Neuroscience Research and Clinical Practice

https://nrcp.cultechpub.com/nrcp

E-ISSN: 3141-1037

Vol. 2, No. 2, November 2026

DOI: 10.64229/8m5kv497

Review

When Digital Biomarkers Drift: A Critical Review of Temporal, Technological, Behavioral, and Population Instability in AI Models for Neurodegenerative Disease

Olaitan Ebenezer Oluwadare[1,*], Anyebe Daniel Ameh[2], Abdul-Basid Yakubu[3], Faith Uzochi Sunday[2]

[1]Applied Mathematics Department, School of Physics, Engineering, Mathematics and Computer Science, Delaware State University, Dover, Delaware, USA

[2]College of Professional Studies, University of Denver, Denver, Colorado, USA

[3]School of Natural Sciences, University of Chester, Chester, England, United Kingdom

[*]Corresponding author: Olaitan Ebenezer Oluwadare, olaitan.oluwadare89@gmail.com

Abstract

Digital biomarkers derived from smartphones, wearables, and connected sensors are increasingly combined with artificial intelligence (AI) to detect and monitor neurodegenerative diseases, yet the signals on which these models depend are not static. This critical review, based on a structured search of biomedical, engineering, and regulatory literature, examines four interacting sources of instability: temporal drift arising from disease progression, aging, treatment, and symptom fluctuation; technological drift caused by changes in devices, sensors, software, and processing pipelines; behavioral and contextual drift associated with routines, adherence, engagement, and missing data; and population drift that arises when deployment cohorts differ from narrow development cohorts. We map these sources onto established machine-learning concepts, including covariate shift, label shift, and concept drift, and summarize disease- and modality-specific evidence in Parkinson disease, Alzheimer disease, Huntington disease, and related conditions. Across the literature, separating genuine disease progression from measurement-related, behavioral, and population-related change remains difficult, and validation based on random data splits, short follow-up, homogeneous cohorts, and single devices can overestimate durability. These limitations can produce false deterioration signals, unstable individual predictions, and inequitable performance after deployment. We propose a Detect → Attribute → Quantify → Adapt → Validate → Monitor (DAQAV-M) framework with illustrative operational criteria, attribution strategies, and version-controlled updating rules, and we position it relative to existing dataset-shift, model-monitoring, and regulatory lifecycle approaches. Trustworthy digital biomarkers require lifecycle validation that separates biomarker validity from model performance and tests stability within individuals, across technologies, and across diverse populations.

Keywords

Digital biomarkers, Neurodegenerative diseases, Artificial intelligence, Biomarker drift, Dataset shift, Longitudinal validation

Article History

Received: 18 September 2026 Revised: 10 October 2026

Accepted: 10 November 2026 Available Online: 16 November 2026

Copyright

© 2026 by the authors. This article is published by the Cultech Publishing Sdn. Bhd. under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/

1. Introduction: The Stability Problem in Digital Biomarker Science

1.1 From Static Biomarkers to Continuously Evolving Digital Phenotypes

Digital biomarkers have emerged as a distinct approach to characterizing health through data generated by digital technologies, including smartphones, wearable devices, connected sensors, and other computational platforms [1-3]. Definitions of the term still vary considerably across the biomedical literature [4], which is one reason why the conditions under which a digital measure remains valid deserve explicit attention. Unlike conventional biomarkers that are often obtained at defined clinical encounters, digital biomarkers can capture physiological, motor, cognitive, and behavioral signals repeatedly or continuously in everyday environments. In neurodegenerative disease, this capability is particularly relevant because clinically meaningful changes may extend beyond what can be observed during brief, episodic assessments. Smartphone and wearable technologies, for example, can capture features related to gait, movement, speech, activity, sleep, and cognitive performance through active assessments or passive sensing [5,6].

The growing volume and complexity of these data have increasingly positioned artificial intelligence and machine learning as essential analytical tools. Rather than relying on a single predefined measurement, AI models can integrate multidimensional signals and identify patterns that may be difficult to recognize through conventional statistical approaches. Recent work has consequently expanded the digital biomarker landscape across neurodegenerative disorders, including Parkinson disease, Alzheimer disease, mild cognitive impairment, Huntington disease, and other conditions [7-10]. Yet this apparent advantage introduces an important conceptual challenge: the digital phenotype is not a fixed biological measurement. It is produced through an interaction between the individual, disease state, behavior, technology, the population in which a model was developed, and the surrounding environment.

1.2 Why Predictive Performance May Not Remain Stable Over Time

This dynamic nature creates the possibility of digital biomarker drift: changes in the distribution, meaning, or reliability of digitally captured signals as their biological, behavioral, technological, or environmental conditions evolve. Consequently, a signal strongly associated with a clinical state during model development may not retain the same relationship after deployment.

This problem is particularly important for AI models because predictive performance depends not only on the algorithm but also on the data-generating environment in which the algorithm operates. Evidence from machine-learning research demonstrates that distributional changes over time can produce measurable deterioration in model performance, while retraining or adaptation may partially restore performance [11-13]. In a clinical example outside neurology, prediction models that performed well within the trial in which they were developed performed close to chance when applied to independent trials of the same condition [14]. In digital health, therefore, high accuracy during retrospective validation cannot by itself establish long-term reliability. Regulatory guidance similarly emphasizes that digitally derived measures require evidence that the technology is fit for purpose and that potential sources of error affecting validity are identified [15].

1.3 Scope, Objectives and Central Argument of the Review

This review critically examines the instability of AI-enabled digital biomarkers for neurodegenerative disease through three interconnected individual-level dimensions: temporal, technological, and behavioral instability. Temporal instability concerns changes in the patient and disease trajectory; technological instability concerns changes in the systems used to generate and process measurements; and behavioral instability concerns changes in how individuals interact with their devices and environments. In response to an important gap in the existing literature, we also treat population-level drift, meaning systematic differences between the cohort in which a model is developed and the populations in which it is deployed, as a fourth, analytically separate dimension that compounds the other three. These dimensions are considered together because they may interact and become difficult to disentangle in real-world longitudinal monitoring.

The central argument is that digital biomarkers should not be conceptualized simply as fixed predictors whose validity is established once through conventional model validation. Rather, they should be understood as dynamic measurement systems whose reliability must be evaluated across changing biological, technological, behavioral, and population conditions. The review therefore moves beyond the question of whether an AI model can predict disease to a more consequential question: whether the model continues to measure what it is intended to measure as the patient, technology, and context change over time. To avoid repeating this argument throughout, subsequent sections focus on the mechanisms, evidence, and methods that follow from it.

1.4 Review Methodology

This article is a critical narrative review rather than a systematic review or meta-analysis. Its aim is conceptual synthesis and methodological appraisal, not a quantitative estimate of effect sizes. To make the scope and evidence base transparent, the literature was identified through a structured search strategy. PubMed/MEDLINE, Scopus, Web of Science, and IEEE Xplore were searched for articles published between January 2009 and May 2026, supplemented by Google Scholar, arXiv, and medRxiv for machine-learning methods literature, and by the websites of the U.S. Food and Drug Administration, the European Medicines Agency, and the European Union for regulatory documents.

Search strings combined three concept blocks with Boolean operators: (i) disease terms ("neurodegenerative", "Parkinson*", "Alzheimer*", "mild cognitive impairment", "dementia", "Huntington*", "amyotrophic lateral sclerosis", "multiple sclerosis"); (ii) measurement terms ("digital biomarker*", "wearable*", "smartphone", "smartwatch", "accelerometer", "remote monitoring", "digital phenotyping", "speech", "gait", "sleep"); and (iii) instability terms ("drift", "dataset shift", "distribution shift", "covariate shift", "concept drift", "temporal validation", "external validation", "generalizability", "recalibration", "model updating", "continual learning", "domain adaptation", "federated learning", "missing data", "adherence", "equity"). Methods papers on dataset shift and model monitoring were included regardless of clinical domain when they informed the framework.

Inclusion criteria were: English-language peer-reviewed original studies, reviews, consensus statements, conference proceedings, or regulatory guidance that (a) reported development, validation, or longitudinal use of digital measures in neurodegenerative or closely related neurological conditions, or (b) addressed the detection, attribution, or management of dataset shift in clinical prediction models. Exclusion criteria were: conference abstracts without full text, studies of digital therapeutics without a measurement component, and studies that did not report how the digital measure was obtained. Titles and abstracts were screened by two authors (O.E.O. and A.D.A.), full texts were assessed against the criteria, and disagreements were resolved by discussion with a third author. Backward and forward citation searching of key articles supplemented the database search. Because the review is critical rather than systematic, no formal risk-of-bias scoring or PRISMA flow diagram was produced; instead, the evidence was appraised qualitatively for study design, follow-up duration, device diversity, population diversity, and validation strategy, and these attributes are summarized in the disease- and modality-specific evidence table presented in Section 3.

2. What Does It Mean for a Digital Biomarker to "Drift"?

2.1 Distinguishing Biological Change from Measurement and Model Drift

The interpretation of longitudinal digital biomarkers depends on a distinction that is often implicit in clinical research but becomes considerably more difficult when measurements are collected continuously outside controlled clinical settings. A change in a digital signal does not necessarily represent a change in the underlying disease state. It may instead reflect an alteration in the way the signal was generated, captured, processed, or interpreted. For AI-enabled digital biomarkers, this distinction is fundamental because the observed output is the product of several interacting processes rather than a direct readout of pathology.

Terminology. Throughout this review, "drift" denotes an observable statistical phenomenon: a change over time, or between settings, in the joint distribution of digital features, reference outcomes, or model outputs. "Biological progression", "technological change", "behavioral change", and "population difference" denote candidate causes of that phenomenon. Temporal drift is therefore a description of when and how the data change, whereas biological progression is one possible explanation for why they change. Keeping phenomenon and cause separate is what makes the attribution step of the proposed framework (Section 5.5) possible.

Biological change represents the most clinically intuitive source of variation. In neurodegenerative disease, changes in motor function, cognition, speech, sleep, or daily activity may accompany disease progression. However, even biological trajectories are rarely uniform. Symptoms can fluctuate, progression can differ substantially between individuals, and treatment may modify the expression of disease without eliminating its underlying pathology. The challenge is therefore not simply to determine whether a digital signal changes, but to establish whether the observed change corresponds to a meaningful change in the biological construct that the biomarker is intended to represent.

Measurement-related change introduces a different problem. Digital biomarkers are generated through technologies that have their own characteristics and sources of variability. A wearable sensor may be positioned differently from one assessment period to another, a smartphone may be replaced by a device with different hardware characteristics, or an operating system update may alter how sensor data are collected or processed [16,17]. These effects may occur without

any corresponding biological change in the individual. If they are not recognized, an AI model may interpret measurement variation as evidence of altered disease status.

Behavioral change occupies an intermediate position because it can be both a meaningful component of the phenotype and a source of measurement distortion. An individual with a neurodegenerative disorder may gradually change physical activity, communication patterns, sleep routines, medication adherence, or the frequency with which digital assessments are completed. Such changes may reflect genuine functional consequences of disease, but they may also arise from changes in motivation, living circumstances, familiarity with technology, or the burden associated with repeated assessments [18]. This creates a fundamental attribution problem: the model may detect a real change in behavior without necessarily identifying its cause.

At the computational level, changes can also occur in the data distribution or in the model itself. An AI system developed using data collected from one population, device configuration, clinical environment, or period may encounter a different distribution after deployment. This phenomenon is commonly discussed within the broader machine learning literature as distribution shift or concept drift, although these terms describe related but distinct mechanisms [19-21]. Importantly, deterioration in predictive performance does not necessarily imply that the original model was poorly constructed. The environment in which the model operates may simply have changed.

If an algorithm produces a higher risk score several months after its initial assessment, the increase could reflect disease progression, a treatment-related change, altered behavior, a technological modification, a change in the population being monitored, or some combination of these factors. Treating every deviation as biological deterioration risks converting an unstable measurement process into an apparently precise clinical signal. Conversely, dismissing all variation as technical noise could conceal genuine changes in disease state. Reliable digital biomarkers must therefore distinguish changes in the person from changes in the system through which the person is being measured.

2.2 A Taxonomy of Digital Biomarker Drift and Its Relationship to Machine-Learning Terminology

Digital biomarker instability is better understood as a multidimensional phenomenon than as a single form of model deterioration. To give the taxonomy greater conceptual precision, let X denote the vector of digital features, Y the clinical reference outcome (for example, a rating-scale score or diagnostic status), and C the measurement context (device, software version, wear location, setting, and population). A digital biomarker model learns an approximation of $P(Y \mid X)$ under the development distribution $P_{dev}(X, Y, C)$. Drift occurs when the deployment distribution $P_{dep}(X, Y, C)$ differs from P_{dev}. Standard machine-learning terminology distinguishes three idealized cases [19,20,22]: covariate shift, in which $P(X)$ changes while $P(Y \mid X)$ is stable; label (prior-probability) shift, in which $P(Y)$ changes while $P(X \mid Y)$ is stable, for example when a model developed in early-stage patients is deployed in a clinic with more advanced disease; and concept drift, in which $P(Y \mid X)$ itself changes, so that the same feature values now imply a different clinical state [21].

The clinical sources of drift map onto these cases in an informative but imperfect way. Biological progression typically changes $P(Y)$ and may change $P(Y \mid X)$ as the phenotype evolves. Technological change alters the measurement function linking the latent clinical state to X, which appears as covariate shift when it is uniform and as concept drift when it interacts with disease (for example, when step-count error is larger in severe disease). Behavioral and contextual change can produce any of the three. Population drift is a structured form of covariate and label shift arising from cohort differences. Missingness drift, a change over time in the probability that data are observed, $P(R \mid X, Y)$, where R indicates observation, is a further mechanism that the conventional triad does not name but that is common in longitudinal monitoring (Section 3.5). Model drift, finally, is not an independent cause; it is the downstream manifestation of any of these changes in model discrimination, calibration, or stability.

The major dimensions of digital biomarker drift, their causal mechanisms, observable indicators, relationship to machine-learning terminology, potential effects, and corresponding detection and mitigation strategies are summarized in Table 1. The last column links each dimension directly to the stages of the framework proposed in Section 5.5.

Table 1. Major dimensions of digital biomarker drift in neurodegenerative disease: causal mechanisms, observable indicators, relationship to machine-learning terminology, and detection and mitigation strategies.

Drift dimension	Main drivers (causal mechanism)	Examples and observable indicators	Relationship to ML terminology	Potential effect on AI-based biomarkers	Detection and mitigation (DAQAV-M stage)
Temporal drift (phenomenon; principal cause: biological progression)	Disease progression, symptom fluctuation, aging, treatment changes	Changes in gait, speech, cognition, sleep, and activity; gradual within-person trends; stage-dependent feature-outcome relationships	Label shift (change in stage mix); concept drift when feature meaning changes with stage	Altered feature distributions or feature-outcome relationships	Temporal holdout validation; within-person mixed-effects trajectories; stage-stratified calibration; clinical anchoring (Detect, Attribute)
Technological drift	Device replacement, sensor differences, software/firmware updates, preprocessing changes	New smartphones or wearables, changes in sampling or data processing; abrupt change points aligned with version changes and shared across users	Covariate shift; concept drift when measurement error interacts with disease state	Measurement shift, reduced calibration, and loss of generalizability	Device and software version logging; device-transition bridging; cross-device validation; harmonization or recalibration (Attribute, Adapt)
Behavioral/ contextual drift	Changes in routine, adherence, motivation, environment, caregiver involvement	Altered activity, sleep, device use, or assessment participation	Covariate shift; concept drift when behavior acts as a proxy for disease	Behavioral proxies and context-dependent measurement changes	Context metadata; adherence monitoring; audits of proxy features; context-adjusted or robust features (Attribute, Quantify)
Population drift (diversity deficit)	Development cohort systematically unlike the deployment population (e.g., WYHU sampling [23]); differences in age, ancestry, language, education, socioeconomic position, comorbidity, and technology access	Deployment covariates outside the development range; subgroup performance gaps	Covariate and label shift; concept drift when feature meaning differs between groups	Systematic, group-concentrated error; inequitable performance	Cross-population validation; subgroup reporting; domain-classifier or density-ratio monitoring; targeted data collection; federated training (Detect, Attribute, Adapt)
Missingness drift	Device non-use, informative dropout, caregiver-mediated collection, technical failures, stage-dependent missingness	Rising or differential missing-data rates; declining wear time or task completion	Change in the missingness mechanism $P(R \mid X, Y)$; missing not at random	Biased progression estimates; reduced data volume misread as decline	Report missingness by stage and subgroup; joint or pattern-mixture models; sensitivity analyses; missingness as a monitored signal (Detect, Quantify)
Model drift	Changes in input distributions or feature-outcome relationships	Performance deterioration after deployment	Downstream manifestation of any of the shifts above, not an independent cause	Reduced predictive stability and reliability	Discrimination and calibration monitoring; control charts; recalibration or updating under version control (Quantify, Adapt, Validate)
Mixed drift	Simultaneous biological, technological, population, and behavioral changes	Disease progression coinciding with device replacement and altered technology use	Combination of shift types	Complex distribution shift that is difficult to attribute to a single source	Multi-source attribution (metadata, within-person controls, causal designs); conservative adaptation; human review (Attribute)

Temporal drift refers to changes in the characteristics of the data or their relationship with the clinical outcome as time passes. In neurodegenerative disease, time is inseparable from biological change because patients may experience progression, symptom fluctuation, treatment modification, or changes in functional capacity. Importantly, temporal drift should not be interpreted exclusively as deterioration. Some features may become more informative as disease manifestations become more pronounced, whereas others may lose their discriminative value as the phenotype changes.

Technological drift arises when the systems used to acquire, transmit, preprocess, or analyze digital data change. Smartphones and wearable devices are not standardized measurement instruments in the same way as many conventional laboratory assays. Differences in sensors, sampling characteristics, hardware, firmware, operating systems, software versions, and preprocessing procedures can alter the characteristics of recorded data, so that an apparent change in a digital biomarker reflects the measurement platform rather than the patient.

Behavioral or contextual drift concerns changes in how individuals generate and interact with digital data and in the circumstances surrounding measurement. Physical activity, sleep, communication, mobility, adherence to prescribed assessments, and patterns of device use may change over time for reasons that may or may not be related to disease. This makes behavioral and contextual drift particularly difficult to isolate from biological progression.

Population drift arises when the individuals in whom a biomarker is deployed differ systematically from those in whom it was developed. Many development cohorts for digital biomarkers are recruited through research centers or app stores and over-represent participants who own recent devices, are digitally literate, and have the time and health to participate. Rudroff has described this pattern as a "WYHU" (Western, Young, Healthy, University-affiliated) sampling bias and has argued that it constitutes one half of a "dual diversity" problem: a population-level diversity deficit that is analytically distinct from, but compounds, the individual-level neurobiological heterogeneity that characterizes neurodegenerative disease [23]. Population drift therefore deserves its own place in the taxonomy rather than being treated only as an implementation or equity concern.

Model drift occurs when the predictive relationship learned by an AI system becomes less appropriate for the data encountered during subsequent use. If the distribution of inputs changes, or if the relationship between those inputs and the clinical outcome changes, a previously calibrated model may become systematically less reliable [24,25]. In many cases, model drift is the downstream manifestation of biological, technological, behavioral, population, or contextual changes that the original development dataset did not capture.

These dimensions can interact in ways that make attribution considerably more difficult. Consider a patient whose mobility declines as a neurodegenerative disorder progresses. Reduced mobility may alter the patient's daily activity and patterns of smartphone use, producing behavioral changes in the data. If the patient subsequently changes from one wearable device to another, the characteristics of the recorded movement signals may also change. A decline in predictive performance observed after deployment could therefore be attributed incorrectly to the model when the underlying problem is a compound change in the system being measured. The reverse is also possible: technological and behavioral changes may conceal genuine biological progression. Drift should therefore be conceptualized as a system-level phenomenon rather than a collection of isolated technical errors.

2.3 Why Neurodegenerative Disease Presents a Special Challenge

Neurodegenerative diseases present a particular challenge for digital biomarker development because the biological system being measured is itself changing. Disease trajectories are progressive but highly heterogeneous, with substantial differences between individuals in the timing and severity of cognitive, motor, behavioral, and functional changes. Symptoms may also fluctuate across days or within the same day, influenced by disease characteristics, fatigue, treatment timing, and other contextual factors [16]. Consequently, a digital signal observed at one point may not have the same clinical meaning later in the disease course. AI models must distinguish meaningful longitudinal changes from normal within-person variability and determine whether a changing signal represents progression, fluctuation, or another transient influence.

Two levels of heterogeneity should be distinguished here. Individual-level heterogeneity refers to differences between and within patients in disease trajectory and symptom expression. Population-level heterogeneity refers to differences in the composition of cohorts, including age, sex, ancestry, socioeconomic position, language, comorbidity, and access to technology. Statistical relationships estimated across a group need not hold within any given individual [26], and relationships estimated in one population need not hold in another [23]. Neuroimaging work comparing fatigue in multiple sclerosis and long COVID illustrates the point: diagnostically distinct populations that share a symptom label showed different patterns of frontal-striatal glucose metabolism [27], suggesting that a single "generalizable" model of a digital fatigue or motor signal could flatten genuinely different signal architectures [28].

Aging, medication effects, and comorbidities further complicate this distinction. Neurodegenerative diseases frequently occur in older populations in whom mobility, sleep, cognition, sensory function, and activity can change independently of neurodegenerative pathology. Cardiovascular, musculoskeletal, psychiatric, and sleep-related conditions may similarly influence digital measures. Treatment introduces another source of variation because medication initiation, dose changes, or treatment response can alter the digital phenotype without necessarily representing a corresponding change in underlying disease progression. The challenge is not simply to control for these factors statistically, but to determine which observed changes retain sufficient disease specificity to support reliable longitudinal interpretation.

The problem becomes still more complex as disease progression affects functional independence and the way individuals interact with technology. Declining mobility or cognition may change daily routines, reduce participation in active assessments, alter patterns of smartphone or wearable use, or increase reliance on caregivers. Changes in data availability or device engagement may therefore carry clinical information while simultaneously introducing measurement instability. For digital biomarkers in neurodegenerative disease, reliability cannot be defined as the preservation of an identical signal over time because the patient and disease are not static; the critical requirement is that the measurement system remains capable of separating clinically meaningful biological change from variation arising from aging, treatment, comorbidity, behavior, and measurement conditions.

2.4 From Accuracy at Validation to Reliability in the Real World

AI-based digital biomarkers are commonly evaluated using accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and related metrics. These measures describe performance under defined conditions but provide limited evidence of longitudinal reliability. A biomarker may retain strong performance shortly after development yet become less reliable as disease characteristics evolve, treatment changes, devices are replaced, or new patterns of behavior emerge. These circumstances raise questions that conventional validation metrics alone cannot answer: does the biomarker remain valid six months or two years after deployment, does it remain comparable when the measurement technology changes, and can its outputs still be interpreted when the individual's behavior changes? Real-world reliability therefore requires evaluation of stability, calibration, generalizability, and interpretability over time [24].

2.5 Distinguishing Biomarker Validity from Model Performance

A decrease in AUC, calibration, or predictive accuracy does not necessarily mean that the digital biomarker itself has lost biological validity, and stable predictive performance does not establish construct validity. These are different properties that are established by different evidence. Following the V3 and V3+ frameworks for sensor-based digital health technologies [29,30], at least five layers can be distinguished. Verification and analytical validity concern whether the sensor and algorithm measure the intended physical or behavioral quantity accurately and reproducibly (for example, whether a step count or gait speed estimate agrees with a reference system). Clinical validity concerns whether the measure is associated with the clinical state or outcome of interest in the intended population. Construct validity concerns whether the measure reflects the theoretical construct it is claimed to represent (for example, bradykinesia rather than general physical activity) and not a proxy such as device ownership or study engagement. Clinical utility concerns whether using the measure improves decisions or outcomes. Model performance, finally, describes how well an algorithm built on the measure predicts a reference outcome in a particular dataset.

Drift can affect each layer separately. A firmware update may impair analytical validity while leaving the underlying construct unchanged; recalibration can then restore model performance without any change in biological meaning. Conversely, progression may change the clinical meaning of a feature (concept drift) while analytical validity remains intact; here recalibration may restore apparent performance but conceal a loss of construct validity. A model may also maintain a high AUC because it has learned a behavioral proxy that happens to be correlated with disease in the development cohort, which is evidence of performance but not of construct validity. The framework proposed in Section 5.5 therefore asks which validity layer has been affected before deciding how to respond.

2.6 Clinical Endpoints, Ground Truth, and Reference Standards

Attribution of change in a digital biomarker depends heavily on the reference standard used to define disease progression. In neurodegenerative disease, ground truth may be defined by clinical rating scales (for example, the MDS-UPDRS in Parkinson disease or the Unified Huntington's Disease Rating Scale), clinician global assessment, structural or molecular imaging, fluid biomarkers, neuropsychological testing, functional outcomes, or composite endpoints. Each has its own error structure. Rating scales are subject to inter-rater and within-subject variability; a real-world analysis of repeated MDS-UPDRS assessments found that within-subject reliability was limited, which constrains its use as a criterion for short-term change [31]. Imaging and fluid biomarkers index pathology but may relate only loosely to day-to-day function. Cognitive tests show practice effects with repetition. Diagnostic criteria themselves are revised over time.

The reference outcome can therefore drift as well. When raters, scale versions, assay platforms, or diagnostic criteria change during a study, an apparent loss of agreement between a digital biomarker and its reference may reflect instability in the reference rather than in the digital measure. Studies should report which reference standard was used, how its reliability was assessed, whether it changed during follow-up, and whether digital measures were evaluated against multiple references. Where digital measures may be more sensitive than episodic clinical scales, agreement with the scale is an imperfect criterion, and complementary evidence such as anchoring to clinically meaningful events, convergence across modalities, and responsiveness to known interventions becomes important.

3. Why Digital Biomarkers Drift: Temporal, Technological, Behavioral, Population, and Missingness Drivers

This section summarizes the mechanisms of each driver and the disease- and modality-specific evidence that illustrates them. The evidence is organized by disease, biomarker modality, technology, analytic method, drift type, clinical consequence, and mitigation strategy in Table 2.

3.1 Temporal Drift: When the Patient and Disease Trajectory Change

Temporal drift occurs when the characteristics of digitally captured data, or their relationship with a clinical outcome, change over time. Disease progression can alter gait, speech, movement, cognition, sleep, activity, and daily functioning, while the rate and pattern of change vary substantially between individuals [5]. An AI model trained predominantly on

an early disease phenotype may therefore encounter later clinical states that were poorly represented during development.

Disease-specific evidence illustrates this. In early Parkinson disease, the WATCH-PD study followed participants with a smartwatch and smartphone over 12 months and reported longitudinal change over that period in several digital measures [32]. Longitudinal wearable gait and postural data combined with machine learning have also been used to identify motor progression in Parkinson disease [33], and proposals to stage Parkinson disease using digital health technology [34] presuppose that stage-specific feature relationships can be characterized and kept stable. In UK Biobank, wrist accelerometry recorded years before diagnosis distinguished individuals who later received a Parkinson disease diagnosis, showing that digital signals evolve during the prodromal phase [35]. In amyotrophic lateral sclerosis, frequent remote speech recordings tracked bulbar decline over time [36], and in Alzheimer disease and mild cognitive impairment, remote unsupervised memory assessments on mobile devices detected cognitive impairment in memory-clinic samples [37,38]. Each of these findings implies that the relationship between digital features and clinical state is stage-dependent, which is the definition of temporal concept drift.

Symptom fluctuation introduces an additional challenge. Continuous digital monitoring can capture variations that episodic clinical assessments may miss, but not every short-term change represents disease progression [16]. Fatigue, sleep, medication timing, daily activity, and other transient factors can influence digital measurements. Treatment changes may similarly alter motor or behavioral signals without corresponding to a change in underlying pathology. Repeated active tasks introduce practice effects, which can mask decline in cognitive measures or exaggerate improvement after a change in task version. Aging and changing functional capacity further modify the data-generating process, and changes in adherence and missing data may reflect clinically meaningful changes while altering the information available to the model.

Temporal drift should therefore not automatically be interpreted as model failure. Some change is expected because the underlying biological system is evolving. For neurodegenerative disease, the goal of a robust digital biomarker is not to produce an identical signal over time, but to maintain a clinically interpretable relationship between changing digital observations and changing disease state.

3.2 Technological Drift: When the Measurement System Changes

Technological drift occurs when changes in the devices, software, sensors, or data processing systems used to generate digital biomarkers alter the characteristics of the measurements over time. Digital biomarkers are often dependent on consumer technologies that are routinely replaced and updated, so a patient may remain biologically stable while the measurement system changes beneath the AI model.

Changes in smartphone models and wearable devices can affect sensor characteristics, sampling behavior, and the quality or quantity of recorded data. A systematic study of motion sensors across many smartphone and smartwatch models documented substantial heterogeneity in sampling rates, sampling stability, and sensor biases between devices, and showed that this heterogeneity degraded activity-recognition performance when models were transferred across devices [39]. Even when two devices are designed to measure the same phenomenon, differences in hardware and sensor configuration may produce systematically different observations [17]. In Parkinson disease, the reliability of consumer-smartwatch step counts has been evaluated separately across disease phenotypes, stages, and severities [40], an approach that is necessary because technological error can interact with disease state rather than acting as uniform noise. A patient survey examined both the acceptability and the technical reliability of wearable devices used for Parkinson disease monitoring [41], underlining that device reliability in daily life cannot be assumed. The Mobilise-D consortium validated wearable-derived real-world walking speed separately across several clinical conditions, including Parkinson disease [42], and wearable gait estimation in Huntington disease has been compared between clinic and free-living conditions [43], reflecting the recognition that laboratory performance does not guarantee free-living performance [16].

Software and platform changes create another source of instability. Operating system updates, firmware modifications, changes to application architecture, and alterations in data collection procedures can affect how sensor information is recorded, transmitted, or processed. Proprietary preprocessing is particularly important because transformations applied before data reach the researcher may not always be fully transparent or remain constant across software versions. This creates a reproducibility problem in which apparently identical digital measurements may not remain technically equivalent across time or platforms. Large foundation models pretrained on wearable signals from many users and devices [44] may reduce sensitivity to some device differences, but they also inherit the device mix and population composition of their pretraining data, which makes documentation of that composition essential.

Technological drift therefore challenges the assumption that repeated digital measurements represent the same underlying measurement process. Validation must establish not only whether a biomarker performs well with a particular device, but whether its interpretation remains stable when the technology generating the measurement changes [45]. Cross-device validation, transparent data processing, logging of device and software versions, and monitoring for technological changes should consequently be considered essential components of longitudinal evaluation.

3.3 Behavioral and Contextual Drift: When Patients Change How They Generate Data

Behavioral and contextual drift arises when changes in how individuals live, interact with technology, or respond to their environment alter the digital signals available to an AI model [18]. Physical activity, sleep, social interaction, smartphone use, medication routines, and adherence to active assessments can all change over time. In neurodegenerative disease, such changes may result directly from disease progression, but they may also reflect fatigue, changes in living circumstances, caregiver involvement, motivation, digital literacy, or adaptation to the technology itself.

Empirical studies show that engagement is itself dynamic. In a health-system cohort, engagement with connected wearable devices was not sustained uniformly across patients, and time to disengagement could be modeled with survival analysis [46]; a pilot cohort of older adults documented variation in adherence to wearable activity monitoring [47]; and a qualitative meta-synthesis of long-term wearable research in dementia described barriers to enrollment and sustained adherence relating to participants, caregivers, and the devices themselves [48]. Speech-based measures in Huntington disease [49] and Parkinson disease [45] depend on recording conditions and task performance, which can change with routine and environment as well as with disease.

This raises the problem of behavioral proxies. An AI model may appear to detect disease because it has learned characteristics associated with how patients behave rather than features directly related to the underlying pathology. Such associations can produce strong predictive performance within the development population while becoming unreliable when behavior, environment, or patterns of technology use change. Behavioral and contextual drift should consequently be treated as part of the measurement process rather than dismissed as noise, and robust digital biomarkers should incorporate sufficient contextual information to interpret unavoidable behavioral changes appropriately.

3.4 Population Drift and the Dual Diversity Problem

Population drift is distinct from within-patient temporal drift. A model can be perfectly stable within each of its development participants and still fail when deployed in a population whose age, ancestry, language, socioeconomic position, comorbidity profile, disease subtype distribution, or technology access differs from the development cohort. In the terms of Section 2.2, population drift produces covariate shift (different feature distributions), label shift (different disease prevalence or stage mix), and, when the meaning of a feature differs between groups, concept drift.

Several features of digital biomarker research make this especially likely. App-based and smartwatch-based studies recruit people who own and use the relevant devices; research-center cohorts over-represent participants who are younger, healthier, more educated, and closer to academic centers [23]; and large volunteer biobanks used for prodromal discovery [35] were not designed to be representative of the populations in which screening would eventually be deployed. Meanwhile, the populations most affected by neurodegenerative disease are older and more diverse, and access to and use of digital technology among older adults varies with sociodemographic factors [50,51]. In a community-based cohort of older adults, the digital divide was itself associated with cognitive disparities [52]. For cognitive digital biomarkers, this implies that features describing device use may partly encode socioeconomic and educational differences rather than neurodegeneration.

Following Rudroff's dual diversity framing [23], two robustness claims should be kept separate. Robustness to individual heterogeneity means that a feature remains interpretable across patients with different trajectories within the sampled population. Robustness to population diversity means that the feature retains its meaning across populations that were not sampled. Evidence for the first does not establish the second. Methods to detect population drift after deployment, such as density-ratio and novelty-detection approaches, are available [53], but they detect difference rather than explain it, which is why population shift is given its own branch in the attribution stage of the proposed framework.

3.5 Missing-Data Drift

Missing data in longitudinal digital monitoring are rarely missing completely at random. Data may be missing at random given observed variables (for example, lower wear time on weekends), or missing not at random, when the probability of missingness depends on the unobserved value itself, for example when patients stop completing gait tasks on days when their mobility is worst [54]. In neurodegenerative disease, several mechanisms make missingness informative: device non-use as dexterity, cognition, or motivation declines; informative dropout of participants who progress fastest; transition to caregiver-mediated data collection, which changes who operates the device and how; device loss, battery, or synchronization failures, which may be more frequent among participants with limited technical support [41]; and differential missingness across disease stages [47,48].

Missingness drift has three consequences for AI models. First, if missingness increases with disease severity and models are trained on available data only, estimates of progression will be biased toward stability. Second, if models use the quantity of data as a feature, a decline in engagement for non-clinical reasons can be misread as deterioration. Third, when missingness patterns differ between development and deployment populations, missingness becomes a vector of population drift. Recommended practice includes reporting missingness by time, stage, and subgroup; modeling missingness explicitly through joint models of the outcome and dropout, pattern-mixture models, or multiple

imputation under stated assumptions; conducting sensitivity analyses under plausible missing-not-at-random scenarios [54]; and monitoring missingness rates as a drift signal in their own right (Section 5.5).

3.6 Interactions Between Biological, Technological, Behavioral, and Population Drift

The most important challenge is that these sources of drift rarely occur independently. In real-world longitudinal monitoring, changes in disease, behavior, and technology can occur simultaneously and become statistically difficult to separate. Disease progression may reduce mobility, which changes physical activity and patterns of smartphone use. Reduced device engagement may then alter the amount and type of data available to the model, producing a distribution shift that is partly biological and partly behavioral. If a new device is introduced during the same period, differences in sensor characteristics may further modify the observed signal. If the deployment population also differs from the development cohort, the model may have no reference for how these changes typically co-occur in that population.

These interactions expose a limitation in treating drift as a purely technical problem. The relevant unit of analysis is not the AI model alone, but the relationship among the patient, disease, behavior, technology, population, and computational model. Detecting drift should therefore be accompanied by determining its likely source before deciding whether recalibration, adaptation, or further clinical assessment is appropriate. Table 2 summarizes the published evidence for these mechanisms by disease and modality.

Table 2. Disease- and modality-specific evidence on sources of drift in digital biomarkers for neurodegenerative and related neurological disease.

Disease	Biomarker modality	Technology	AI / analytic method	Drift type illustrated	Evidence	Clinical consequence	Mitigation strategy
Parkinson disease (early)	Gait, motor, tremor, activity	Smartwatch and smartphone (WATCH-PD)	Digital feature extraction; longitudinal change analysis	Temporal (progression)	Longitudinal change over 12 months in several digital measures [32]	Stage-dependent feature-outcome relationships; baseline models may miscalibrate later	Temporal holdout; within-person mixed-effects models
Parkinson disease (prodromal)	Wrist accelerometry, activity, sleep	Research-grade wrist accelerometer (UK Biobank)	Machine-learning classifiers	Temporal (prodromal); population	Accelerometry distinguished people later diagnosed with Parkinson disease years before diagnosis [35]	Volunteer cohort may not represent screening populations	Cross-population validation before screening use
Parkinson disease	Gait and posture	Body-worn inertial sensors	Machine-learning classification of progression	Temporal	Motor progression identified from longitudinal sensor data [33]	Progression labels depend on the reference standard	Report the reference standard and its reliability [31]
Parkinson disease	Daily step count	Consumer smartwatch	Proprietary on-device algorithm	Technological x biological	Step-count reliability examined by phenotype, stage, and severity [40]	Stage-dependent measurement error could mimic or mask decline	Stage-stratified analytical validation
Parkinson disease	Motor virtual exam (tremor, bradykinesia)	Research smartwatch	Algorithmic motor measures	Behavioral/contextual (unsupervised home use)	Frequent remote measurements reliable relative to in-clinic assessment [55]	Home context and adherence vary over time	Adherence and context monitoring; within-person reliability
Parkinson disease	Wearable monitoring (various)	Multiple wearable devices	Not applicable (survey)	Technological; missingness	Patient acceptability and technical reliability surveyed [41]	Technical failures produce missing or biased data	Device logs; missingness monitoring
Several conditions, including Parkinson disease	Real-world walking speed	Lower-back wearable (Mobilise-D)	Multi-stage gait algorithms	Contextual; population	Walking-speed estimation validated across several clinical conditions [42]	Algorithm accuracy may differ between conditions	Condition-specific validation
Alzheimer disease / mild cognitive impairment	Episodic memory	Smartphone (unsupervised remote tasks)	Composite scores; ROC analysis	Behavioral/ contextual; temporal (practice effects)	Remote composite detected cognitive impairment in memory-clinic samples [37]; scoping review of smartphone and tablet tools [38]	Practice effects and changing engagement could mask decline	Parallel test forms; modeling of practice effects; adherence tracking

Disease	Biomarker modality	Technology	AI / analytic method	Drift type illustrated	Evidence	Clinical consequence	Mitigation strategy
Alzheimer disease	Speech and language	Recorded narrative speech	Machine learning on linguistic features	Population (language, education); technological (recording)	Linguistic features identified Alzheimer disease in English narrative speech [56]	Transportability across languages and education levels not established	Language- and education-stratified validation
Older adults / cognitive decline	Device use and engagement	Smartphones and internet	Cohort analysis	Population (digital divide)	Digital divide associated with cognitive disparities [52]	Device-use features may encode socioeconomic position rather than neurodegeneration	Adjust for and stratify by access; audit proxy features
Huntington disease	Cognitive and motor tasks	Smartphone and smartwatch application	Feature-based scores	Cross-sectional (longitudinal drift untested)	Reliability and validity of a remote monitoring platform (cross-sectional) [57]	Longitudinal stability not yet established	Longitudinal and cross-device validation
Huntington disease	Gait	Wearable sensors	Gait-parameter algorithms	Contextual (clinic vs free-living)	Gait estimation compared in clinic and free-living conditions [43]	Clinic-derived models may misestimate at home	Context-specific validation
Huntington disease	Speech	Digital speech recordings	Acoustic feature analysis	Behavioral; technological	Digital assessment of speech [49]	Recording conditions affect features	Standardized recording protocols; metadata
Amyotrophic lateral sclerosis	Speech (bulbar function)	Remote smartphone recordings	Algorithmic speech metrics	Temporal	Frequent remote speech analysis tracked bulbar change [36]	Rapid progression requires frequent review of thresholds	Within-person trajectories
Neurological fatigue (multiple sclerosis, long COVID)	Fatigue-related physiological and behavioral signals	Wearables and AI (review); FDG-PET (mechanistic)	Various	Population (diagnostic heterogeneity)	Distinct metabolic patterns across fatigue populations [27]; review of digital fatigue biomarkers [28]	A single model may flatten disease-specific signal architectures	Disease-specific validation; dual diversity reporting [23]
Cross-domain (activity recognition)	Motion sensors	Many smartphone and smartwatch models	Activity-recognition classifiers	Technological	Device heterogeneity in sampling and sensor biases degraded cross-device performance [39]	Device replacement can cause false change	Cross-device validation; harmonization

4. Consequences for AI Models and Clinical Validity

4.1 Distribution Shift, Dataset Shift and Loss of Model Generalizability

AI-based digital biomarkers assume that deployment data remain sufficiently comparable to development data. When patients, disease stage, behavior, technology, clinical setting, or population change, distribution shift can reduce the reliability of predictions even when the model architecture remains unchanged [11,12]. The formal types of shift relevant to digital biomarkers were defined in Section 2.2 and Table 1; here we focus on their consequences. Although these concepts are computationally distinct, they can occur simultaneously in clinical practice. Empirical work in medical imaging has shown that monitoring aggregate model performance alone can miss clinically relevant data drift, and that the sensitivity of drift detectors depends strongly on sample size and the type of shift [58,59].

The clinical consequence is reduced generalizability. Models developed in selected populations, standardized devices, and controlled settings may encounter greater variability in routine use, and the resulting performance loss may not be apparent from original validation. For longitudinal digital biomarkers, generalizability also involves maintaining a meaningful relationship between digital observations and clinical outcomes across changing temporal, technological, behavioral, and population contexts. The critical issue is therefore not whether distribution shift can be eliminated, but whether it can be detected, characterized, and managed without compromising clinical meaning [12,60].

4.2 Hidden Failure Modes in Longitudinal Monitoring

The consequences of digital biomarker drift become most important when model outputs are interpreted as evidence of clinical change. A change in device, sensor behavior, physical activity, sleep pattern, or engagement with digital

assessments may alter the measured signal and lead an AI model to generate a higher risk estimate [24]. Such false deterioration is clinically problematic because it can trigger unnecessary assessments or interventions and may undermine confidence in the monitoring system. The opposite problem is equally important. If disease-related changes are masked by technological or behavioral variation, the model may fail to detect genuine deterioration, creating false reassurance.

Repeated predictions can also become unstable when the underlying data distribution changes. A patient may receive substantially different risk scores across successive periods without a corresponding clinical change, particularly when measurements are sparse, noisy, or affected by changes in context. Over time, inconsistent outputs can affect disease classification and progression estimates, particularly when the model was developed around relatively static diagnostic categories but is subsequently used to characterize a continuously evolving disease process.

These problems extend beyond statistical performance to the way digital information is incorporated into clinical decision making. Frequent alerts generated from unstable measurements may contribute to alarm fatigue, reducing the likelihood that clinically important signals receive appropriate attention [12]. Conversely, an apparently stable digital score may be given excessive weight despite uncertainty about whether the measurement process has remained comparable. A reduction in smartphone activity, for example, may reflect worsening function, but it may also result from a change in routine, device use, or technical accessibility. Without sufficient contextual information, the model may detect the change accurately while attributing it incorrectly.

4.3 Why Conventional Validation May Overestimate Durability

Many AI-based digital biomarkers are evaluated using development datasets divided into training and test sets, often through random splitting or cross-validation. Although appropriate for estimating performance under similar data conditions, these approaches can overestimate durability when observations from the same individuals, devices, or periods are distributed across training and testing sets [61]. For a biomarker intended for longitudinal monitoring, a more informative test is whether a model trained on earlier observations can maintain performance on genuinely later observations that reflect changes in disease, behavior, and technology [62].

The problem is reinforced by reliance on retrospective datasets, short follow-up periods, and single time-point assessments. A model validated over a relatively short interval may appear stable while failing to capture gradual changes in disease phenotype, treatment, adherence, or technology [13,63]. Limited device diversity creates a related problem because an algorithm may be tested on a narrow technological configuration and subsequently applied to devices with different sensors or processing characteristics.

Homogeneous study populations deserve separate treatment because they are conceptually distinct from the other limitations listed here. Short follow-up and random splitting limit evidence about temporal and individual-level stability; homogeneous populations limit evidence about population-level stability. Importantly, external validation does not by itself address the latter. A cohort can be external, recruited at a different site or in a different study, while being as demographically and technologically narrow as the development cohort. In this review, "external validation" refers to cross-site or cross-cohort generalizability in the conventional sense, and "cross-population validation" refers to deliberate testing in populations that differ in the demographic, socioeconomic, linguistic, clinical, and technological characteristics that are plausible sources of drift.

A further distinction concerns within-person and between-person generalizability. A model can discriminate well between patients, performing well across different individuals at one time point, while performing poorly when monitoring the same patient longitudinally, because between-person differences in a feature can be large relative to within-person change. Conversely, a personalized model may track an individual very stably yet fail to generalize to new patients. Group-level associations are not guaranteed to hold within individuals [26]. Validation designed for monitoring should therefore report both between-person metrics (for example, discrimination across participants) and within-person metrics (for example, test-retest reliability, intraclass correlation, minimal detectable change, responsiveness to known within-person change, and the stability of individual predictions over periods without clinical change).

These limitations do not make conventional validation methods inappropriate; rather, they demonstrate that predictive accuracy and longitudinal robustness are different properties. Demonstrating durability therefore requires temporal validation, external validation, cross-population validation, diverse technologies, longer follow-up, and testing under conditions that approximate the variability of clinical use [62,64,65].

4.4 Implications for Clinical Translation and Equity

The consequences of unstable digital biomarkers extend directly into clinical translation. In clinical decision making and remote monitoring, an unreliable change in a digital measure could influence whether a patient is assessed, referred, or considered to have deteriorated. In clinical trials, drift may affect endpoint consistency across participants or study periods, particularly when devices or software change during a study. Health systems consequently require evidence not only that a digital biomarker works, but that its performance can be monitored and maintained as conditions evolve.

There is also an important equity dimension, which follows directly from the population drift mechanism described in Section 3.4 rather than standing apart from it. Technology-dependent biomarkers may perform differently across populations because access to newer smartphones or wearable devices, patterns of digital engagement, internet connectivity, living environments, and familiarity with technology are not uniformly distributed [66]. Beyond device access, relevant factors include socioeconomic status and the ability to replace devices; educational background and digital literacy, which shape how active tasks are performed [50,51]; language and cultural context, which are critical for speech- and language-based biomarkers developed predominantly in English-speaking cohorts [56]; rural versus urban environments, which affect connectivity, synchronization, and patterns of mobility; caregiver support, which influences whether data are collected at all and by whom; and accessibility requirements, since tremor, visual impairment, hearing loss, or cognitive impairment may make standard device interfaces harder to use and thereby alter both data quality and missingness.

Obermeyer and colleagues showed how an algorithm that used health-care costs as a proxy for health needs systematically underestimated the needs of Black patients, because unequal access produced lower costs at the same level of illness [67]. The analogous risk for digital biomarkers is that data volume, engagement, or device-use features act as proxies for disease. When lower data availability reflects limited access rather than disease, it may be interpreted as poorer adherence or as decline, and behavioral patterns associated with limited technology use may be treated as disease-related signals. This is a population drift problem with an equity consequence: the error is not random but concentrated in groups under-represented in development cohorts.

Equitable implementation therefore requires attention to the entire measurement ecosystem rather than model performance alone. Validation across diverse populations, devices, environments, and patterns of technology use is necessary to determine whether a digital biomarker measures the intended clinical construct consistently rather than reproducing differences in access or digital behavior. Reporting performance by relevant subgroups is now recommended in reporting guidance for AI-based prediction models [65] and in international consensus guidelines for trustworthy medical AI [68].

4.5 Regulatory Implications of Post-Deployment Modification

Drift management implies that models will sometimes be modified after deployment, which raises regulatory questions that are specific to adaptive AI. A "locked" algorithm returns the same output for the same input and is changed only through discrete, reviewed updates; an "adaptive" algorithm changes its behavior through continued learning after deployment. Most regulated AI-enabled devices to date have been locked, in part because each substantive modification could otherwise require a new marketing submission [69].

Regulators have responded with mechanisms that allow pre-specified change. Predetermined change control plans (PCCPs), introduced as guiding principles by the U.S. Food and Drug Administration, Health Canada, and the UK Medicines and Healthcare products Regulatory Agency [70] and finalized in FDA guidance in December 2024 [71], allow a manufacturer to describe in advance the modifications it intends to make (for example, retraining on new data or recalibration for a new device), the protocol for developing, validating, and implementing them, and the assessment of their impact. FDA draft guidance on AI-enabled device software functions further emphasizes a total product lifecycle approach, including performance monitoring after deployment and transparency about model versions [72]. In the European Union, the Artificial Intelligence Act classifies many medical AI systems as high-risk and requires post-market monitoring, risk management, logging, and human oversight [73].

Three implications follow for digital biomarkers. First, software and firmware changes in consumer devices are often outside the control of the biomarker developer; a PCCP or monitoring plan should therefore anticipate device and operating-system changes, define how equivalence will be re-established, and specify what happens to scores collected during the transition. FDA guidance on digital health technologies in clinical investigations similarly asks sponsors to consider how updates during a trial will be managed [15]. Second, in clinical trials using digital endpoints, a mid-study change in device or algorithm may require bridging analyses so that pre- and post-change measurements remain comparable. Third, the Attribute stage of the proposed framework provides the type of evidence that a change control plan requires: a documented reason for the modification and an assessment of whether it addresses the identified source of drift.

5. Building Drift-Resilient Digital Biomarkers

5.1 Longitudinal Validation as a Standard Rather Than an Optional Feature

If digital biomarkers are intended to monitor neurodegenerative disease over extended periods, validation should reflect the temporal conditions under which they will actually be used. Longitudinal validation should therefore be treated as a fundamental component of digital biomarker evaluation rather than an additional step performed after model development: time itself should become a validation dimension.

Temporal holdout testing is particularly informative because it evaluates a model on observations collected at a later period than those used for development, providing a more realistic assessment of temporal generalizability than random

splitting alone [62]. Longer follow-up is also important because some forms of drift may emerge gradually; multi-year datasets can capture changes in disease stage, treatment, functional capacity, behavior, and technology that short studies are unlikely to reveal. Prospective studies provide an additional level of evidence, external validation tests whether performance is preserved in independent cohorts and clinical environments [62,64], cross-population validation tests whether it is preserved in populations that differ in plausible sources of drift, and cross-device validation tests whether it transfers across smartphones, wearables, or technological generations. Recurring local validation, in which performance is re-established at each deployment site and periodically thereafter, has been proposed as a complement to one-time external validation [74].

Longitudinal validation should also evaluate more than discrimination. Changes in calibration [25], missingness, data quality, subgroup performance, and the stability of individual predictions can provide important evidence about whether a biomarker remains clinically interpretable. A model that maintains an acceptable overall AUC while producing increasingly unstable patient-level estimates may not be suitable for longitudinal decision-making. Complementary validation and monitoring strategies, with examples of statistical and AI methods for each, are summarized in Table 3.

Table 3. Validation and monitoring strategies for drift-resilient digital biomarkers, with example statistical and AI methods.

Strategy	**Primary purpose**	**What it can reveal**	**Key limitation**	**Example statistical / AI methods**
Random train/test splitting	Estimate performance under similar data conditions	Baseline discrimination and predictive performance	May overestimate longitudinal robustness	k-fold cross-validation with subject-level (grouped) splitting to prevent leakage of repeated measurements [61]
Temporal holdout validation	Test performance on later observations	Temporal distribution shift and durability	Requires adequate longitudinal data	Train on an earlier period and test on a later one; rolling-origin evaluation; calibration slope tracked over time [24]
External validation	Assess performance in independent populations or settings	Population and clinical-context generalizability	May not capture within-person temporal drift; an external cohort may be as demographically narrow as the development cohort	Independent-cohort testing of discrimination and calibration; reporting according to TRIPOD+AI [65]
Cross-population validation	Test performance in populations that differ in plausible sources of drift	Population diversity deficit; subgroup performance gaps	Requires deliberate recruitment of diverse cohorts	Performance stratified by age, sex, ancestry, language, education, and socioeconomic position; domain classifier quantifying cohort differences; group fairness metrics
Cross-device validation	Evaluate robustness across technologies	Sensitivity to hardware, sensors, and platforms	Cannot fully reproduce all real-world technology changes	Leave-one-device-out validation; Bland-Altman agreement; device-specific calibration; domain adaptation [39]
Within-person (longitudinal reliability) validation	Assess stability of repeated measurements in the same individual	Test-retest reliability, minimal detectable change, responsiveness	Requires dense repeated data and periods of known clinical stability	Intraclass correlation; linear mixed-effects models; minimal detectable change; within-person prediction variance [26]
Prospective validation	Evaluate performance under predefined real-world conditions	Deployment reliability and longitudinal behavior	Resource- and time-intensive	Pre-registered protocol; silent (shadow) deployment; recurring local validation [74]
Post-deployment monitoring	Detect changes after implementation	Data drift, calibration changes, uncertainty, and performance deterioration	Requires ongoing surveillance and suitable reference outcomes	Population stability index; maximum mean discrepancy or domain-classifier tests; CUSUM/EWMA control charts; label-shift estimation [22,59,75]
Recalibration/ model updating	Respond to confirmed performance deterioration	Whether adaptation can restore predictive reliability	May reduce comparability across time	Logistic recalibration of intercept and slope; nonparametric updating; transfer learning; continual learning [13,63,76]
Revalidation	Confirm continued clinical validity after adaptation	Whether updated models remain reliable across relevant conditions	Requires new evaluation data	Champion-challenger comparison on recent held-out data; acceptance criteria defined in a predetermined change control plan [71]

5.2 Monitoring Drift After Deployment

Longitudinal validation before deployment cannot guarantee that a digital biomarker will remain reliable once exposed to the changing conditions of routine use. Drift monitoring should therefore continue after implementation, with attention to both changes in the data and changes in model performance [60,77]. Monitoring the distribution of incoming data can provide an early indication that the population, technology, behavior, or measurement environment has diverged from the conditions represented during development. However, detecting distributional change does not necessarily establish that clinical performance has deteriorated [58]. The most useful monitoring strategy therefore connects changes in the data with changes in calibration, discrimination, uncertainty, and clinically relevant outcomes whenever appropriate reference information becomes available [24]. A recent case study of a deployed clinical model illustrated the role of routine data drift detection within a structured model governance process [78].

Drift detection should be accompanied by uncertainty estimation and appropriate human oversight. When a model encounters data substantially different from those used during development, communicating increased uncertainty, or abstaining from a prediction, may be more responsible than producing a confident prediction. Human review can provide contextual information that automated systems cannot reliably infer, particularly when technological changes, treatment modifications, or unusual behavioral circumstances coincide with a change in model output.

5.3 Managing the Consequences of Model Updating

When performance begins to deteriorate, recalibration may be preferable to complete model redevelopment if the underlying relationship remains broadly valid but predictions have become systematically misaligned [13]. More substantial changes may require model updating or retraining [63]. Adaptation can help a model remain responsive to changing conditions, but frequent updating can also alter the model's behavior and make predictions difficult to compare across time. In longitudinal monitoring, comparability is itself clinically valuable because clinicians need to determine whether a change in a patient's score reflects genuine disease progression or merely a change in the model.

Several strategies can preserve comparability while allowing adaptation. (i) Version control: every model, preprocessing pipeline, and feature definition receives a version identifier, and every score is stored with the identifiers that produced it, so that any score can be traced and reproduced. (ii) Model-freezing periods: within a defined monitoring window, such as the duration of a clinical trial or a treatment-response assessment, the model is held fixed unless a safety threshold is crossed. (iii) Parallel ("shadow" or champion-challenger) evaluation: a candidate model runs alongside the deployed model on the same data, and its outputs are compared against reference outcomes before it replaces the deployed version. (iv) Predefined update criteria: the magnitude of degradation that triggers recalibration versus retraining, the data required, and the acceptance criteria for the updated model are specified in advance, consistent with a predetermined change control plan [71]. (v) Preservation and bridging of historical scores: historical scores are retained under their original version and, where feasible, re-scored with the updated model, so that trajectories can be displayed on a consistent scale with the version change marked. Continual-learning approaches that update models incrementally must also guard against "catastrophic forgetting" of earlier patient groups or disease stages [76].

5.4 Designing Biomarkers Around Invariance and Robustness

A complementary strategy is to reduce vulnerability to drift at the biomarker design stage. Not every digital feature that improves prediction is necessarily a robust representation of disease. Some features may depend heavily on a particular device, environment, behavioral routine, or pattern of technology use, and their predictive value may disappear when the conditions that generated them change.

This motivates the concept of biologically anchored features, referring here to digital characteristics whose clinical relevance is less dependent on a particular device or behavioral context [56]. The aim is not to assume that any digital feature is completely invariant, but to identify signals that retain useful relationships with the underlying clinical construct across reasonable changes in technology, environment, and behavior [17,45]. Methods such as invariant risk minimization seek representations whose predictive relationship with the outcome is stable across training environments [79], and causal reasoning about how data were generated can identify which shifts a model should be invariant to [11,80].

An important qualification follows from the dual diversity argument [23]. A feature that is invariant across devices and sessions within a narrow cohort has demonstrated robustness to technological and individual variation in that cohort, but not robustness to population diversity. Invariance learned across "environments" can only be as broad as the environments represented in training. Claims that a feature is biologically anchored should therefore state the populations across which invariance was tested, and invariance across diagnostically distinct groups that share symptom labels [27] should not be assumed.

The broader principle is to treat robustness as a design objective, not merely a property assessed after an algorithm has been developed. This approach complements post-deployment monitoring by reducing the sources of instability to which the model is exposed in the first place.

5.5 A Proposed Framework for Drift-Resilient Digital Biomarkers

The preceding analysis suggests that managing digital biomarker drift requires a shift from episodic model evaluation toward a structured process capable of identifying and responding to changes throughout the biomarker lifecycle. On this basis, this review proposes a Detect, Attribute, Quantify, Adapt, Validate, and Monitor (DAQAV-M) framework for drift-resilient digital biomarkers. Figure 1 provides an integrated overview of the major sources and consequences of digital biomarker drift and illustrates the proposed six-stage DAQAV-M framework for maintaining model reliability throughout the digital biomarker lifecycle [69,70]. Operational criteria for each stage are summarized in Table 4.

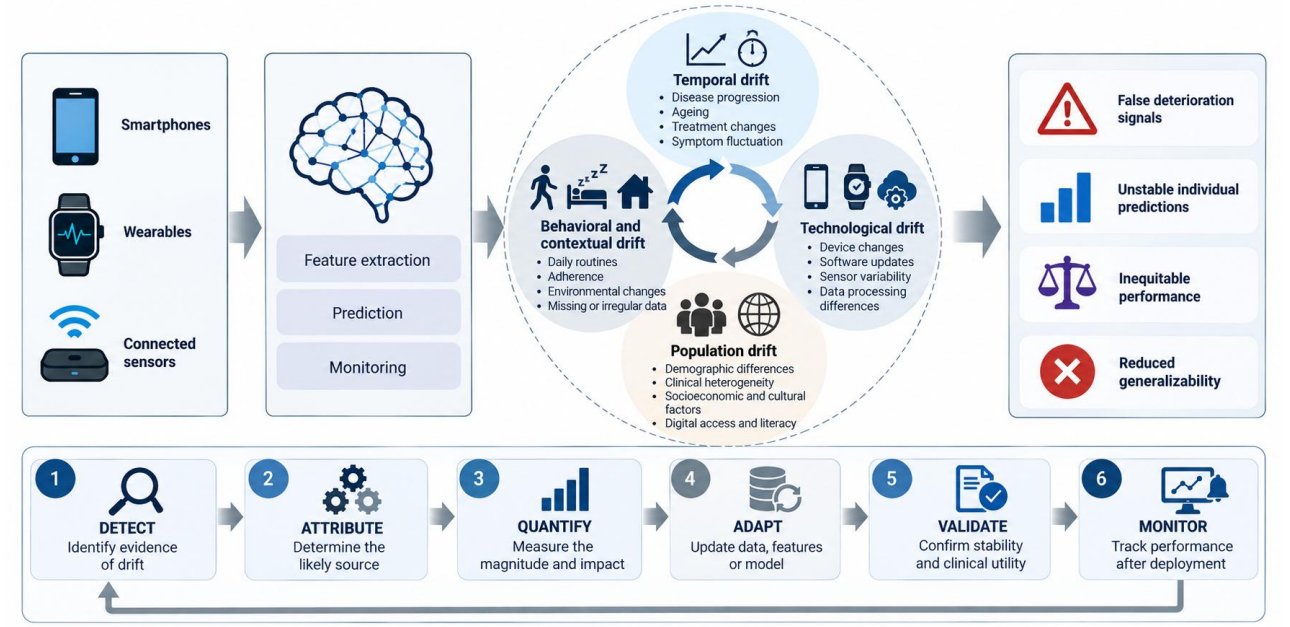


Figure 1. Sources, consequences, and management of digital biomarker drift in AI-enabled neurodegenerative disease monitoring.

Schematic representation of the interacting sources of digital biomarker drift and the proposed Detect-Attribute-Quantify-Adapt-Validate-Monitor (DAQAV-M) framework. Digital signals collected through smartphones, wearable devices, and connected sensors are processed by artificial intelligence (AI) models for feature extraction, prediction, and longitudinal monitoring. Four interacting sources of instability—temporal, technological, behavioral/contextual, and population drift—can alter the distribution, interpretation, and reliability of digital biomarkers, potentially resulting in false deterioration signals, unstable individual predictions, inequitable model performance, and reduced generalizability. The DAQAV-M framework provides a structured, iterative approach to detecting distributional changes, attributing their likely causes, quantifying their magnitude and clinical significance, implementing appropriate model adaptations, validating subsequent performance, and continuously monitoring reliability after deployment. The feedback loop emphasizes the need for ongoing evaluation and controlled adaptation to maintain clinical validity and robustness across changing biological, technological, behavioral, and population conditions.

Table 4. Operational criteria for the DAQAV-M framework. Thresholds are illustrative and should be pre-specified for each intended use.

Stage	Key question	Example methods and statistics	Illustrative trigger to proceed
Detect	Has anything changed relative to development and the last validation?	Population stability index; Kolmogorov-Smirnov or chi-squared tests with false-discovery-rate control; maximum mean discrepancy or domain-classifier tests [75]; label-shift estimation [22]; CUSUM/EWMA control charts on features, predictions, and missingness [59]	For example, population stability index > 0.2 for a key feature; domain-classifier AUC > 0.6; a control-chart signal sustained over a pre-specified window; a change in missing-data rate > 10 percentage points
Attribute	Why has it changed: biological, technological, behavioral, population, missingness, or mixed?	Alignment of change points with device and software metadata; interrupted time series around updates; mixed-effects models with device/version terms; negative-control features; difference-in-differences; comparison of deployment and development covariates; targeted clinical reassessment	A documented attribution (single or mixed source) with the supporting evidence recorded
Quantify	Does the change matter clinically, and which validity layer is affected?	Change in AUC against a non-inferiority margin; calibration-in-the-large and calibration slope [25]; expected calibration error; within-person variability compared with the minimal clinically important difference; Bland-Altman agreement between devices; subgroup metrics	For example, AUC decrease > 0.05; calibration slope outside 0.8-1.2; within-person variability during clinical stability exceeding the minimal clinically important difference; subgroup performance gap beyond a pre-specified tolerance
Adapt	What is the least disruptive adequate response?	No model change plus clinical review (biological source); harmonization or recalibration (technological); targeted data collection and retraining (population); explicit missingness modeling; versioned retraining	Only when a Quantify threshold is met; method matched to the attributed source; conducted under a change control plan [71]
Validate	Has adaptation restored validity without breaking comparability?	Temporal, cross-device, and cross-population revalidation; champion-challenger comparison; agreement between old and new versions on historical data	Acceptance criteria met on recent held-out data in all pre-specified subgroups; bridging between versions documented
Monitor	Is the system remaining stable?	Ongoing control charts; recurring local validation [74]; audit logs of versions and scores	Return to Detect on any signal; scheduled review at a pre-specified interval (for example, quarterly)

Detect. The first stage involves identifying evidence that the data or model behavior has changed from the conditions established during development and previous validation. Practical approaches include univariate tests of feature distributions (Kolmogorov-Smirnov or chi-squared tests with correction for multiple testing, or the population stability index), multivariate two-sample tests such as maximum mean discrepancy or a "domain classifier" trained to distinguish development from deployment data [75], methods for detecting label shift from model outputs without new labels [22], novelty-detection methods for population drift [53], and sequential monitoring with statistical process control charts (such as cumulative sum or exponentially weighted moving average charts) applied to feature means, prediction distributions, missingness rates, and, when labels are available, calibration and discrimination [59,77]. Detection should function as an early-warning process rather than an assumption that every statistical change represents clinical deterioration.

Attribute. The second stage determines the likely source of the observed change. This requires distinguishing biological progression from technological, behavioral, contextual, population, or mixed causes. Attribution is particularly important because the appropriate response depends on the source: a change associated with genuine disease progression should not be "corrected away" by recalibration, whereas a change caused by a new device or software update usually should be. Because the attribution problem is the least developed aspect of current practice, we outline concrete strategies. (a) Metadata collection: device model, operating system, firmware and application versions, wear location, setting, and who operated the device are recorded with every observation, so that change points can be aligned with metadata events. (b) Device-transition periods: when devices are replaced, an overlap period in which both devices are worn, or a bridging substudy, allows device effects to be estimated directly. (c) Within-person controls: an abrupt change that coincides with a software or device event and is shared across many participants is more consistent with technological drift, whereas a gradual change confined to individuals with clinical progression is more consistent with biology; interrupted time-series analysis around update dates formalizes this comparison. (d) Negative-control features: features that are expected to be insensitive to disease but sensitive to measurement, such as sensor noise floors

or sampling irregularity, help distinguish technical from biological change. (e) Mixed-effects models with person-level random effects and device or software version as fixed or random effects separate within-person trajectories from measurement effects. (f) Causal inference designs: explicit causal diagrams of the data-generating process [80], difference-in-differences comparisons between participants who did and did not experience a device change, and instrumental use of exogenous events such as scheduled updates. (g) Clinical anchoring: targeted clinical reassessment of a sample of participants whose scores changed. (h) Population comparison: domain classifiers or density-ratio estimates comparing deployment and development covariates indicate whether the deployment cohort differs systematically from the development cohort, which defines the population-shift branch. (i) Sensitivity analyses: conclusions are tested under alternative assumptions about missingness and measurement error. Where multiple sources occur simultaneously, the framework recognizes that drift may be mixed rather than forcing it into a single category.

Quantify. The third stage assesses how substantially the identified change affects biomarker reliability and which validity layer (Section 2.5) is affected. This involves examining whether drift has altered discrimination, calibration, prediction stability, subgroup performance, or the clinical interpretation of the biomarker. Useful quantities include the change in AUC relative to a pre-specified non-inferiority margin; calibration-in-the-large and calibration slope [25]; expected calibration error; within-person score variability over periods without clinical change, compared with the minimal clinically important difference; agreement between scores from old and new devices; and performance in pre-specified subgroups. Thresholds should be defined in advance for the intended use rather than adopted generically; the values in Table 4 are illustrative. Quantification provides a basis for determining whether the observed change is sufficiently consequential to require intervention rather than responding to every detectable deviation.

Adapt. The fourth stage concerns the controlled modification of the system when intervention is justified. Depending on the source and magnitude of drift, adaptation may involve recalibration, modification of preprocessing procedures, device-specific harmonization, model updating, or retraining. However, adaptation should not be automatic. Excessive updating may make longitudinal predictions difficult to compare and may introduce new forms of instability, so adaptation should follow the version-control and comparability rules in Section 5.3. When the attributed source is biological progression, the appropriate action may be no model change at all, but rather clinical review; when it is population shift, the appropriate action may be targeted data collection in the under-represented population before any model change.

Validate. The fifth stage determines whether adaptation has actually restored clinically meaningful performance. An updated model should not be assumed to be improved simply because its statistical performance increases on newly available data. Validation should assess whether the revised system remains reliable across relevant populations, technologies, time periods, and clinical contexts [64,65], and whether it preserves comparability with historical scores.

Monitor. The final stage establishes ongoing surveillance after validation and deployment. Monitoring should track changes in data characteristics, model performance, uncertainty, missingness, and relevant clinical outcomes where available. Importantly, the framework is cyclical rather than linear: new evidence of drift returns the system to detection and attribution.

Relationship to existing frameworks and novelty. The DAQAV-M framework builds on, rather than replaces, existing work. Dataset-shift detection methods [59,75] supply tools for the Detect stage; calibration-drift and model-updating methods [13,24,81] and continual-learning approaches [76] supply tools for the Adapt stage; quality-improvement and monitoring frameworks [77] and recurring local validation [74] address the Monitor stage; V3 and V3+ [29,30] define the validity layers; reporting guidance [65] and consensus guidelines [68] define what should be documented; and regulatory lifecycle approaches [71,72] govern how changes are authorized. What these approaches generally do not provide, and what DAQAV-M adds, is the following: (1) an explicit Attribute stage placed before adaptation, which is essential for digital biomarkers because a model that is "corrected" for biological progression has stopped measuring the disease; (2) a source taxonomy specific to digital measurement that includes technological, behavioral, population, and missingness drift and maps them onto machine-learning shift types; (3) a requirement to identify which validity layer is affected, so that restoring model performance is not mistaken for restoring construct validity; and (4) longitudinal comparability as an explicit constraint on adaptation. Table 5 summarizes this comparison. The novelty is therefore integrative: the framework organizes established tools around the attribution problem that is distinctive to longitudinal digital measurement.

Table 5. Relationship of the DAQAV-M framework to existing approaches.

Existing approach	Main focus	Contribution to DAQAV-M	What it typically does not address
Dataset-shift detection [59,75]	Statistical detection of distribution change	Methods for the Detect stage	The cause of a shift and its clinical significance
Calibration-drift monitoring and model updating [13,24,81]	Maintaining predictive performance	Tools for the Quantify and Adapt stages	Whether change reflects biological progression that should not be corrected
Continual learning and invariant or domain-adaptive learning [76,79]	Adapting models to new data or environments	Methods for the Adapt and design stages	Longitudinal comparability of scores; which validity layer is affected
AI quality improvement and recurring local validation [74,77]	Governance of deployed models	Structure for the Monitor and Validate stages	A source taxonomy specific to digital measurement
V3 and V3+ frameworks [29,30]	Fit-for-purpose validity of sensor-based digital health technologies	Definitions of analytical and clinical validity layers	Management of drift after deployment
TRIPOD+AI and FUTURE-AI [65,68]	Reporting and trustworthiness	Documentation and subgroup reporting standards	Operational procedures for attributing drift
Predetermined change control plans and lifecycle guidance [71,72]	Regulatory authorization of change	Governance of the Adapt and Validate stages	A scientific method for identifying drift sources
DAQAV-M (this review)	Lifecycle management of digital biomarker drift	Integrates the above around an explicit Attribute stage, a digital-measurement source taxonomy, validity-layer assessment, and comparability constraints	Not yet evaluated prospectively; thresholds require calibration to each intended use

5.6 Published Examples of Drift-Resilient Approaches

Although few digital biomarker studies in neurodegenerative disease have implemented a complete lifecycle approach, published examples exist for most of its components. Temporal validation and calibration-drift analysis have been demonstrated in clinical prediction, where calibration of several model types deteriorated over years of deployment [24] and where the robustness of features to non-stationarity varied markedly [61]; a systematic review catalogued approaches for preserving performance under temporal dataset shift [81]. Recalibration and updating have been implemented with nonparametric updating methods [13] and with transfer learning to correct temporal performance drift [63]. Cross-device validation and mitigation were examined systematically for smartphone and smartwatch motion sensors [39], and multi-condition validation of wearable-derived walking speed was undertaken by the Mobilise-D consortium [42]. Domain adaptation and invariant representation learning have been proposed to handle device and setting differences [79]. Federated learning, which trains models across sites without pooling data, can widen the population represented in training [82], and has been explored for Parkinson disease data with heterogeneous (non-independent and identically distributed) distributions across sites [83]. Foundation models pretrained on large and varied wearable datasets offer representations that may transfer across tasks and devices [44]. Drift detection has been evaluated empirically in real-world medical imaging data [58], in simulated post-market surveillance [59], and for population drift in deployed diagnostics [53], and a governance case study showed how drift detection supported model maintenance in practice [78]. Uncertainty-aware and selective prediction, in which the model abstains or flags reduced confidence when inputs are unfamiliar, is a further approach consistent with the Monitor stage. The gap in neurodegenerative disease is less a lack of methods than a lack of studies that combine them prospectively over clinically relevant time scales.

6. Research Priorities and the Future of Digital Biomarker Validation

6.1 What Should Future Studies Measure?

Future studies of digital biomarkers should provide greater detail about the conditions under which measurements are collected and how those conditions change over time. Reporting the duration and completeness of follow-up is particularly important because short observation periods may provide limited evidence about longitudinal stability. Studies should also document changes in smartphones, wearable devices, sensors, software platforms, operating systems, and other components of the measurement system [29,30]. Without this information, it can be difficult to determine whether changes in biomarker performance reflect biological evolution or changes in the technology generating the data.

Behavioral and clinical context should receive similar attention. Changes in physical activity, sleep, medication, daily routines, adherence, and living circumstances can influence digital measurements and should not be treated as incidental

sources of variation. Reporting patterns of missingness, by time, disease stage, and subgroup, together with the assumed missing-data mechanism, is also essential.

Future validation should additionally report population diversity and performance across relevant demographic, socioeconomic, linguistic, clinical, technological, and temporal subgroups [4,65]. Development and validation cohorts should be described with sufficient detail to judge how far they depart from the populations in which the biomarker will be used, including the dimensions summarized in Section 3.4. Temporal performance should be evaluated explicitly, including whether discrimination, calibration, prediction stability, and data quality remain acceptable across the intended monitoring period, and both within-person and between-person performance should be reported.

6.2 From Static Validation to Lifecycle Validation

Validation of digital biomarkers should move beyond the assumption that acceptable development and external-test performance establishes permanent reliability. Because patients, technologies, behaviors, and clinical environments change, validation is better understood as a lifecycle process spanning development, deployment, monitoring, recalibration, and revalidation [69,70,72]. Validation should therefore be regarded as an ongoing claim that must be supported by evidence over the biomarker's operational lifetime, rather than a permanent designation achieved at a single point in time.

6.3 Priority Questions for the Field

Several unresolved questions should shape the next generation of digital biomarker research. First, which digital features retain a stable relationship with disease across different stages of neurodegeneration, and which are highly sensitive to changes in behavior or technology? Second, what degree of performance degradation or distributional change should trigger recalibration, and what evidence should be required before an updated model is returned to clinical use? Third, when does incremental model updating become inappropriate, requiring replacement with a substantially redesigned system? Fourth, which features remain valid across populations that differ in age, ancestry, language, education, and access to technology, and how large and diverse must validation cohorts be to support such claims?

A further challenge is determining how biological progression can be separated from behavioral adaptation, particularly when behavior is itself influenced by disease. Future studies must also establish how much longitudinal evidence is necessary to demonstrate durability across devices, populations, and clinical environments. Finally, regulators and health systems will need clearer criteria for determining when evidence of longitudinal stability is sufficient for clinical adoption. Addressing these questions will require closer integration of AI methodology, clinical science, digital measurement, and implementation research.

7. Conclusion

What current evidence establishes. Digital measures of gait, motor function, speech, cognition, and activity can detect and track change in Parkinson disease, Alzheimer disease, Huntington disease, amyotrophic lateral sclerosis, and related conditions, including change that episodic clinical assessment may miss. It is also established that the data on which these measures depend vary with device, setting, disease stage, and engagement, and that clinical prediction models lose calibration and generalizability under temporal and population shift.

What remains uncertain. Few studies have followed digital biomarkers in neurodegenerative disease over multiple years while devices, software, and populations changed, so the magnitude of each drift source in practice is largely unknown. It is unclear which features are robust across populations as well as across individuals, how best to attribute observed change to its source, what thresholds should trigger adaptation, and how adaptive models can be updated without compromising longitudinal comparability or regulatory accountability.

What should change in future studies. Studies should (1) use temporal, cross-device, and cross-population validation in addition to random splitting and conventional external validation; (2) report within-person as well as between-person performance; (3) log device, software, and context metadata with every observation and plan device-transition bridging; (4) report and model missingness by stage and subgroup; (5) specify the reference standard and its own stability; (6) distinguish analytical, clinical, and construct validity from model performance; and (7) pre-specify monitoring, attribution, and update rules, such as those in the proposed DAQAV-M framework (Figure 1, Table 4), with version control that preserves the interpretability of historical scores. Digital biomarkers that meet these standards will be more than accurate models: they will be measurement systems whose meaning can be trusted as patients, technologies, and populations change.

Acknowledgements

The authors have no additional acknowledgements to declare.

Funding

No specific funding was received for this work.

Ethical Statement

Ethical approval was not required because this study is a critical review of previously published literature and did not involve human participants, animals, or identifiable patient data.

Informed Consent

Not applicable.

Data Availability Statement

No new datasets were generated or analyzed for this critical review. All information discussed was derived from the published literature cited in the manuscript.

Author Contributions

Olaitan Ebenezer Oluwadare (O.E.O.): Conceptualization, Methodology, Investigation (literature search and screening), Formal analysis (evidence synthesis), Visualization, Writing - original draft, Writing - review & editing, Project administration. Anyebe Daniel Ameh (A.D.A.): Investigation (literature search and screening), Formal analysis (evidence synthesis), Validation, Writing - review & editing. Abdul-Basid Yakubu (A.B.Y.): Investigation, Formal analysis (critical evaluation of the evidence), Writing - review & editing. Faith Uzochi Sunday (F.U.S.): Investigation, Validation (interpretation of the evidence), Writing - review & editing. All authors contributed to the conception and development of the work, critically reviewed and revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

Conflict of Interest

The authors declare that they have no competing interests.

Generative AI Statement

The authors acknowledge using artificial intelligence (AI)-based tools solely for language-level editing and polishing of the manuscript, including improvements to grammar, spelling, sentence structure, clarity, and readability. AI tools were not used to generate manuscript content, develop scientific arguments, analyse or interpret data, or create figures and images. The authors are responsible for all intellectual contributions, scientific content, interpretations, and conclusions, and they reviewed and approved the final manuscript.

References

[1] Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. The first frontier: Digital biomarkers for neurodegenerative disorders. Digital Biomarkers, 2017, 1(1), 6-13. DOI: 10.1159/000477383

[2] Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. npj Digital Medicine, 2019, 2, 14. DOI: 10.1038/s41746-019-0090-4

[3] Vasudevan S, Saha A, Tarver ME, Patel B. Digital biomarkers: Convergence of digital health technologies and biomarkers. npj Digital Medicine, 2022, 5, 36. DOI: 10.1038/s41746-022-00583-z

[4] Macias Alonso AK, Hirt J, Woelfle T, Janiaud P, Hemkens LG. Definitions of digital biomarkers: A systematic mapping of the biomedical literature. BMJ Health & Care Informatics, 2024, 31(1), e100914. DOI: 10.1136/bmjhci-2023-100914

[5] Espay AJ, Bonato P, Nahab FB, Maetzler W, Dean JM, Klucken J, et al. Technology in Parkinson's disease: Challenges and opportunities. Movement Disorders, 2016, 31(9), 1272-1282. DOI: 10.1002/mds.26642

[6] Hansen C, Sanchez-Ferro A, Maetzler W. How mobile health technology and electronic health records will change care of patients with Parkinson's disease. Journal of Parkinson's Disease, 2018, 8(s1), S41-S45. DOI: 10.3233/JPD-181498

[7] Andrzejewski KL, Dowling AV, Stamler D, Felong TJ, Harris DA, Wong C, et al. Wearable sensors in Huntington disease: A pilot study. Journal of Huntington's Disease, 2016, 5(2), 199-206. DOI: 10.3233/JHD-160197

[8] Chudzik A, Śledzianowski A, Przybyszewski AW. Machine learning and digital biomarkers can detect early stages of neurodegenerative diseases. Sensors (Basel), 2024, 24(5), 1572. DOI: 10.3390/s24051572

[9] Song J, Cho E, Lee H, Lee S, Kim S, Kim J. Development of neurodegenerative disease diagnosis and monitoring from traditional to digital biomarkers. Biosensors (Basel), 2025, 15(2), 102. DOI: 10.3390/bios15020102

[10] Li H, Zecca M, Huang J. Evaluating the utility of wearable sensors for the early diagnosis of Parkinson disease: Systematic review. Journal of Medical Internet Research, 2025, 27, e69422. DOI: 10.2196/69422

[11] Subbaswamy A, Saria S. From development to deployment: Dataset shift, causality, and shift-stable models in health AI. Biostatistics, 2020, 21(2), 345-352. DOI: 10.1093/biostatistics/kxz041
[12] Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. The New England Journal of Medicine, 2021, 385(3), 283-286. DOI: 10.1056/NEJMc2104626
[13] Davis SE, Greevy RA Jr, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. Journal of the American Medical Informatics Association, 2019, 26(12), 1448-1457. DOI: 10.1093/jamia/ocz127
[14] Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. Science, 2024, 383(6679), 164-167. DOI: 10.1126/science.adg8538
[15] U.S. Food and Drug Administration. Digital health technologies for remote data acquisition in clinical investigations: Guidance for industry, investigators, and other stakeholders. Silver Spring (MD): FDA. 2023. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations
[16] Warmerdam E, Hausdorff JM, Atrsaei A, Zhou Y, Mirelman A, Aminian K, et al. Long-term unsupervised mobility assessment in movement disorders. The Lancet Neurology, 2020, 19(5), 462-470. DOI: 10.1016/S1474-4422(19)30397-7
[17] Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. npj Digital Medicine, 2020, 3, 18. DOI: 10.1038/s41746-020-0226-6
[18] Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology, 2016, 41(7), 1691-1696. DOI: 10.1038/npp.2016.7
[19] Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognition, 2012, 45(1), 521-530. DOI: 10.1016/j.patcog.2011.06.019
[20] Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. Dataset shift in machine learning. Cambridge (MA): MIT Press, 2009. DOI: 10.7551/mitpress/9780262170055.001.0001
[21] Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Computing Surveys, 2014, 46(4), 44. DOI: 10.1145/2523813
[22] Lipton ZC, Wang YX, Smola AJ. Detecting and correcting for label shift with black box predictors. Proceedings of the 35th International Conference on Machine Learning, 2018, 80, 3122-3130. Available from: https://proceedings.mlr.press/v80/lipton18a.html
[23] Rudroff T. The dual diversity crisis in physiology. The Journal of Physiology, 2026, 604(12), 4626-4630. DOI: 10.1113/JP290947
[24] Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. Journal of the American Medical Informatics Association, 2017, 24(6), 1052-1061. DOI: 10.1093/jamia/ocx030
[25] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: The Achilles heel of predictive analytics. BMC Medicine, 2019, 17(1), 230. DOI: 10.1186/s12916-019-1466-7
[26] Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(27), E6106-E6115. DOI: 10.1073/pnas.1711978115
[27] Rudroff T. Frontal-striatal glucose metabolism and fatigue in patients with multiple sclerosis, long COVID, and COVID-19 recovered controls. Experimental Brain Research, 2024, 242(9), 2125-2136. DOI: 10.1007/s00221-024-06882-z
[28] Rudroff T. Digital biomarkers and AI for remote monitoring of fatigue progression in neurological disorders: Bridging mechanisms to clinical applications. Brain Sciences, 2025, 15(5), 533. DOI: 10.3390/brainsci15050533
[29] Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): The foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). npj Digital Medicine, 2020, 3, 55. DOI: 10.1038/s41746-020-0260-4
[30] Bakker JP, Barge R, Centra J, Cobb B, Cota C, Guo CC, et al. V3+ extends the V3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. npj Digital Medicine, 2025, 8(1), 51. DOI: 10.1038/s41746-024-01322-2
[31] Evers LJW, Krijthe JH, Meinders MJ, Bloem BR, Heskes TM. Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS. Movement Disorders, 2019, 34(10), 1480-1487. DOI: 10.1002/mds.27790
[32] Adams JL, Kangarloo T, Gong Y, Khachadourian V, Tracey B, Volfson D, et al. Using a smartwatch and smartphone to assess early Parkinson's disease in the WATCH-PD study over 12 months. Journal of Parkinson's Disease, 2024, 10(1), 112. DOI: 10.1038/s41531-024-00721-2
[33] Sotirakis C, Su Z, Brzezicki MA, Conway N, Tarassenko L, FitzGerald JJ, et al. Identification of motor progression in Parkinson's disease using wearable sensors and machine learning. Journal of Parkinson's Disease, 2023, 9, 142. DOI: 10.1038/s41531-023-00581-2
[34] Templeton JM, Poellabauer C, Schneider S, Rahimi M, Braimoh T, Tadamarry F, et al. Modernizing the staging of Parkinson disease using digital health technology. Journal of Medical Internet Research, 2025, 27, e63105. DOI: 10.2196/63105
[35] Schalkamp AK, Peall KJ, Harrison NA, Sandor C. Wearable movement-tracking data identify Parkinson's disease years before clinical diagnosis. Nature Medicine, 2023, 29(8), 2048-2056. DOI: 10.1038/s41591-023-02440-2
[36] Stegmann GM, Hahn S, Liss J, Shefner J, Rutkove S, Shelton K, et al. Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis. npj Digital Medicine, 2020, 3, 132. DOI: 10.1038/s41746-020-00335-x
[37] Berron D, Glanz W, Clark L, Basche K, Grande X, Güsten J, et al. A remote digital memory composite to detect cognitive impairment in memory clinic samples in unsupervised settings using mobile devices. npj Digital Medicine, 2024, 7(1), 79. DOI: 10.1038/s41746-024-00999-9
[38] van den Berg RL, van der Landen SM, Keijzer MJ, van Gils AM, van Dam M, Ziesemer KA, et al. Smartphone- and tablet-based tools to assess cognition in individuals with preclinical Alzheimer disease and mild cognitive impairment: Scoping review. Journal of Medical Internet Research, 2025, 27, e65297. DOI: 10.2196/65297
[39] Stisen A, Blunck H, Bhattacharya S, Prentow TS, Kjærgaard MB, Dey A, et al. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys '15). New York: ACM, 2015, 127-140. DOI: 10.1145/2809695.2809718

[40] Bianchini E, Rinaldi D, De Carolis L, Galli S, Alborghetti M, Hansen C, et al. Reliability of average daily steps measured through a consumer smartwatch in Parkinson disease phenotypes, stages, and severities: Cross-sectional study. JMIR Formative Research, 2025, 9, e63153. DOI: 10.2196/63153
[41] Rookes TA, Batla A, Armstrong M, Ambler G, Walters K, Schrag A. Patient acceptability and technical reliability of wearable devices used for monitoring people with Parkinson disease: Survey study. JMIR Formative Research, 2025, 9, e63704. DOI: 10.2196/63704
[42] Kirk C, Küderle A, Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, Ullrich M, et al. Mobilise-D insights to estimate real-world walking speed in multiple conditions with a wearable device. Scientific Reports, 2024, 14(1), 1754. DOI: 10.1038/s41598-024-51766-5
[43] Lozano-Garcia M, Doheny EP, Mann E, Morgan-Jones P, Drew C, Busse-Morris M, et al. Estimation of gait parameters in Huntington's disease using wearable sensors in the clinic and free-living conditions. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2024, 32, 2239-2349. DOI: 10.1109/TNSRE.2024.3407887
[44] Abbaspourazad S, Elachqar O, Miller AC, Emrani S, Nallasamy U, Shapiro I. Large-scale training of foundation models for wearable biosignals. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), 2024. Available from: https://arxiv.org/abs/2312.05409
[45] Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 2009, 56(4), 1015-1022. DOI: 10.1109/TBME.2008.2005954
[46] Clark A, Gresham G, Pevnick J, Duncan R, Kamrava M, Sobolev M. Predicting engagement patterns with connected wearable devices in a health system: Survival analysis. Journal of Medical Internet Research, 2025, 27, e78507. DOI: 10.2196/78507
[47] Ding H, Ho K, Searls E, Low S, Li Z, Rahman S, et al. Assessment of wearable device adherence for monitoring physical activity in older adults: Pilot cohort study. JMIR Aging, 2024, 7, e60209. DOI: 10.2196/60209
[48] Peterson CM, St Louis RM, Flannagan C. Enhancing enrollment and adherence in long-term wearable research on dementia: Qualitative systematic review and meta-synthesis. JMIR Aging, 2025, 8, e63768. DOI: 10.2196/63768
[49] Nunes AS, Pawlik M, Mishra RK, Waddell E, Coffey M, Tarolli CG, et al. Digital assessment of speech in Huntington disease. Frontiers in Neurology, 2024, 15, 1310548. DOI: 10.3389/fneur.2024.1310548
[50] Yang R, Gao S, Jiang Y. Digital divide as a determinant of health in the U.S. older adults: Prevalence, trends, and risk factors. BMC Geriatrics, 2024, 24(1), 1027. DOI: 10.1186/s12877-024-05612-y
[51] Jaana M, Tamim H, Paré G. eHealth and the digital divide among older Canadians: Insights from a national cross-sectional study. Journal of Medical Internet Research, 2025, 27, e72274. DOI: 10.2196/72274
[52] Li Y, Liu C, Sun J, Zhang J, Li X, Zhang Z. The digital divide and cognitive disparities among older adults: Community-based cohort study in China. Journal of Medical Internet Research, 2024, 26, e59684. DOI: 10.2196/59684
[53] Jones WS, Farrow DJ. One-class support vector machines for detecting population drift in deployed machine learning medical diagnostics. Scientific Reports, 2025, 15, 12157. DOI: 10.1038/s41598-025-94427-x
[54] Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. The New England Journal of Medicine, 2012, 367, 1355-1360. DOI: 10.1056/NEJMsr1203730
[55] Burq M, Rainaldi E, Ho KC, Chen C, Bloem BR, Evers LJW, et al. Virtual exam for Parkinson's disease enables frequent and reliable remote measurements of motor function. npj Digital Medicine, 2022, 5(1), 65. DOI: 10.1038/s41746-022-00607-8
[56] Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease, 2016, 49(2), 407-422. DOI: 10.3233/JAD-150520
[57] Lipsmeier F, Simillion C, Bamdadian A, et al. A remote digital monitoring platform to assess cognitive and motor symptoms in Huntington disease: Cross-sectional validation study. Journal of Medical Internet Research, 2022, 24(6), e32997. DOI: 10.2196/32997
[58] Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. Nature Communications, 2024, 15(1), 1887. DOI: 10.1038/s41467-024-46142-w
[59] Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. npj Digital Medicine, 2024, 7, 120. DOI: 10.1038/s41746-024-01085-w
[60] Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: Implications and potential remedies. The British Journal of Radiology, 2023, 96(1150), 20220878. DOI: 10.1259/bjr.20220878
[61] Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. Proceedings of Machine Learning Research, 2019, 106, 381-405. Available from: https://proceedings.mlr.press/v106/nestor19a.html
[62] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: What, why, how, when and where? Clinical Kidney Journal, 14(1), 49-58. DOI: 10.1093/ckj/sfaa188
[63] Zhang X, Xue Y, Su X, Chen S, Liu K, Chen W, et al. A transfer learning approach to correct the temporal performance drift of clinical prediction models: Retrospective cohort study. JMIR Medical Informatics, 2022, 10(11), e38053. DOI: 10.2196/38053
[64] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. Annals of Internal Medicine, 2015, 162, W1-W73. DOI: 10.7326/M14-0698
[65] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. The BMJ, 2024, 385, e078378. DOI: 10.1136/bmj-2023-078378
[66] Nouri S, Khoong EC, Lyles CR, Karliner L. Addressing equity in telemedicine for chronic disease management during the Covid-19 pandemic. NEJM Catalyst Innovations in Care Delivery, 2020, 1(3), 1-13. DOI: 10.1056/CAT.20.0123
[67] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 2019, 366(6464), 447-453. DOI: 10.1126/science.aax2342
[68] Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. The BMJ, 2025, 388, e081554. DOI: 10.1136/bmj-2024-081554
[69] U.S. Food and Drug Administration. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Silver Spring (MD): FDA, 2021. Available from: https://www.fda.gov/media/145022/download

[70] U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Predetermined change control plans for machine learning-enabled medical devices: Guiding principles. 2023. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles
[71] U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and Food and Drug Administration staff. Silver Spring (MD): FDA, 2024. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence
[72] U.S. Food and Drug Administration. Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations. Draft guidance for industry and Food and Drug Administration staff. Silver Spring (MD): FDA, 2025. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing
[73] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024. Available from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[74] Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. Nature Medicine, 2023, 29(11), 2686-2687. DOI: 10.1038/s41591-023-02540-z
[75] Rabanser S, Günnemann S, Lipton ZC. Failing loudly: an empirical study of methods for detecting dataset shift. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019. Available from: https://arxiv.org/abs/1810.11953
[76] Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health, 2020, 2(6), e279-e281. DOI: 10.1016/S2589-7500(20)30102-3
[77] Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. npj Digital Medicine, 2022, 5(1), 66. DOI: 10.1038/s41746-022-00611-y
[78] Verhoef C, Gommers D, van de Sande D, van Genderen ME, Hilling DE, Daams F, et al. Importance of model governance in clinical AI models: Case study on the relevance of data drift detection. BMJ Digital Health & AI, 2025, 1(1), e000046. DOI: 10.1136/bmjdhai-2025-000046
[79] Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. arXiv, 2019. Available from: https://arxiv.org/abs/1907.02893
[80] Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nature Communications, 2020, 11, 3673. DOI: 10.1038/s41467-020-17478-w
[81] Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. Applied Clinical Informatics, 2021, 12(4), 808-815. DOI: 10.1055/s-0041-1735184
[82] Austin JA, Lobo EH, Samadbeik M, Engstrom T, Philip R, Pole JD, et al. Decades in the making: The evolution of digital health research infrastructure through synthetic data, common data models, and federated learning. Journal of Medical Internet Research, 2024, 26, e58637. DOI: 10.2196/58637
[83] Tanim SA, Aurnob AR, Mridha MF, Safran M, Alfarhood S, Che D. Secure federated learning for Parkinson's disease: Non-IID data partitioning and homomorphic encryption strategies. IEEE Access, 2024, 12, 127309-127327. DOI: 10.1109/ACCESS.2024.3454690