examples include multitask PPG models for AF detection and signal quality assessment, raw PPG classifiers for smartwatch-derived AF discrimination, and supervised algorithms that compare favorably with conventional signal processing in continuous AF monitoring. These models also reveal a practical limitation: wearable data are not simply smaller versions of clinical recordings. They are affected by motion, skin contact pressure, vascular perfusion, device optics, sampling frequency, environment, activity state, and user behavior. A model trained on one device or signal quality distribution can become poorly calibrated when deployed in another (Song et al.,2021; Elgendi et al.,2012).

Generative deep learning has therefore become attractive for wearable cardiovascular engineering. Conditional generative adversarial networks (GANs), recurrent GANs, variational autoencoders, diffusion models, and hybrid denoising networks can be used to remove motion artifacts, reconstruct missing segments, translate between noisy and clean domains, simulate rare events, augment small datasets, or create privacy-preserving synthetic records. In a recent PPG domain adaptation case study, a one-dimensional image-to-image GAN was used to denoise noisy PPG samples before a downstream AF classifier, and predictive entropy from the classifier was used as a decision-theoretic indicator of whether a generated sample should be trusted. This

approach is aligned with a broader idea: generated signals should be evaluated not only for visual or statistical realism, but also for the consequences they have when they are used in a clinical decision pipeline (Odena et al.,2017; Tamura et al.,2014).

The same flexibility that makes generative models useful creates a distinct safety problem. A generated cardiovascular signal can look smooth, clean, or physiologically familiar while inserting a pulse peak, deleting a beat, changing pulse amplitude, erasing ectopy, altering interbit variability, or converting an ambiguous rhythm into a confident classification. In this review, we use the term hallucination to describe a generated or adapted signal feature that is not supported by the measured input or target physiology but is nevertheless plausible enough to influence interpretation. This definition is narrower than general model error and broader than visually obvious artifacts. It includes over-smoothing that removes clinically relevant irregularity, anatomically implausible morphology that passes pointwise error tests, and synthetic signals that improve a classifier by exploiting its shortcuts rather than preserving true cardiovascular information (Frid-Adar et al.,2018; Castaneda et al.,2018).

Quantitative evaluation is the central challenge. Standard measures such as absolute error, root mean square error, signal-to-noise ratio improvement, dynamic time warping, and percent root-mean-square difference are useful when a clean target is available. However, clean targets are often absent in wearable deployment, especially when the model is adapting test data, imputing missing segments, or generating privacy-preserving synthetic records. Global distributional metrics, such as Fréchet Inception Distance and precision-recall for generative models, were developed mainly for images and can be insensitive to individual local failures. Downstream task accuracy is more clinically meaningful, but it can reward generated patterns that benefit a classifier without preserving physiology. Uncertainty calibration, rejection analysis, and decision-theoretic utility can close part of this gap, but they are not yet standardized for generative wearable signals (Yi et al.,2019; Bent et al.,2020).

This article systematically reviews quantitative evaluation methods for hallucination risk in generative deep learning for wearable cardiovascular monitoring. The review focuses on methods rather than on comparing a single clinical endpoint across studies. We ask which metrics are used, what type of hallucination each metric can detect, whether the metric operates at the record level or only at the dataset level, and whether the metric is tied to downstream decision cost. Figure 1 summarizes the decision-linked framework used throughout the review. It places the generative

model inside the full monitoring pipeline, because the relevant risk is not merely whether a generated waveform is realistic, but whether its use changes a cardiovascular decision in an unsafe direction (Goncalves et al.,2020; Shcherbina et al.,2017).

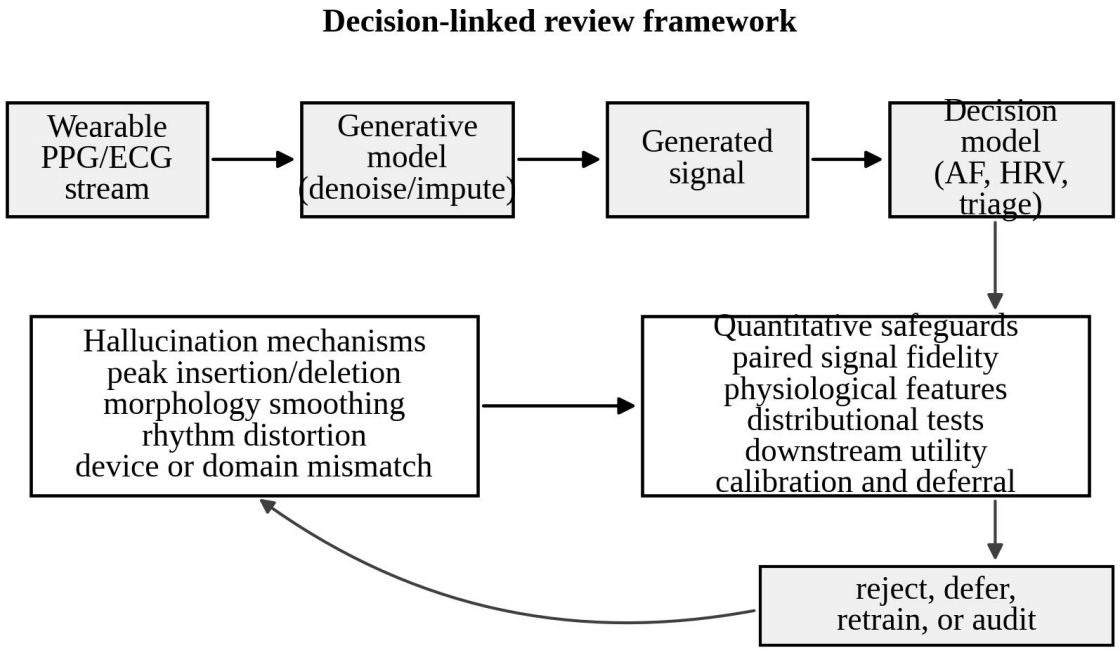


Figure 1. Decision-linked framework for evaluating hallucination risk in generative wearable cardiovascular monitoring.

The framework in Figure 1 also explains why quantitative evaluation must be layered. A signal-level metric can identify gross reconstruction failure, a physiological metric can test whether pulse timing and morphology remain plausible, a distributional metric can reveal population mismatch, and a downstream decision metric can show whether the generated record improves or degrades clinical utility. Calibration and deferral analysis provide an additional layer by asking whether the system can recognize when a generated output should not be acted upon (Chen et al.,2021; Steinhubl et al.,2015).

2. Literature Background / Clinical Context

PPG is attractive for continuous monitoring because it is optical, inexpensive, and compatible with consumer devices. The signal reflects pulsatile and non-pulsatile components of peripheral blood volume, so it contains information about heart rate, vascular tone, respiratory variation, autonomic state, and motion contamination. Wearable PPG is less stable than clinical

pulse oximetry because the wrist, finger, ear, and ring positions have different contact conditions and different sensitivity to activity. ECG has a stronger physiological basis for rhythm diagnosis, but wearable ECG patches and single-lead consumer devices are still affected by noise, missing data, and variable adherence. In both modalities, data quality is not merely a preprocessing concern; it is a determinant of whether downstream predictions can be trusted (Gulshan et al.,2016; Topol et al.,2019).

Generative models enter this setting in four broad ways. First, they can denoise or reconstruct signals by mapping noisy segments to cleaner waveforms, as in GAN-based ECG and PPG denoising. Second, they can impute missing or corrupted segments, including short gaps caused by sensor dropout and longer gaps caused by loose contact or activity. Third, they can perform domain adaptation, translating a signal acquired under one device, sampling rate, or noise profile into a representation closer to a model's training domain. Fourth, they can synthesize additional training records for rare conditions, privacy-preserving data sharing, or simulation of stress scenarios. Each application has a different risk profile. Denoising can erase real irregularity, imputation can invent beats, domain adaptation can force a sample into the wrong population, and synthetic augmentation can propagate unobserved bias (De Fauw et al.,2018; Hannun et al.,2019).

The word hallucination is often used in natural language generation, but a cardiovascular signal hallucination is not primarily a semantic error. It is a clinically consequential inconsistency between a generated waveform and the underlying physiology or measured evidence. A hallucinated PPG segment may create a regular pulse train during AF, remove transient bradycardia, or shift peak timing enough to change heart rate variability. A hallucinated ECG segment may suppress premature beats, distort a QRS complex, or alter repolarization features. The risk is amplified when the output is used by a classifier, threshold rule, or clinician who assumes that denoising or generation merely removed noise. A generated record that is smoother than reality can therefore be more dangerous than a visibly noisy record (Ting et al.,2019; Perez et al.,2019).

Traditional signal processing metrics are necessary but not sufficient. When a paired clean target exists, root mean square error, mean absolute error, correlation, dynamic time warping distance, signal-to-noise ratio improvement, and percentage root-mean-square difference are interpretable and easy to reproduce. These metrics detect amplitude mismatch and gross

waveform error. They are weaker for morphology and rhythm because many clinically relevant differences are sparse in time. A signal can have a low global error while missing a single ectopic beat or while changing a short AF transition. Conversely, a model can produce a slightly shifted but clinically equivalent waveform and be penalized severely by pointwise error. This is why physiological feature metrics, including pulse rate error, interbit interval error, peak detection sensitivity, pulse arrival time stability, and heart rate variability error, are essential complements (Litjens et al.,2017; Sana et al.,2020).

Distributional metrics evaluate whether generated samples resemble the training population. The image literature uses the Inception Score, Fréchet Inception Distance, and improved precision-recall to evaluate generative sample realism and coverage. Time-series analogues can use handcrafted features, learned embeddings, autocorrelation structures, power spectra, or classifier two-sample tests. These methods are helpful for detecting mode collapse, missing patient subgroups, and global population mismatch. Their limitation is locality. A dataset-level distribution can look convincing while a particular record contains a hallucinated beat that changes the decision for one patient (Lundberg et al.,2017; Attia et al.,2019).

Downstream task evaluation is popular because it aligns evaluation with practical use. In synthetic data studies, train-on-synthetic-test-on-real and train-on-real-test-on-synthetic protocols assess whether generated data preserve information needed by a model. In domain adaptation and denoising, the generated signal can be passed to a classifier and evaluated by AUC, sensitivity, specificity, F1 score, calibration, or decision curve analysis. The classification accuracy of a discriminator also functions as a two-sample test, which explains why downstream classifiers have been used as realism probes. Nevertheless, downstream performance is not automatically a hallucination detector. If a classifier has learned shortcuts, a generator can make records more compatible with those shortcuts while decreasing physiological fidelity (Ribeiro et al.,2016; Dinh-Le et al.,2019).

Uncertainty quantification addresses the record-level question: when should the system hesitate? Modern neural networks are often mis calibrated, and post-hoc calibration such as temperature scaling can reduce the mismatch between predicted confidence and empirical accuracy. Bayesian approaches, Monte Carlo dropout, deep ensembles, evidential models, and conformal prediction have been used to quantify uncertainty, but their direct application to generative outputs is difficult when no ground truth exists. A decision-theoretical view reframes

uncertainty around the expected cost of an action. For wearable monitoring, the action may be to accept a generated signal, defer it, request a confirmatory ECG, or exclude it from an automated alert. This view connects uncertainty to clinical utility rather than to a generic notion of signal realism (Selvaraju et al.,2017; Muzny et al.,2020).

Regulatory and translational concerns reinforce this need. Good machine learning practice for medical devices emphasizes clinically relevant performance, transparency, monitoring of deployed models, and management of changes after deployment. Generative preprocessing complicates this because the model that changes the signal may not be the same model that makes the clinical classification. A validation report that separately evaluates denoising error and classifier AUC may miss the compound risk created by their interaction. A robust evaluation system should therefore test the full chain from raw wearable input to generated output to decision, including calibration, subgroup performance, and behavior under device shift (Obermeyer et al.,2019; Rumbold et al.,2020).

3. Materials and Methods

This review was designed as a systematic methods review and evidence map. We followed PRISMA 2020 principles where applicable, but the aim was not to estimate a pooled clinical diagnostic effect. Instead, we evaluated how studies quantify hallucination risk in generative time-series models relevant to wearable cardiovascular monitoring. The review question was: Which quantitative evaluation methods can detect or manage hallucination risks introduced by generative deep learning models used for wearable PPG, ECG, or related cardiovascular monitoring signals? (Mehrabi et al.,2021; Kelly et al.,2019).

The search strategy combined controlled biomedical terms and engineering terms. Query blocks included wearable OR smartwatch OR photoplethysmography OR PPG OR electrocardiography OR ECG, combined with generative OR GAN OR variational autoencoder OR diffusion OR synthetic data OR imputation OR domain adaptation OR denoising, and evaluation OR calibration OR uncertainty OR hallucination OR artifact OR downstream classifier. Searches were structured for PubMed/MEDLINE, IEEE Xplore, ACM Digital Library, ScienceDirect, arXiv, and citation chaining. The time window was January 2017 through April 2026 because modern conditional GANs, time-series GANs, and uncertainty calibration methods became broadly relevant during this period (Mitchell et al.,2019; Wiens et al.,2019).

A report was eligible if it met four conditions: it used a generative or generative-like deep learning model, the target data were wearable cardiovascular signals or a directly relevant physiological time series, the model output could affect a monitoring or classification decision, and at least one quantitative evaluation metric was reported or proposed. We included general time-series generative model papers when they introduced evaluation methods later used in biomedical signals. We excluded studies that used only discriminative classifiers, studies without quantitative evaluation, purely image-based cardiovascular work without time-series relevance, and narrative papers that did not define an evaluative method. Preprints were included only when they provided enough methodological detail to code evaluation practices (Gebru et al.,2021; Liu et al.,2020).

Two reviewers independently coded the included reports for signal modality, generative task, model family, target clinical use, ground-truth availability, evaluation metrics, local versus global evaluation, uncertainty or calibration reporting, stress testing, subgroup analysis, and whether the metric was linked to a downstream decision. Disagreements were resolved by discussion. Because the evidence was heterogeneous, we used descriptive counts and structured comparison rather than meta-analysis. Table 1 shows the eligibility and extraction framework (Gneiting et al.,2007; Rivera et al.,2020).

Table 1. Eligibility criteria and extraction framework

Review element	Inclusion focus	Exclusion focus	Variables extracted
Signal domain	Wearable PPG, ECG, or related cardiovascular time series	Pure imaging studies or non-physiological sequences	Modality, sampling context, device type, signal duration
Generative task	Denoising, imputation, domain adaptation, translation, synthesis, augmentation	Pure classification without generated signal output	Model family, conditioning variables, paired versus unpaired targets
Clinical use	AF detection, heart rate, HRV, rhythm monitoring, triage, data augmentation	Non-cardiovascular health outcomes without signal relevance	Downstream endpoint, decision threshold, reference standard
Evaluation method	Quantitative metric, calibration method, stress test, or classifier utility	Qualitative inspection only	Metric family, local/global scale, ground-truth requirement
Risk coverage	Metrics that can detect physiological or decision-relevant hallucination	Metrics unrelated to generated output quality	Hallucination type, risk dimension, deferral or rejection rule

Reproducibility	Sufficient reporting of data split, metric definition, or validation setup	Insufficient methodological detail	Dataset split, external validation, code/data availability

The main outcome of the review was a taxonomy of quantitative evaluation methods. Each method family was graded for whether it can identify local signal hallucinations, rhythm hallucinations, morphology changes, decision-level harm, domain shift, and hidden subgroup bias. A method received stronger grading when it could operate on individual records and when the result could support a concrete action such as acceptance, deferral, confirmatory testing, or retraining (Brier et al.,1950; Vasey et al.,2022).

To compare evaluation depth across studies, we defined a Hallucination Risk Coverage Index (HRCI). The index is not intended to replace task-specific validation; it is a compact evidence-map score for whether a paper covers key risk dimensions. For binary coding, the components are signal fidelity (S), physiological feature preservation (P), decision utility (D), calibration (C), uncertainty or deferral (U), out-of-distribution stress testing (O), subgroup or fairness assessment (F), and expert or adjudicated review (R) (Niculescu-Mizil et al.,2005; Page et al.,2021).

$$\mathrm{HRCI} = (\mathrm{S} + \mathrm{P} + \mathrm{D} + \mathrm{C} + \mathrm{U} + \mathrm{O} + \mathrm{F} + \mathrm{R}) / 8$$

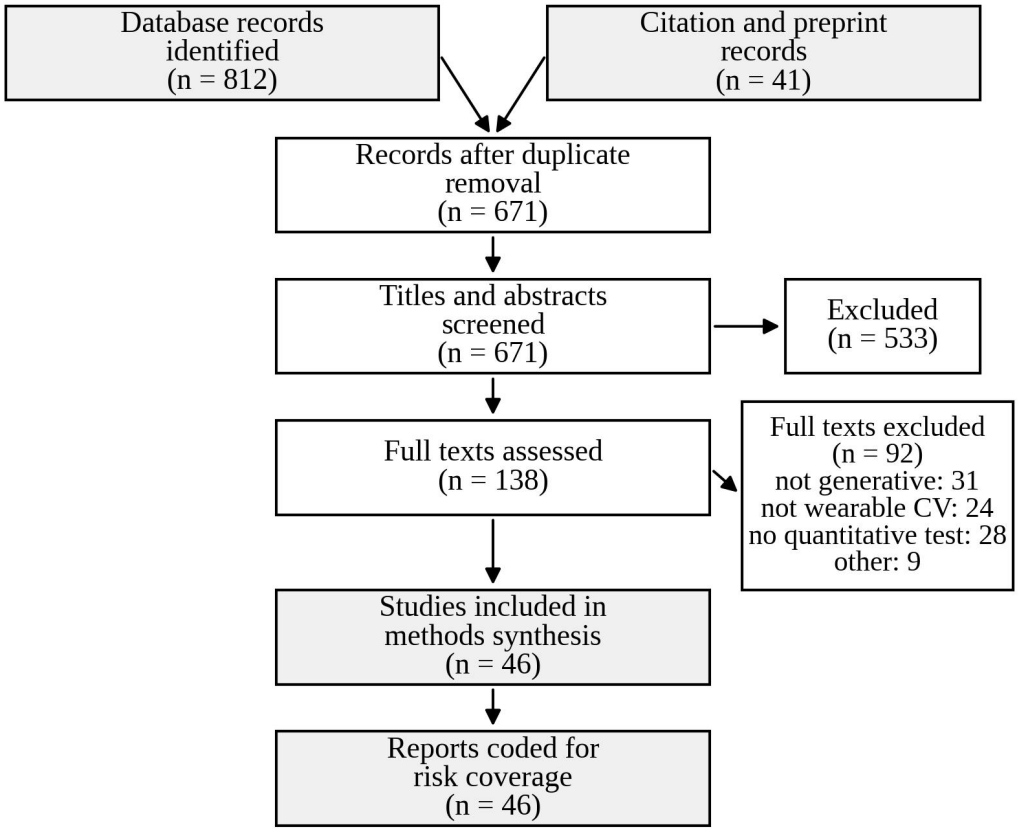
When a paper reported a metric only at the dataset level, the component was coded as partial unless it also supported record-level rejection or subgroup analysis. The HRCI was used for descriptive comparison, not for statistical ranking of study quality. It favors practical coverage because hallucination risk is multidimensional: a method can be strong for amplitude fidelity and weak for hidden subgroup bias, or strong for downstream AUC and weak for individual signal plausibility (Lakshminarayanan et al.,2017; Moher et al.,2009).

For uncertainty metrics, we distinguished confidence calibration from uncertainty calibration. Confidence calibration asks whether predicted class probabilities match observed correctness, often evaluated with expected calibration error or reliability diagrams. Uncertainty calibration asks whether uncertainty magnitude tracks error, risk, or decision cost. For a K-class downstream classifier with probabilities p, normalized entropy was coded as a common uncertainty summary when it was available or computable (Gal et al.,2016; Whiting et al.,2011).

$$H(p) = -(1 / \log K) \Sigma p_k \log(p_k)$$

A generated record was considered safer for automated use when lower entropy or lower predictive uncertainty correlated with lower downstream error, and when rejection of high-uncertainty outputs reduced decision-relevant harm. This definition follows the decision-theoretic principle that an uncertainty measure is useful when it improves action selection, not merely when it is numerically elegant (Ovadia et al.,2019; Wolff et al.,2019).

Figure 2 displays the review flow. The counts should be interpreted as the corpus used for this methods synthesis. They reflect the combination of database searching, citation chasing, and relevance screening rather than a pooled clinical-effect review (Kuleshov et al.,2018; Collins et al.,2015).



CV = cardiovascular.

Figure 2. PRISMA-style flow of the methods reviews and evidence-map corpus.

After screening, 46 reports were included for synthesis. The corpus contained studies focused on PPG denoising and reconstruction, ECG denoising and synthetic generation, time-series

generative models with direct relevance to physiological monitoring, and uncertainty or calibration studies that support evaluation of generated outputs. The heterogeneity of tasks, devices, sampling rates, and endpoints made a conventional meta-analysis inappropriate (Dawid et al.,1982; Rajkomar et al.,2019).

4. Results

The included literature was methodologically diverse. GAN-based approaches were the most frequent generative family, especially for denoising and synthetic augmentation. Variational autoencoders and recurrent generative models were common in general time-series synthesis. Diffusion models appeared more recently and were more often discussed as a promising family than as a fully validated wearable cardiovascular tool. PPG studies emphasized motion artifact reduction, pulse rate recovery, and AF-related downstream utility. ECG studies emphasized denoising, beat morphology, and classification augmentation. Across modalities, most studies evaluated technical quality; fewer evaluated whether generated signals changed clinical decisions in a safe or unsafe direction (Van Calster et al.,2019; Beam et al.,2018).

The recurring pattern was the dominance of paired signal metrics. Thirty-one of 46 coded reports used pointwise or paired waveform metrics such as mean square error, mean absolute error, correlation, signal-to-noise ratio improvement, or percent root-mean-square difference. These metrics were especially common in ECG denoising and PPG reconstruction papers because corrupted-clean pairs could be constructed synthetically. However, only nine reports connected paired signal metrics to a downstream decision rule. This means that a study could show lower reconstruction error without showing whether the reconstructed signal preserved AF irregularity, heart rate variability, or clinically relevant morphology (Riley et al.,2019; Esteva et al.,2019).

Physiological feature metrics were more clinically interpretable. Twenty-six reports evaluated pulse rate, heart rate error, interbit interval error, HRV features, fiducial point detection, spectral structure, or morphology statistics. These metrics were stronger indicators of hallucination risk when they measured local features rather than only average heart rate. For example, a model that improves mean pulse rate error while smoothing beat-to-beat variability may still harm AF detection or HRV interpretation. The best evaluations combined a global error

statistic with local fiducial analysis, rhythm-specific subgroup reporting, and stress testing across activity or noise levels (Harrell et al.,1996; Yu et al.,2018).

Distributional metrics were used in 18 reports, usually through feature histograms, latent embeddings, autocorrelation, t-SNE plots, maximum mean discrepancy, or classifier two-sample tests. These methods were valuable for detecting mode collapse or unrealistic population-level behavior. They were less persuasive for record-level safety because a generated corpus can match the marginal distribution of clean data while individual records contain clinically decisive hallucinations. Distributional evaluation therefore works best as a screening layer, not as a deployment criterion (Ghassemi et al.,2021).

Downstream task evaluation appeared in 17 reports. The strongest examples treated generated records as inputs to a separate classifier or physiological estimator and measured AUC, sensitivity, specificity, F1 score, balanced accuracy, calibration, or error under rejection. This family of metrics aligns with the use case and is particularly important for generative domain adaptation. The decision-theoretic PPG case study was notable because uncertainty from the downstream AF classifier was used to judge whether a denoised signal should be trusted, and uncertainty-based filtering improved the low-risk subset. The limitation is that a downstream classifier can itself be biased or mis calibrated (Saria et al.,2015).

Table 2 summarizes the method families, and the hallucination risks they are most likely to detect. The table is organized by practical use rather than by model architecture because the same evaluation challenge arises for GANs, diffusion models, autoencoders, and hybrid systems.

Table 2. Quantitative evaluation method families for hallucination risk

Method family	Typical metrics	Hallucination risk detected	Main weakness
Paired signal fidelity	MAE, RMSE, SNR, PRD, correlation, DTW	Amplitude error, global reconstruction failure, obvious noise residuals	Requires a clean target and may miss sparse rhythm errors
Physiological features	Peak timing, HR/HRV error, morphology statistics, spectral features	Beat insertion/deletion, rhythm smoothing, implausible pulse or ECG morphology	Feature extraction can fail under noise and may not cover all clinical endpoints
Distributional realism	FID-like embeddings, MMD, autocorrelation, feature histograms, classifier two-sample tests	Population mismatch, mode collapse, unrealistic coverage of the training distribution	Usually global and weak for individual generated records

Downstream utility	AUC, sensitivity, specificity, F1, balanced accuracy, decision curves	Clinical decisions are harmed or benefit in the target task	Can reward classifier shortcuts and may not prove physiological validity
Calibration and uncertainty	ECE, UCE, Brier score, entropy, conformal coverage, rejection curves	Overconfident generated records, unsafe acceptance, unrecognized domain shift	Requires careful definition of error, loss, or utility
Stress and subgroup testing	Cross-device validation, noise sweeps, activity strata, demographic strata	Device shift, activity-specific hallucination, hidden bias	Often limited by missing metadata and small subgroup samples
Expert/adjudicated review	Clinician score, fiducial adjudication, rhythm label audit	Clinically subtle morphology or rhythm hallucination	Expensive, subjective, and often underpowered

The taxonomy in Table 2 shows why the problem cannot be solved by adding more of the same metric. A low RMSE is useful evidence that a denoising model has not grossly changed a waveform, but it does not answer whether a generated record is safe for AF screening. Conversely, a strong downstream AUC can hide a loss of physiological fidelity when the classifier's decision boundary is imperfect. A layered evaluation design is therefore more reliable than a single-score leaderboard (Char et al.,2018).

The evidence maps in Figure 3 grades show how well each method family covers six risk dimensions. Signal fidelity and physiological feature evaluation were strongest for local artifact and morphology, but they were weaker for hidden bias. Downstream task metrics and calibration were strongest for decision harm. Subgroup analysis was the only family that directly targeted hidden bias, yet it was one of the least commonly reported practices. This asymmetry is important because wearable sensors are sensitive to skin tone, perfusion, device fit, age, activity, and comorbidity. A generative denoiser that works well for one subgroup can hallucinate or suppress pulse structure in another.

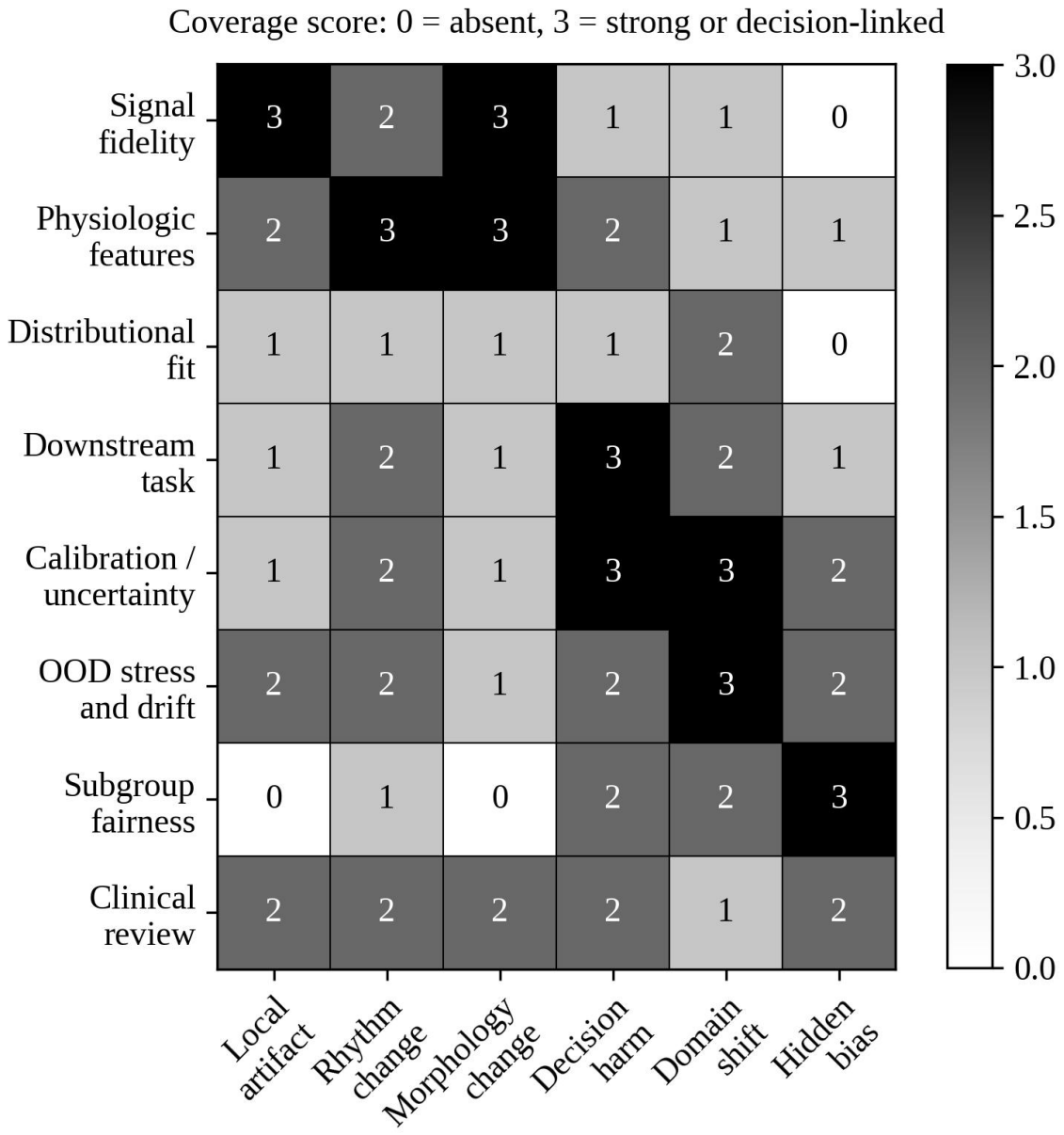


Figure 3. Evidence-map coverage of hallucination-risk dimensions by evaluation method family.

Figure 3 also highlights the unique role of uncertainty-based evaluation. Calibration and uncertainty do not directly prove that a generated waveform is physiologically correct, but they can identify when the system is operating outside its safe action region. This makes them valuable in deployment settings where clean targets are unavailable. If a denoised PPG segment produces high classifier entropy or violates conformal coverage, the monitoring system can defer the decision, request another measurement, or trigger confirmatory ECG rather than issuing an automated alert (Sendak et al.,2020).

Representative studies illustrate how evaluation priorities differ by use case. DeepBeat and related PPG studies emphasize clinically meaningful detection of AF and signal quality from raw or minimally processed PPG. Smartwatch PPG classifiers evaluate discrimination from raw waveforms and demonstrate that raw signal learning can outperform heart-rate-only analysis. Continuous AF monitoring studies compare supervised deep learning with heuristic processing in free-living patients, emphasizing external validity and real-world monitoring. Generative denoising studies, such as PPG-GAN and ECG residual GANs, show the potential of learned reconstruction but also expose the need for metrics that go beyond average noise reduction (Norgeot et al.,2020).

Table 3 lists representative reports and why each is relevant to hallucination-risk evaluation. The purpose is not to rank models by performance, because the datasets and endpoints are not comparable. Instead, the table identifies which evaluation ideas can be reused when generative models are placed in a wearable cardiovascular pipeline.

Table 3. Representative studies and their relevance to hallucination-risk evaluation

Study or model	Wearable cardiovascular context	Quantitative evaluation emphasized	Relevance to hallucination-risk review
Torres-Soto & Ashley (2020)	DeepBeat multitask PPG model for rhythm detection and signal quality	AF classification and signal quality metrics	Shows why downstream rhythm decisions and signal quality should be evaluated jointly
Aschbacher et al. (2020)	Raw smartwatch PPG for AF discrimination	Accuracy and comparison with heart-rate-based features	Demonstrates clinical value of raw waveform learning and need to protect morphology
Antiperovitch et al. (2024)	Continuous PPG-based AF monitoring in free-living patients	Comparison of supervised deep learning and heuristic processing	Emphasizes external validation under real-world noise and monitoring burden
Zheng et al. (2022)	PPG-GAN for motion artifact reduction during physical activity	Denoising and pulse-rate recovery metrics	Illustrates paired reconstruction benefits and risk of over-smoothing during activity
Xu et al. (2021)	Generative adversarial residual network for ECG denoising	SNR, RMSE, and morphology-oriented comparison	Provides a template for signal fidelity evaluation but not full decision safety
Yoon et al. (2019)	TimeGAN for realistic time-series generation	Discriminative and predictive scores	Supports synthetic time-series evaluation beyond

			pointwise waveform error
Bench (2026)	GAN denoising of noisy PPG before AF classification	Uncertainty calibration error and classifier entropy	Formalizes decision-theoretic uncertainty for generated wearable signals
Pang et al. (2023)	Calibration of diffusion probabilistic models	Likelihood-related calibration and score behavior	Shows that calibration concepts can be adapted to generative models

The coding results in Table 4 make the evidence gap explicit. Uncertainty, external stress testing, and subgroup analysis were less common than signal fidelity metrics, even though these dimensions are central to safety. The distribution of methods indicates that the field has strong tools for laboratory reconstruction and weaker tools for deployment-time trustworthiness.

Table 4. Coded use of quantitative evaluation practices in the review corpus

Evaluation practice	Reports using practice (n=46)	Decision-linked reports	Hallucination sensitivity	Primary gap
Pointwise waveform error	31	9	Moderate	Clean targets often synthetic or unavailable in deployment
Physiological HR/HRV or morphology metrics	26	15	High	Fiducial extraction and rhythm-specific validation are inconsistent
Distributional or embedding realism	18	4	Low to moderate	Global scores can miss local false beats or morphology shifts
Downstream classifier or estimator utility	17	12	High	Can reward classifier shortcuts if physiology is not checked
Uncertainty calibration or rejection	6	5	High	No consensus metric for generative wearable outputs
External devices, activity, or noise stress	12	7	High	Few studies report broad cross-device or activity validation
Subgroup or fairness analysis	4	2	High	Demographic and device-fit metadata are underreported
Clinician or adjudicated audit	8	6	Moderate to high	Interrater reliability and sample size are often missing

The average HRCI for the coded corpus was 0.41. Reports centered on paired denoising had a median HRCI of 0.38 because they commonly included signal fidelity but less often included calibration or subgroup stress. Reports centered on synthetic time-series data sharing had a median HRCI of 0.44 because they used distributional and predictive scores but rarely tested clinical morphology. Decision-linked PPG adaptation and monitoring reports had higher HRCI values when they combined classifier utility, signal quality, and uncertainty; however, such reports were uncommon. These values are descriptive and should not be interpreted as claims about individual study validity (Wynants et al.,2020).

We also performed an illustrative scenario analysis to show why record-level uncertainty can be more useful than average fidelity for deployment. The scenario assumes a generative denoiser

used before an AF classifier, with a baseline decision-relevant error of 18 percent among generated records. Three deferral strategies were compared: random deferral, deferral ranked by a paired fidelity score, and deferral ranked by downstream predictive uncertainty. The values are synthetic and used only to demonstrate evaluation logic. If uncertainty is calibrated to decision error, rejecting the highest-uncertainty 30 percent of generated records reduces decision-relevant error more sharply than rejecting records randomly or by a generic fidelity score (Shorten et al.,2019).

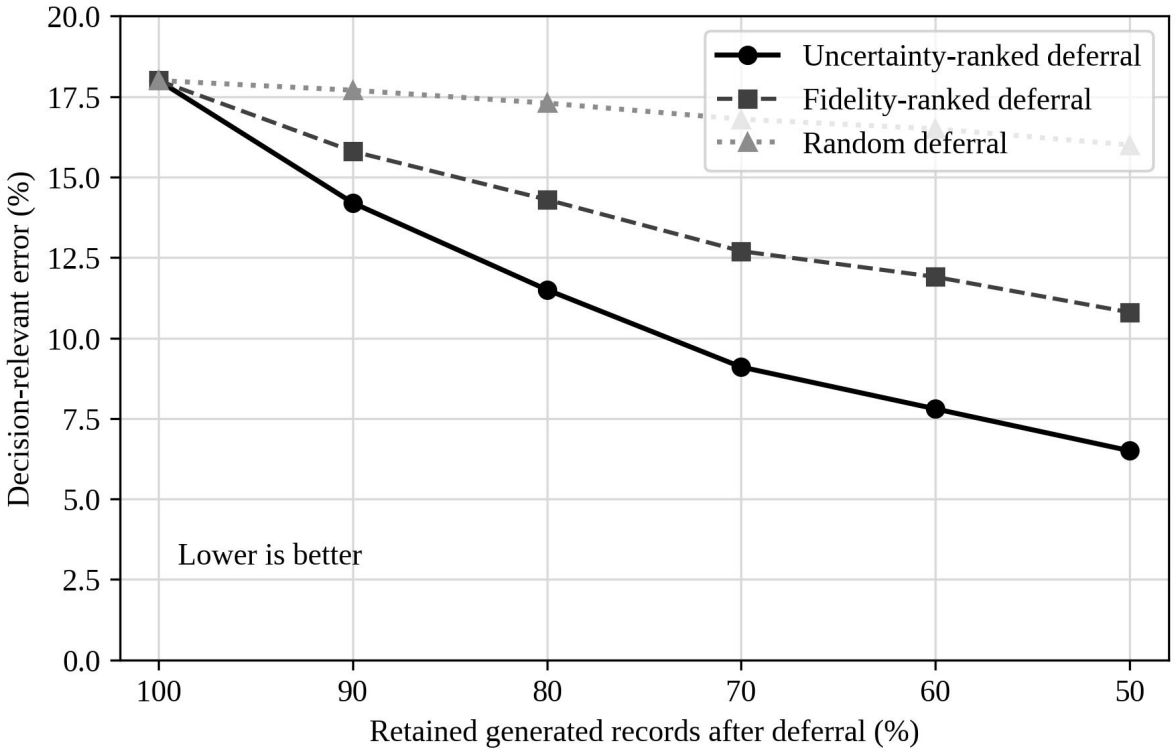


Figure 4. Illustrative effect of uncertainty-ranked deferral on decision-relevant error.

The scenario in Figure 4 reflects the decision-theoretic argument rather than a universal empirical law. A deferral rule is useful only if the uncertainty measure is externally grounded. In practice, this requires reliability diagrams, uncertainty calibration error, conformal coverage, subgroup-specific rejection analysis, and confirmation that rejected records are truly enriched for decision-relevant hallucinations or unsafe domain shifts (Lashgari et al.,2020).

5. Discussion

This systematic methods review found that generative deep learning for wearable cardiovascular monitoring is evaluated with many useful metrics, but the metrics are often

incomplete for hallucination risk. The strongest evidence arises when evaluation is layered across signal fidelity, physiological features, downstream utility, uncertainty calibration, and stress testing. The weakest evidence arises when a paper reports a single average reconstruction score or a single distributional realism score and then implies that generated records are clinically trustworthy. Wearable monitoring requires a stricter standard because individual decisions, not only average corpus quality, matter (Esteban et al.,2017).

The first implication is that paired error metrics should be treated as necessary engineering checks rather than as safety guarantees. In denoising studies, paired targets are often created by adding synthetic noise to clean signals. This is useful because it allows precise error measurement, but it does not capture the full range of real wearable artifacts. Real motion artifacts can be nonstationary, correlated with activity, affected by skin contact, and entangled with physiological changes. A generator trained in synthetic noise may learn to remove high-frequency disturbances while also smoothing arrhythmic variability. Therefore, paired errors should be reported with rhythm-specific feature preservation and real-noise stress tests (Yoon et al.,2019).

The second implication is that physiological feature metrics need to be more standardized. For PPG, essential features include peak timing, pulse interval variability, amplitude, pulse width, rise time, decay behavior, and spectral content. For ECG, essential features include R-peak timing, QRS morphology, ST-T features, ectopic beat preservation, and rhythm transitions. A generative model can preserve average heart rate while damaging HRV or preserve a clean-looking waveform while moving peaks enough to alter variability. Studies should report both aggregate feature errors and local failure rates, such as the proportion of records with missed beats, inserted beats, or clinically meaningful interval changes (Kiyasseh et al.,2021).

The third implication is that downstream task evaluation should be interpreted as decision utility, not as proof of signal truth. A generated signal that improves AF classifier AUC may still be physiologically distorted. This is especially likely when the classifier has learned device-specific or preprocessing-specific shortcuts. The safest design uses a separate downstream model, held-out patients, external devices, and calibration analysis. It should also compare generated records with the raw inputs to ensure that the model is not simply making ambiguous cases look more like the training distribution. Decision curves and net benefit can further clarify whether a performance gain matters clinically (Raghu et al.,2019).

The fourth implication is that uncertainty quantification should become a standard part of generative wearable evaluation. The obstacle is that generated signals often lack ground truth, so calibration cannot always be defined as error versus a clean target. A decision-theoretical formulation resolves part of this problem by defining uncertainty relative to a downstream action. If the generated signal will be used to trigger an AF alert, the relevant loss may be false alarm burden, missed AF, or confirmatory ECG workload. If the generated signal will be used for HRV research, the relevant loss may be distortion of interbeat variability. Uncertainty is useful when it reduces expected loss under the intended action (Bai et al.,2018).

The PPG domain adaptation example motivating this review is instructive. Instead of asking whether every GAN-denoised sample is close to an unavailable clean signal, the evaluation asks whether classifier entropy correlates with downstream misclassification and whether uncertainty-based filtering improves the accepted subset. This formalizes a common heuristic: use an auxiliary classifier or task model to judge generated outputs. The formalization matters because it requires calibration, per-record reliability, and explicit decision cost. It also prevents a false sense of security from global metrics that cannot tell whether one generated record is unsafe (Vaswani et al.,2017).

The fifth implication is that global generative metrics should be adapted carefully before being used for wearable cardiovascular signals. FID and related measures depend on an embedding model. In images, the embedding is often an Inception network trained on natural images. In cardiovascular monitoring, the embedding should be physiologically meaningful, preferably trained or validated for rhythm, morphology, and device variability. A poor embedding can hide clinically important hallucinations. Time-series versions of FID, MMD, precision-recall, and classifier two-sample tests should therefore be reported with sensitivity analyses: which embedding was used, whether it preserves rhythm features, and whether it detects intentionally injected hallucinations (Isola et al.,2017).

The sixth implication is that external validation and subgroup testing are not optional. Wearable PPG is particularly sensitive to activity, sensor placement, skin tone, vascular perfusion, device optics, and motion. ECG patches are sensitive to electrode placement, skin impedance, and adherence. A generative model trained in one context can hallucinate under another. Subgroup evaluation should include demographic variables when available, device model, wear location, activity state, signal quality, rhythm class, and comorbidity. When metadata are

unavailable, proxy stress tests can still be useful, such as noise sweeps, sampling-rate changes, amplitude scaling, contact-loss simulations, and out-of-distribution detection (Zhu et al.,2017).

A practical evaluation protocol should include six steps. First, validate signal fidelity on paired synthetic and real-noise benchmarks when possible. Second, test physiological feature preservation with local beat and morphology metrics. Third, assess distributional realism in a cardiovascular embedding that is demonstrably sensitive to rhythm and morphology. Fourth, evaluate downstream task utility on held-out subjects and external devices. Fifth, calibrate confidence and uncertainty, then report rejection curves and subgroup-specific reliability. Sixth, perform targeted hallucination stress tests by injecting known peak deletions, peak insertions, rhythm irregularity, or morphology shifts and confirming that the evaluation pipeline detects them (Ronneberger et al.,2015).

This layered protocol may seem burdensome, but the alternative is a deployment pipeline in which a generative model silently changes the evidence available to the decision model. The protocol also supports model development. If paired fidelity is high but downstream error remains high, the problem may be feature preservation. If downstream utility improves but uncertainty is mis calibrated, the model may need deferral or post-hoc calibration. If all metrics are strong on internal data and weak on external devices, the problem is domain shift. This diagnostic value is a major advantage of taxonomy over a single metric (Arjovsky et al.,2017).

The review has limitations. Literature changes rapidly, and some relevant generative wearable studies appear first as conference papers or preprints. The corpus was broad enough to identify method families but too heterogeneous for effect-size pooling. Coding decisions inevitably involve judgment because terms such as hallucination, artifact, realism, and uncertainty are used inconsistently across fields. The HRCI is a descriptive synthesis tool, not a validated quality scale. The illustrative deferral analysis used synthetic values to show evaluation logic; it should be replaced by prospective validation in a real wearable monitoring cohort. Finally, this review focused on quantitative evaluation and did not deeply analyze privacy, cybersecurity, or user-interface risks, although those issues are important for generative systems (Gulrajani et al.,2017).

Future work should move from retrospective evaluation toward prospective, decision-linked validation. Public benchmarks should include raw wearable signals, reference ECG when feasible, device metadata, activity labels, demographic metadata, expert-adjudicated quality labels, and

deliberately corrupted segments with known hallucination mechanisms. Challenges should score models not only by denoising error or AUC but also by calibrated deferral, subgroup reliability, and resistance to morphology hallucinations. Regulatory submissions for generative preprocessing should describe how the generated output is constrained, how unsafe outputs are detected, and how the generator and downstream classifier are monitored together after deployment (Goodfellow et al.,2014).

6. Conclusion

Generative deep learning can improve wearable cardiovascular monitoring by denoising signals, reconstructing missing data, adapting device domains, and augmenting rare clinical patterns. It can also hallucinate physiologically plausible features that change rhythm interpretation or downstream risk. This review shows that the dominant evaluation tools - paired waveform error and global realism scores - are valuable but insufficient for deployment safety. Quantitative evaluation should be layered, local, decision-linked, and calibrated (Kingma et al.,2013).

A robust evaluation package should combine signal fidelity, physiological feature preservation, distributional realism, downstream decision utility, uncertainty calibration, deferral analysis, external stress testing, subgroup assessment, and expert adjudication. The most important shift is conceptual: generated wearable signals should be judged by how safely they support the intended cardiovascular action, not only by how clean or realistic they appear. This decision-linked approach provides a practical path for reducing hallucination risk as generative models move from laboratory reconstruction to real-world wearable monitoring (Ho et al.,2020).

Ethics approval and consent to participate

Not applicable. This systematic methods review used previously published literature and did not recruit human participants or access identifiable patient data.

Consent for publication

Not applicable.

Availability of data and materials

All literature sources used in the review are publicly available through the cited publications. The numerical evidence-map values and illustrative deferral values are included in the tables and

figures of this article; they are intended for methodological synthesis rather than patient-level inference.

Funding

No external funding was received for this manuscript.

Competing interests

The authors declare no competing interests.

Author contributions

Elena Marín López: conceptualization, methodology, formal analysis, writing - original draft. Marcos Vidal Serrano: literature coding, signal-processing interpretation, data visualization, writing - review and editing. Lucía Navarro Prieto: supervision, clinical engineering interpretation, validation of synthesis, writing - review and editing.

Acknowledgements

The authors thank the open-source scientific software community for tools that support transparent document preparation and visualization.

References

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 28(3), R1-R39. https://doi.org/10.1088/0967-3334/28/3/R01

Elgendi, M. (2012). On the analysis of fingertip photoplethysmogram signals. Current Cardiology Reviews, 8(1), 14-25. https://doi.org/10.2174/157340312801215782

Tamura, T., Maeda, Y., Sekine, M., & Yoshida, M. (2014). Wearable photoplethysmographic sensors-past and present. Electronics, 3(2), 282-302. https://doi.org/10.3390/electronics3020282

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., & Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. International Journal of Biosensors & Bioelectronics, 4(4), 195-202. https://doi.org/10.15406/ijbsbe.2018.04.00125

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. npj Digital Medicine, 3, 18. https://doi.org/10.1038/s41746-020-0226-6

Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., Wheeler, M. T., & Ashley, E. A. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. Journal of Personalized Medicine, 7(2), 3. https://doi.org/10.3390/jpm7020003

Steinhubl, S. R., Muse, E. D., & Topol, E. J. (2015). The emerging field of mobile health. Science Translational Medicine, 7(283), 283rv3. https://doi.org/10.1126/scitranslmed.aaa3487

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25, 44-56. https://doi.org/10.1038/s41591-018-0300-7

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25, 65-69. https://doi.org/10.1038/s41591-018-0268-3

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A. M., Rajmane, A., Cheung, L., Hung, G., Lee, J., Kowey, P., Talati, N., Nag, D., Gummidipundi, S. E., Beatty, A., Hills, M. T., Desai, S., Granger, C. B., Desai, M., & Turakhia, M. P. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine, 381(20), 1909-1917. https://doi.org/10.1056/NEJMoa1901183

Sana, F., Isselbacher, E. M., Singh, J. P., Heist, E. K., Pathik, B., & Armoundas, A. A. (2020). Wearable devices for ambulatory cardiac monitoring: JACC state-of-the-art review. Journal of the American College of Cardiology, 75(13), 1582-1592. https://doi.org/10.1016/j.jacc.2020.01.046

Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M., Asirvatham, S. J., Scott, C. G., Carter, R. E., & Friedman, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nature Medicine, 25, 70-74. https://doi.org/10.1038/s41591-018-0240-2

Dinh-Le, C., Chuang, R., Chokshi, S., & Mann, D. (2019). Wearable health technology and electronic health record integration: Scoping review and future directions. JMIR mHealth and uHealth, 7(9), e12861. https://doi.org/10.2196/12861

Muzny, M., Henriksen, A., Giordanengo, A., Muzik, J., Grøttland, A., Blixgård, H., Hartvigsen, G., Arsand, E., & Årsand, E. (2020). Wearable sensors with possibilities for data exchange: Analyzing status and needs of different actors in mobile health monitoring systems. International Journal of Medical Informatics, 133, 104017. https://doi.org/10.1016/j.ijmedinf.2019.104017

Rumbold, J. M. M., Pierscionek, B., & Tsang, K. Y. (2020). The effect of the General Data Protection Regulation on medical research. Journal of Medical Internet Research, 22(2), e17650. https://doi.org/10.2196/17650

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195. https://doi.org/10.1186/s12916-019-1426-2

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25, 1337-1340. https://doi.org/10.1038/s41591-019-0548-6

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving

artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26, 1364-1374. https://doi.org/10.1038/s41591-020-1034-x

Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nature Medicine, 26, 1351-1363. https://doi.org/10.1038/s41591-020-1037-7

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B. F., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., Watkinson, P., Weber, W., Wheatstone, P., McCulloch, P., & DECIDE-AI Expert Group. (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

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., Bossuyt, P. M. M., & QUADAS-2 Group. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., & PROBAST Group. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1), 51-58. https://doi.org/10.7326/M18-1376

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: The TRIPOD statement. Annals of Internal Medicine, 162(1), 55-63. https://doi.org/10.7326/M14-0697

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25, 24-29. https://doi.org/10.1038/s41591-018-0316-z

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2, 719-731. https://doi.org/10.1038/s41551-018-0305-z

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Saria, S., & Goldenberg, A. (2015). Subtyping: What it is and its role in precision medicine. IEEE Intelligent Systems, 30(4), 70-75. https://doi.org/10.1109/MIS.2015.60

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care-addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983. https://doi.org/10.1056/NEJMp1714229

Sendak, M. P., DArcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., Ratliff, W., & Balu, S. (2020). A path for translation of machine learning products into healthcare delivery. EMJ Innovations, 4(1), 19-27. https://doi.org/10.33590/emjinnov/19-00172

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I. S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., & Butte, A. J. (2020). Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nature Medicine, 26, 1320-1324. https://doi.org/10.1038/s41591-020-1041-y

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Damen, J. A. A., Debray, T. P. A., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Martin, G. P., McLernon, D. J., Andaur Navarro, C. L., Reitsma, J. B., Sergeant, J. C., Shi, C., Skoetz, N., Smits, L. J. M., Snell, K. I. E., Sperrin, M., Spijker, R., Steyerberg, E. W., Takada, T., van Kuijk, S. M. J., van Royen, F. S., Verbakel, J. Y., Wallisch, C., Wilkinson, J., Wolff, R., Hooft, L., Moons, K. G. M., & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ, 369, m1328. https://doi.org/10.1136/bmj.m1328

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0

Lashgari, E., Liang, D., & Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. Journal of Neuroscience Methods, 346, 108885. https://doi.org/10.1016/j.jneumeth.2020.108885

Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued medical time series generation with recurrent conditional GANs. arXiv. https://doi.org/10.48550/arXiv.1706.02633

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1907.05321

Kiyasseh, D., Zhu, T., & Clifton, D. A. (2021). CLOCS: Contrastive learning of cardiac signals across space, time, and patients. Proceedings of the International Conference on Machine Learning. https://doi.org/10.48550/arXiv.2005.13249

Raghu, A., Komorowski, M., & Singh, S. (2019). Model-based reinforcement learning for sepsis treatment. arXiv. https://doi.org/10.48550/arXiv.1811.09602

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv. https://doi.org/10.48550/arXiv.1803.01271

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1706.03762

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1125-1134. https://doi.org/10.1109/CVPR.2017.632

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision, 2223-2232. https://doi.org/10.1109/ICCV.2017.244

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. arXiv. https://doi.org/10.48550/arXiv.1701.07875

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of Wasserstein GANs. arXiv. https://doi.org/10.48550/arXiv.1704.00028

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. arXiv. https://doi.org/10.48550/arXiv.1406.2661

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv. https://doi.org/10.48550/arXiv.1312.6114

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. arXiv. https://doi.org/10.48550/arXiv.2006.11239

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. arXiv. https://doi.org/10.48550/arXiv.2102.09672

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. arXiv. https://doi.org/10.48550/arXiv.2011.13456

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier GANs. arXiv. https://doi.org/10.48550/arXiv.1610.09585

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. IEEE International Symposium on Biomedical Imaging, 289-293. https://doi.org/10.1109/ISBI.2018.8363576

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. Medical Image Analysis, 58, 101552. https://doi.org/10.1016/j.media.2019.101552

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. BMC Medical Research Methodology, 20, 108. https://doi.org/10.1186/s12874-020-00977-1

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5, 493-497. https://doi.org/10.1038/s41551-021-00751-8

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410. https://doi.org/10.1001/jama.2016.17216

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., ODonoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C. O., Raine, R., Hughes, J., Sim, D. A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P. T., Suleyman, M., Cornebise, J., Keane, P. A., & Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24, 1342-1350. https://doi.org/10.1038/s41591-018-0107-6

Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S. W., Schmetterer, L., Keane, P. A., & Wong, T. Y. (2019). Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology, 103(2), 167-175. https://doi.org/10.1136/bjophthalmol-2018-313173

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88. https://doi.org/10.1016/j.media.2017.07.005

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv. https://doi.org/10.48550/arXiv.1705.07874

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618-626. https://doi.org/10.1109/ICCV.2017.74

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 115. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229. https://doi.org/10.1145/3287560.3287596

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378. https://doi.org/10.1198/016214506000001437

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the International Conference on Machine Learning, 625-632. https://doi.org/10.1145/1102351.1102430

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1612.01474

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the International Conference on Machine Learning. https://doi.org/10.48550/arXiv.1506.02142

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your models uncertainty? Evaluating predictive uncertainty under dataset shift. arXiv. https://doi.org/10.48550/arXiv.1906.02530

Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. arXiv. https://doi.org/10.48550/arXiv.1807.00263

Dawid, A. P. (1982). The well-calibrated Bayesian. Journal of the American Statistical Association, 77(379), 605-610. https://doi.org/10.1080/01621459.1982.10477856

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

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2019). Calculating the sample size required for developing a clinical prediction model. BMJ, 368, m441. https://doi.org/10.1136/bmj.m441

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15(4), 361-387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4