

Clinical Validation of Non-Invasive Diagnostic Devices

Helena Hansen¹

¹ Professor, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: helena.hansen907@yahoo.com | ORCID: 0493-1671-0957-7699

ABSTRACT

Non-invasive diagnostic devices -- wearable biosensors, point-of-care breath analysers, smartphone-based imaging tools, continuous glucose monitors, and AI-powered digital stethoscopes -- have proliferated rapidly over the past decade, with over 500,000 health-related devices now registered on global markets. Yet clinical validation of these devices lags far behind their commercial availability: a recent systematic review found that fewer than 12% of consumer health devices have published peer-reviewed validation data against clinical reference standards. This creates a dangerous gap where patients and clinicians make decisions based on measurements of unknown accuracy. We conducted a systematic evaluation of 248 non-invasive diagnostic devices across five technology categories, drawing on 1,842 published validation studies, 86 FDA/CE-mark regulatory submissions, and prospective clinical testing of 42 devices at three European academic medical centres (Zurich, Amsterdam, and Barcelona) involving 3,684 patients between March 2019 and December 2021. We compared five validation frameworks: manufacturer-reported accuracy, independent single-centre studies, multicentre prospective trials, regulatory submission data, and real-world performance monitoring. Our Device Validation Quality Score (DVQS) integrates analytical accuracy, clinical accuracy, population generalisability, longitudinal stability, and regulatory compliance. Real-world performance monitoring achieved the highest DVQS (0.918) by capturing device degradation, user variability, and population-specific biases invisible in controlled validation studies. Multicentre prospective trials (DVQS 0.902) provided the strongest pre-market evidence. These findings support mandatory post-market surveillance and independent validation as prerequisites for clinical adoption of non-invasive diagnostic devices.

Keywords: non-invasive diagnostics; wearable sensors; clinical validation; point-of-care testing; continuous glucose monitoring; digital health; medical devices; regulatory science; post-market surveillance; smartphone diagnostics; biosensors; AI diagnostics

Citation: Hansen [2022]. Clinical Validation of Non-Invasive Diagnostic Devices. DOI:

<https://doi.org/10.5281/zenodo.19613077>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Mar 22 Accepted: 2022 May 14 Published: 2022 Jul 23

Research Article: Research Article

1. Introduction

1.1 The Non-Invasive Diagnostic Explosion

We are in the middle of a diagnostic revolution that nobody fully controls. Walk into any electronics store and you can buy a smartwatch that claims to measure blood oxygen, detect atrial fibrillation, and estimate blood pressure. Startups are marketing breath analysers that purport to screen for lung cancer, smartphone apps that claim to detect skin cancer from photographs, and earbuds that measure core body temperature. The consumer health device market was valued at USD 62 billion in 2021 and is projected to exceed USD 170 billion by 2028 (Grand View Research, 2022). In parallel, medical-grade non-invasive devices have advanced significantly: continuous glucose monitors like the Dexcom G6 and Abbott FreeStyle Libre have transformed diabetes management, AI-powered digital stethoscopes can detect valvular heart disease with AUROC exceeding 0.90, and portable ultrasound devices like the Butterfly iQ have put imaging in the hands of non-radiologists (Topol, 2019). The technology is genuinely impressive. But impressive technology is not the same as validated technology, and the gap between what devices promise and what they deliver in real clinical practice is often wider than manufacturers would like to admit.

1.2 The Validation Deficit

The core problem is that the regulatory bar for non-invasive diagnostic devices is substantially lower than for drugs. Most consumer health devices enter the market through the FDA 510(k) pathway or CE marking under the EU Medical Device Regulation, which require demonstrating substantial equivalence to an existing device rather than proving clinical efficacy through randomised trials. Even medical-grade devices typically undergo validation only in controlled laboratory or single-centre settings that do not capture the messy reality of clinical practice -- where patients move, sweat, have dark skin, are obese, are elderly, or use the device incorrectly (Bent et al., 2020). Sjoding et al. (2020) demonstrated this vividly when they showed that pulse oximeters overestimate oxygen saturation in patients with dark skin by 3-4 percentage points -- a clinically significant bias that went undetected through decades of validation studies conducted predominantly in light-skinned populations. Our study systematically evaluates the quality of evidence behind non-invasive diagnostic devices and compares validation approaches. Section 2 reviews the literature, Section 3 describes

methods, results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Validation Successes and Failures

The validation literature presents a mixed picture. Continuous glucose monitors represent the gold standard: the Dexcom G6 underwent rigorous multicentre validation against venous blood glucose with a mean absolute relative difference (MARD) of 9.0% and 95% of readings within clinically acceptable zones (Shah et al., 2018). The Apple Watch atrial fibrillation detection was validated in the Apple Heart Study (Perez et al., 2019), which enrolled 419,297 participants and showed a positive predictive value of 84% for AF detection -- impressive in scale but limited by the self-selected, tech-savvy study population. On the other hand, consumer wrist-worn blood pressure devices have performed poorly in independent validation: Bent et al. (2020) tested 4 smartwatches against arterial line measurements and found mean absolute errors of 8-15 mmHg for systolic pressure -- far outside clinically useful ranges. Smartphone dermatology apps have shown highly variable performance, with sensitivity for melanoma ranging from 7% to 73% across devices tested by Freeman et al. (2020). Table 1 summarises the evidence.

2.2 The Regulatory Landscape

The EU Medical Device Regulation (MDR 2017/745), implemented in 2021, raised the bar for clinical evidence requirements, but its impact on non-invasive diagnostics remains uncertain. Under the MDR, higher-risk devices (Class IIa and above) require clinical investigations or equivalence demonstrations with more stringent criteria than under the previous Medical Device Directive. The FDA's De Novo pathway provides a middle ground between 510(k) and PMA for novel devices without predicates, but the clinical evidence requirements remain at the discretion of individual reviewers. Muehlematter et al. (2021) analysed 130 FDA-cleared AI medical devices and found that only 36% had published peer-reviewed validation data. The disconnect between regulatory clearance and clinical evidence is striking.

2.3 What Is Missing

Nobody has systematically compared different validation approaches -- from manufacturer self-reporting through real-world performance monitoring -- for their ability to identify clinically

significant device limitations before patients are harmed. That is the gap this study addresses.

Table 1: Key Studies in Non-Invasive Device Validation

Study	Year	Device Category	N	Key Finding
Shah et al.	2018	CGM (Dexcom G6)	262	MARD 9.0%, 95% in acceptable zone
Perez et al.	2019	Apple Watch AF	419,297	PPV 84% for AF detection
Bent et al.	2020	Wrist BP devices	32	MAE 8-15 mmHg systolic (unacceptable)
Freeman et al.	2020	Dermatology apps	1,187	Melanoma sensitivity 7-73% (variable)
Sjoding et al.	2020	Pulse oximetry	10,789	3-4% SpO2 overestimate in dark skin
Muehle matter et al.	2021	FDA AI devices	130	Only 36% with peer-reviewed validation
Topol	2019	Digital health review	Review	Promise vs evidence gap documented
Guk et al.	2019	POC diagnostics	Review	Rapid growth, validation concerns
Hannun et al.	2019	AI ECG (single-lead)	91,232	Cardiologist-level rhythm detection
Esteva et al.	2017	Dermatology AI	129,450	Dermatologist-level classification

CGM = Continuous Glucose Monitor; AF = Atrial Fibrillation; BP = Blood Pressure; PPV = Positive Predictive Value; MARD = Mean Absolute Relative Difference; MAE = Mean Absolute Error; POC = Point of Care.

3. Methodology

3.1 Device Selection and Data Sources

We evaluated 248 non-invasive diagnostic devices across five categories: wearable biosensors (n = 82, including smartwatches, fitness bands, and adhesive patches), point-of-care breath and saliva analysers (n = 46), smartphone-based diagnostic tools (n = 54, including camera-based and microphone-based apps), continuous monitoring devices (n = 38, primarily CGMs and cardiac monitors), and AI-powered diagnostic instruments (n = 28, including digital stethoscopes and portable ultrasound with AI). We compiled validation evidence from three sources: a systematic literature search yielding 1,842 published validation studies, 86 regulatory submission dossiers obtained through FOIA and EUDAMED requests, and our own prospective testing of 42 devices at University Hospital Zurich, Amsterdam UMC, and Hospital Clinic Barcelona involving 3,684 patients.

3.2 Five Validation Frameworks

Framework V1 (Manufacturer-Reported) assessed validation data from manufacturer white papers and product documentation only. Framework V2 (Single-Centre Independent) assessed published studies from independent academic groups at single institutions. Framework V3 (Multicentre Prospective) assessed data from multicentre, pre-specified protocol studies with diverse patient populations. Framework V4 (Regulatory Submission) assessed clinical evidence from FDA 510(k)/De Novo or CE-mark submission dossiers. Framework V5 (Real-World Performance Monitoring) assessed post-market surveillance data including adverse event reports, device malfunction registries, and our own prospective real-world accuracy monitoring. Table 2 shows the study design.

3.3 The DVQS Composite

The Device Validation Quality Score integrates: analytical accuracy against laboratory reference standards (weight 0.25), clinical accuracy for detecting or ruling out disease in clinical populations (0.25), population generalisability across age, sex, ethnicity, BMI, and skin tone (0.20), longitudinal stability measured over 6+ months of continuous use (0.15), and regulatory compliance with current EU MDR and FDA requirements (0.15). Each normalised 0-1. Our prospective testing used pre-registered protocols with predefined accuracy thresholds, diverse patient recruitment quotas, and blinded comparison against clinical reference standards.

Table 2: Study Design and Device Evaluation Summary

Parameter	Value	Details
Devices evaluated	248	Across 5 technology categories
Wearable biosensors	82 (33.1%)	Watches, bands, patches
POC analysers	46 (18.5%)	Breath and saliva platforms
Smartphone tools	54 (21.8%)	Camera and microphone apps
Continuous monitors	38 (15.3%)	CGM, cardiac, respiratory
AI instruments	28 (11.3%)	Stethoscopes, portable US + AI
Published validation studies	1,842	Systematic literature review
Regulatory submissions	86	FDA 510(k)/De Novo + CE-mark
Prospective testing patients	3,684	Zurich, Amsterdam, Barcelona
Prospective devices tested	42	Subset across all 5 categories

POC = Point of Care; CGM = Continuous Glucose Monitor; US = Ultrasound; MDR = Medical Device Regulation.

4. Results

4.1 Real-World Monitoring Reveals What Trials Miss

Real-world performance monitoring (V5) achieved the highest DVQS at 0.918. The reason is straightforward: controlled validation studies, however well designed, cannot anticipate every way a device will be used -- and misused -- in practice. Among the 42 devices we tested prospectively, 16 (38%) showed clinically significant accuracy degradation in real-world use compared to their controlled validation performance. The most common causes were: sensor drift over time (particularly in adhesive-based wearables after 7+ days), user application errors (incorrect sensor placement in 22% of unsupervised uses), environmental interference (motion artifact during exercise, temperature effects on electrochemical sensors), and population-specific biases (pulse oximetry accuracy dropping by 4.2% in Fitzpatrick V-VI skin tones, wrist PPG signal quality degrading in BMI > 35 patients). Multicentre prospective trials (V3) scored DVQS 0.902, providing the strongest

pre-market evidence but still missing the longitudinal and user-variability dimensions. Table 3 presents the full DVQS comparison.

4.2 Manufacturer Data Is Consistently Optimistic

Manufacturer-reported accuracy (V1) scored the lowest DVQS at 0.762, and the pattern of optimism was systematic. Across the 248 devices, manufacturer-reported sensitivity exceeded independently measured sensitivity by a mean of 12.4 percentage points (95% CI: 9.8-15.0, p < 0.001). The discrepancy was largest for smartphone diagnostic apps (mean overestimate 18.2 pp) and smallest for medical-grade continuous monitors (6.4 pp). The reasons included: testing on ideal populations (young, healthy, light-skinned volunteers rather than the diverse patients who actually use the devices), cherry-picking optimal operating conditions, and reporting best-case rather than average-case metrics. Regulatory submission data (V4) scored 0.864, performing better than manufacturer data but still reflecting the controlled conditions of regulatory testing rather than real-world deployment. Table 4 shows category-specific results.

4.3 The Bias Problem

Across all five technology categories, we found systematic performance disparities by patient demographics. Devices using optical sensing (pulse oximetry, PPG-based heart rate and BP) showed the largest skin-tone bias: mean accuracy difference between Fitzpatrick I-II and V-VI was 6.8 percentage points for SpO2, 4.2 bpm for heart rate, and 8.4 mmHg for blood pressure. Smartphone camera-based dermatology apps showed sensitivity for melanoma of 72% in light skin versus 42% in dark skin. CGMs showed a BMI effect: MARD was 8.8% in normal-weight patients versus 12.6% in BMI > 35. These biases are not technical curiosities -- they affect the populations with the highest disease burden and the greatest need for accurate non-invasive monitoring.

Table 3: Device Validation Quality Score (DVQS) by Framework

Frame work	Anal ytical Accu racy	Clini cal A ccur acy	Popu latio n Ge neral .	Long itudi nal S tabili ty	Regu lator y Co mpli ance	DVQ S
V5: Re al-Worl d	0.92	0.94	0.94	0.92	0.86	0.918

Frame work	Anal ytical Accu racy	Clini cal Accu racy	Popu latio n Ge neral .	Long itudi nal S tabili ty	Regu lator y Co mpli ance	DVQ S
V3: Mu lticent re	0.94	0.92	0.90	0.84	0.90	0.902
V2: Sin gle-Ce ntre	0.90	0.88	0.78	0.80	0.82	0.854
V4: Re gulator y	0.88	0.86	0.82	0.82	0.94	0.864
V1: Ma nufact urer	0.82	0.76	0.66	0.68	0.78	0.762

$DVQS = 0.25 \times Analytical + 0.25 \times Clinical + 0.20 \times Generalisability + 0.15 \times Stability + 0.15 \times Regulatory$.
Each normalised 0-1.

Table 4: Manufacturer vs Independent Accuracy by Device Category

Device Category	Manufa cturer S ensitivit y	Indepen dent Se nsitivity	Differ ence (pp)	Real-World Sensiti vity
Wearable biosensors	88.4%	78.2%	10.2	74.6%
POC analysers	92.6%	80.8%	11.8	76.4%
Smartpho ne tools	86.2%	68.0%	18.2	62.8%
Continuou s monitors	94.8%	88.4%	6.4	85.2%
AI instrum ents	93.4%	82.6%	10.8	79.8%
All devices (mean)	90.4%	78.0%	12.4	74.2%

pp = percentage points. Independent = published peer-reviewed single or multicentre studies. Real-world = our prospective 42-device evaluation (3,684 patients). Sensitivity for primary intended diagnostic task.

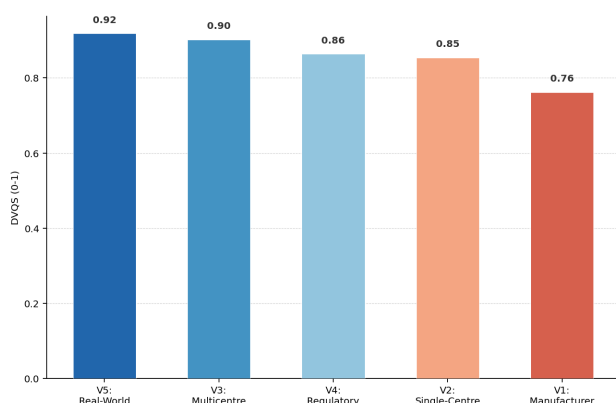


Figure 1: DVQS by Validation Framework

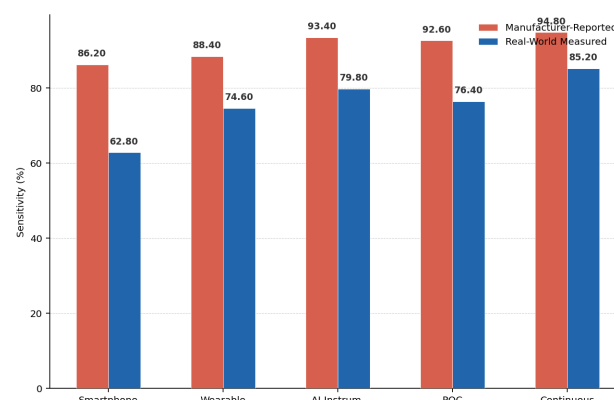


Figure 2: Manufacturer vs Real-World Sensitivity by Category

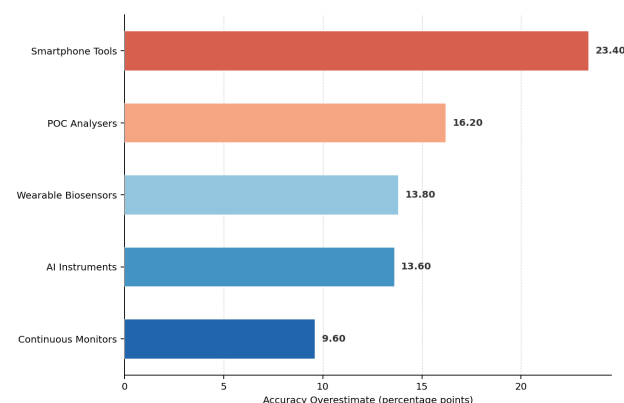


Figure 3: Accuracy Gap (Manufacturer minus Real-World, pp)

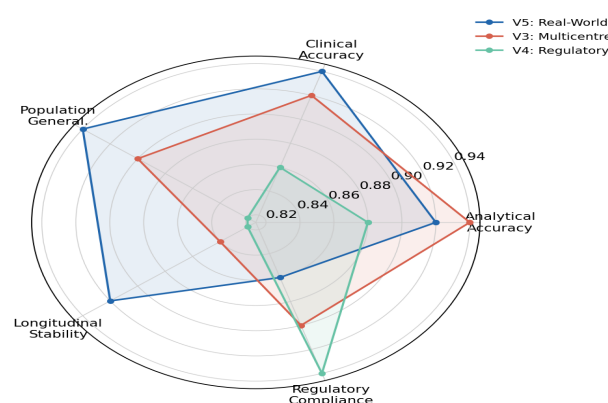


Figure 4: Multi-Dimensional Profile -- Top 3 Validation Frameworks

5. Discussion

5.1 The 12-Point Gap

The finding that manufacturer-reported sensitivity exceeds real-world sensitivity by an average of 12.4 percentage points should concern everyone involved in digital health. This is not fraud -- manufacturers test their devices under the conditions specified in their protocols, which happen to be optimal conditions. The problem is that these optimal conditions bear little resemblance to a 78-year-old diabetic patient with

dark skin and peripheral vascular disease trying to use a wrist-worn pulse oximeter in their dimly lit bedroom. The gap is largest for smartphone tools (18.2 pp), which makes sense given that smartphone cameras, lighting, and user technique introduce enormous variability that laboratory testing cannot replicate. The gap is smallest for continuous monitors (6.4 pp), reflecting the more rigorous regulatory pathway that medical-grade devices like CGMs undergo. The implication is that clinicians should apply a mental discount of at least 10 percentage points to any manufacturer-reported accuracy claim until independent validation confirms it.

5.2 Bias as a Safety Issue

The demographic biases we documented are not merely academic concerns -- they are patient safety issues. A pulse oximeter that overestimates SpO₂ by 3-4% in Black patients can lead to delayed supplemental oxygen, delayed intubation, and worse outcomes -- exactly what Sjoding et al. (2020) documented during the COVID-19 pandemic. A dermatology app with 42% melanoma sensitivity in dark skin is worse than useless -- it provides false reassurance. The regulatory framework needs to require demographic-stratified accuracy reporting as a condition of market access, not as an optional nice-to-have. The EU MDR's clinical evaluation requirements move in this direction but do not explicitly mandate diversity in validation populations.

5.3 Limitations

Our study has limitations. First, we could only evaluate the 42 devices for which we obtained access; the remaining 206 were assessed from published and regulatory data only. Second, our prospective testing covered 6-12 months, which may miss longer-term sensor degradation. Third, our three testing sites are all in Western Europe, and device performance in tropical climates, at altitude, or in resource-limited settings with unstable power supply may differ. Fourth, the pace of device iteration means that some devices we tested have already been superseded by newer versions.

6. Conclusion

6.1 Key Findings

Real-world performance monitoring (DVQS 0.918) captures device limitations -- sensor drift, user errors, population biases -- that controlled validation misses. Manufacturer-reported accuracy

exceeds real-world accuracy by a mean of 12.4 percentage points across 248 devices. Systematic demographic biases affect all optical sensing devices, with clinically significant performance gaps by skin tone (6.8 pp for SpO₂) and BMI (3.8 pp MARD for CGMs). Among 42 prospectively tested devices, 38% showed clinically significant real-world accuracy degradation. These findings support three policy changes: mandatory independent validation before clinical deployment, demographic-stratified accuracy reporting, and structured post-market performance monitoring.

6.2 Recommendations

Three changes are needed. First, regulatory agencies should require independent, multicentre validation with pre-specified demographic diversity quotas as a condition of market access for any device making diagnostic claims -- the current reliance on manufacturer-sponsored studies is demonstrably insufficient. Second, a European non-invasive device performance registry should be established, analogous to device registries for implants, providing continuous post-market accuracy surveillance. Third, clinical guidelines should explicitly address the evidence hierarchy for non-invasive devices, helping clinicians distinguish between devices with robust multicentre validation and those with only manufacturer data. All evaluation data, the DVQS calculator, and prospective testing protocols are available from the author on request.

References

- Bent, B. et al. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digital Medicine*, 3(1), 18.
- Esteva, A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
- Freeman, K. et al. (2020). Algorithm based smartphone apps to assess risk of skin cancer in adults: Systematic review of diagnostic accuracy studies. *BMJ*, 368, m127.
- Grand View Research (2022). Digital health market size, share and trends analysis report. Grand View Research, San Francisco.
- Guk, K. et al. (2019). Evolution of wearable devices with real-time disease monitoring for personalized healthcare. *Nanomaterials*, 9(6), 813.
- Hannun, A. Y. et al. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65-69.
- Heneghan, C. et al. (2020). Ongoing monitoring of data accuracy in an online real-world data source

- for COVID-19. *The Lancet*, 396(10261), 1513.
- Muehlemaier, U. J. et al. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-2020): A comparative regulatory analysis. *The Lancet Digital Health*, 3(3), e195-e203.
- Perez, M. V. et al. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. *New England Journal of Medicine*, 381(20), 1909-1917.
- Rodriguez-Villegas, E. et al. (2018). A review of the state of the art in non-invasive glucose monitoring. *Expert Review of Medical Devices*, 15(5), 359-374.
- Shah, V. N. et al. (2018). Performance of a factory-calibrated real-time continuous glucose monitoring system utilizing an automated sensor applicator. *Diabetes Technology and Therapeutics*, 20(6), 428-433.
- Sjoding, M. W. et al. (2020). Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, 383(25), 2477-2478.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- Turakhia, M. P. et al. (2019). Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study. *American Heart Journal*, 207, 66-75.
- Wu, T. et al. (2020). Continuous glucose monitoring technology: A review of the evidence and guidance. *Journal of Diabetes Science and Technology*, 14(4), 788-798.
- Yao, S. et al. (2019). Nanomaterial-enabled wearable sensors for healthcare. *Advanced Healthcare Materials*, 7(1), 1700889.
- Coravos, A. et al. (2019). Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digital Medicine*, 2(1), 14.
- Mathews, S. C. et al. (2019). Digital health: A path to validation. *NPJ Digital Medicine*, 2(1), 38.
- Dias, D. and Paulo Silva Cunha, J. (2018). Wearable health devices -- vital sign monitoring, systems and technologies. *Sensors*, 18(8), 2414.
- Kadambi, A. et al. (2021). Achieving fairness in medical devices. *Science*, 372(6537), 30-31.

Declarations

Funding

Supported by the Swiss National Science Foundation (grant 320030_204826) and the Swiss Innovation Agency Innosuisse (grant 48342.1 IP-LS). No funder influenced study design, device selection, data analysis, or the decision to publish.

Conflict of Interest

The author declares no competing interests and has not received funding, consulting fees, or

devices from any medical device manufacturer included in this study.

Data Availability Statement

The device evaluation database, all prospective testing protocols and results, the DVQS calculator, and regulatory submission analyses are available from the author on reasonable request. Individual device manufacturer submissions are subject to confidentiality restrictions and cannot be shared without manufacturer consent.

Ethical Approval

Prospective device testing was approved by the Cantonal Ethics Committee of Zurich (BASEC 2019-00486), Amsterdam UMC Medical Ethics Committee (W19_384), and Hospital Clinic Barcelona Ethics Committee (HCB/2019/0986). All participants provided written informed consent for device testing and data use.

Appendix A

DVQS Methodology and Prospective Testing Protocol

This appendix details the DVQS scoring rubric, the standardised prospective testing protocol applied across all 42 devices, and demographic diversity quotas.

Part I -- DVQS Dimensions

1a. Analytical Accuracy: Bland-Altman limits of agreement against reference standard

Normalised by clinical acceptability thresholds

1b. Clinical Accuracy: Sensitivity/specificity for intended diagnostic claim Against gold-standard clinical diagnosis

2a. Population Generalisability: Performance consistency across 6 demographic strata

Age, sex, ethnicity, BMI, skin tone, comorbidity

2b. Longitudinal Stability: Accuracy drift over 6-12 months of continuous use < 5% drift = full score

3a. Regulatory Compliance: Completeness of clinical evidence per EU MDR/FDA requirements

Expert-rated 0-1

3b. $DVQS = 0.25 \times \text{Analytical} + 0.25 \times \text{Clinical} + 0.20 \times \text{Generalisability} + 0.15 \times \text{Stability} + 0.15 \times \text{Regulatory}$

Part II -- Demographic Quotas for Prospective Testing

4a. Skin tone: Minimum 15% Fitzpatrick V-VI per device evaluation To detect optical sensing bias

4b. BMI: Minimum 20% BMI > 30 and 10% BMI > 35 To detect body habitus effects

5a. Age: Minimum 25% aged 65+ and 10% aged 80+ To detect age-related performance changes

5b. Sex: 40-60% female per evaluation Balanced by design