Paper AS10

Ensuring Reliability of Medical Devices through Statistical Oversight in the Age of AI

Author: Shrutimita Pokhariyal, Principal Statistician, Symbio, Chandigarh, India
Co Author: Parveen Kumar, Director Statistics, Geninvo, Chandigarh, India

ABSTRACT

As medical devices like smart injectables and wearables become more data-driven, ensuring their reliability is critical-especially in clinical and real-world settings. While Artificial Intelligence (AI) and Machine Learning (ML) offer predictive capabilities to detect device malfunctions or failures early, they must be supported by rigorous statistical frameworks to ensure accuracy and regulatory compliance. This paper highlights the central role of biostatistics in monitoring and maintaining device performance. Techniques such as Survival Analysis (SA) help estimate time-to-failure, while control charts and Statistical Process Control (SPC) methods detect shifts or trends in device metrics like injection force or temperature. Automated data-quality monitoring to prevent missingness and reduce noise (e.g., alerts for late data, out-of-range values, and incomplete records) can further improve the integrity of downstream analyses. These approaches provide early warnings of malfunction and support predictive maintenance models. Aligned with ICH E6(R3), E9(R1), and United States Food and Drug Administration (FDA) guidance on digital health technologies, these statistical tools enhance transparency, traceability, and patient safety.

INTRODUCTION

Today's medical devices-from smart injectables to wearable monitors-generate continuous, high-frequency data in both clinical trials and routine care. Embedded sensors (e.g., injection force, pressure, temperature) and wireless connectivity improve convenience and outcomes, but they also heighten the need to demonstrate sustained reliability and safety over time. Here, reliability means the device performs its intended function consistently, without unexpected failures. AI/ML can help detect early signs of malfunction or drift, but to be clinically credible and regulator-ready these alerts must be underpinned by prespecified statistical methods that quantify accuracy, timeliness, and actionability. In other words, AI/ML are most effective when integrated into rigorous statistical oversight rather than operating in isolation.

Ensuring device reliability is therefore a cross-functional responsibility spanning biomedical engineers, clinicians, data scientists, and biostatisticians. In modern clinical development, investigational products increasingly rely on connected devices for drug delivery, physiologic monitoring, or endpoint capture; any device malfunction or data-quality degradation can compromise trial integrity, bias treatment-effect estimation, and create patient-safety risk. Accordingly, clinical and statistical teams should embed device evaluation, data-quality controls, and performance monitoring into protocol design, SAPs, and risk-based monitoring plans. This paper describes how established statistical tools-survival analysis, control charts, and statistical process control-can be adapted to device telemetry and paired with AI/ML to form a practical early-warning system, aligned with regulatory expectations for transparency, traceability, and documented governance.

A clear example of prespecified AI/ML governance-highlighting the device-signal vs false-positive screening tradeoff-comes from Pfizer's Acute Respiratory Illness Surveillance (AcRIS) study (NCT04748445). AcRIS is a low-interventional, decentralized "bring-your-own-device" digital study that uses a mobile app to capture daily voice recordings and symptom reports, then develops and validates ML models to classify participants as "well" versus "sick," with additional analyses to characterize false-positive classifications among RT-PCR–negative participants.

Regulatory developments further reinforce this lifecycle, statistically grounded approach. FDA's AI/ML program for device software emphasizes total product lifecycle risk management, including controlled model evolution through Predetermined Change Control Plans (PCCPs) and clearer expectations for documenting development, evaluation, monitoring, and modification of AI-enabled device software functions. European Medicine Agency (EMA's) reflection paper on AI across the medicinal product lifecycle similarly emphasizes human-centric, transparent, context-of-use–driven deployment supported by robust data governance and ongoing performance oversight. Most recently, FDA and EMA's joint Guiding Principles of Good AI Practice in Drug Development (January 2026) signal converging expectations that AI used for evidence generation should be fit-for-purpose, risk-based, and supported by well-

characterized evaluation, monitoring, and accountable human oversight-principles that align directly with prespecified statistical methods for reliability, data quality, and decision traceability.

THE RISE OF DATA-DRIVEN MEDICAL DEVICES AND AI IN MONITORING

Data-Driven Devices:

Smart injectables-such as connected insulin pens, smart auto-injectors for biologics, and pre-filled syringes with dose-logging-together with widely used wearables like continuous glucose monitors (CGMs) and Echocardiogram (ECG) patches are transforming how clinical data are captured. These devices do more than deliver medication or measure vital signs; many generate a continuous stream of metadata about how, when, and how well they are used. For example, a connected insulin pen can record dose, timing, and injection conditions, while a CGM or ECG patch can track glucose levels, heart rate and rhythm, activity periods, and device status (e.g., connectivity, battery). In clinical trials, this near real-time, longitudinal information allows data science and biostatistics teams to monitor both patient status and device performance remotely, enabling earlier detection of malfunction, data drift, or protocol non-adherence. Regulators are also embracing this shift: the FDA's guidance on Digital Health Technologies (DHT) for clinical trials emphasizes that these tools must be "fit-for-purpose"-meaning they need to reliably and accurately capture the clinical events they're intended to measure. To meet that bar, thorough testing and validation of the devices is essential-before they're used and throughout their lifecycle.

AI and Predictive Monitoring:

AI/ML are powerful tools for making sense of the constant flood of data these devices generate. These algorithms are particularly good at spotting subtle patterns that humans might miss. For example, if a smart injector gradually starts requiring more force to deliver medication, it may signal wear or partial blockage; if a skin temperature sensor begins reporting inconsistent values, it could indicate impending hardware failure. Similarly, for cardiac technologies such as implantable pacemakers, Implantable Cardioverter-Defibrillator (ICDs), or external ECG patches, AI models can flag early signs of malfunction by detecting unusual noise in the ECG trace, atypical pacing patterns, or gradual changes in lead impedance and battery performance. By continuously analyzing these streams, AI systems can issue early alerts-well before a complete failure occurs-so devices can be serviced or replaced in time, reducing trial disruptions and enhancing patient safety.

Human-in-the-Loop and Decision Support:

Even with increasingly sophisticated AI/ML techniques, human expertise remains central to safe and reliable device use. In current regulatory thinking, AI is expected to augment rather than replace clinical, engineering, data science, and statistical judgement. When a monitoring algorithm flags a potential issue, clinicians, biomedical engineers, data scientists, and statisticians should verify the underlying data, scrutinize the signal, and determine the appropriate course of action in a governed workflow. In this paper, "framework" refers to a prespecified, risk-based lifecycle scaffold with six pillars:

1) Context-of-use and decision governance: define the intended use of AI signals (screening vs decision support), review ownership, escalation paths, and decision rights.

2) End-to-end operational process: specify data flow, roles and responsibilities, timelines, and standard operating procedures from device capture to analysis and action.

3) Fit-for-purpose validation with prespecified operating characteristics: define the reference standard, performance targets (e.g., sensitivity, false-alert rate, calibration), and the evaluation analysis plan.

4) Documentation and traceability: ensure auditability from raw device measurements through preprocessing, feature generation, model output, and downstream decisions.

5) Ongoing statistical monitoring for data integrity and performance drift: implement prespecified centralized monitoring (e.g., SPC) for data quality, signal stability, and drift detection.

6) Controlled change management: prespecify how device firmware, algorithms, thresholds, and model updates are versioned, assessed, and implemented (including triggers for revalidation).

These pillars can be operationalized by specifying what belongs in the protocol, SAP, and operational quality documents, and by linking each component to applicable FDA/ICH expectations (Table 1).

Pillar	What to prespecify (examples)	Typical placement	Regulatory hook (examples)
Context-of-use & governance	Decision purpose; trigger thresholds; human review; escalation paths; documentation of actions.	Protocol + Monitoring Plan	ICH E6(R3) risk-based quality management; FDA DHT fit-for-purpose approach.
Operational process	Data flow; RACI; timelines; adjudication/CAPA workflow for alerts; communication plan.	Protocol + Ops SOPs	GCP data governance; auditability and traceability expectations.
Fit-for-purpose validation	Reference standard; subject-level split; performance targets; CIs; calibration; drift criteria.	SAP + Validation report	FDA DHT guidance on validation; GMLP principles (performance evaluation).
Traceability & provenance	Source-to-output lineage; preprocessing/feature specifications; version-stamping; reproducible pipelines.	SAP + Ops/QMS	Audit trails/records controls (as applicable); GCP traceability.
Ongoing monitoring (SPC/DQ)	Control limits; alert rate targets; re-baselining rules; dashboards; escalation and remediation.	SAP + Monitoring Plan	FDA DHT: operational specifications to reduce missing data and prevent data loss.
Controlled change management	Firmware/model update policy; PCCP elements; revalidation triggers; user-facing change communication.	QMS + SAP addendum	FDA PCCP guidance for AI-enabled devices; total product lifecycle (TPLC) approach.

The updated ICH E6(R3) Good Clinical Practice (GCP) guideline requires sponsors to implement fit-for-purpose, validated systems and to ensure data integrity, traceability, and auditability from raw device measurements through to analysis and decision-making, with statisticians playing a key role in defining validation strategies and documenting analytical workflows. For software-driven and AI-enabled devices, regulators such as the FDA, Health Canada, the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA), and the International Medical Device Regulators Forum (IMDRF) have issued Good Machine Learning Practice (GMLP) and transparency principles that emphasize rigorous model development, statistically robust performance evaluation, continuous performance monitoring, and clear information for users, including how human experts should interpret and oversee AI outputs. In parallel, the FDA's AI/ML-based Software as a Medical Device (SaMD) discussion paper and Action Plan stress the need for explainable, well-characterized algorithms, predefined change-control plans, and quality systems that support lifecycle evaluation of model performance. Taken together, these expectations imply that AI-powered monitoring tools in clinical research must rest on a sound statistical and methodological foundation-transparent assumptions, quantified uncertainty, pre-specified operating characteristics, and traceable links between device data, algorithm outputs, and human decisions-within a broader framework of documented quality and governance.

PROPOSED STATISTICAL FRAMEWORK FOR AI-ENABLED DEVICE RELIABILITY

From a statistician's point of view, this paper proposes a "framework" which is a structured way to ensure that device data and AI/ML outputs are reliable, traceable, and prespecified-so the study remains GCP-compliant, and results are defensible for regulators (aligned with ICH E6(R3) data integrity/traceability and ICH E9(R1) principles for pre-specification and robust inference). Below is a 6-pillar framework, with practical examples of where to place it in a protocol and SAP.

Pillar	What this pillar *locks* (statistician view)	What the statistician *specifies / delivers* (plain language)	Where it lives in the Protocol	Where it lives in the SAP
1. Context-of-Use	Meaning of AI output, operational action, and "which error matters" (false alarms vs missed true issues)	Defines AI outputs as analyzable variables (risk score/flag), sets **guardrails** ("monitoring only"), and prespecifies escalation logic & error tradeoff rationale.	Objectives/Endpoints; RBM/Data Integrity; DHT/Device section.	Section 7: AI-derived variables + use (monitoring summaries vs covariate/supportive), planned summaries, concordance outputs.
2. Required Processes	The end-to-end **data→quality checks→AI alert review→closure** workflow	Prespecifies data flow (expected frequency/windows, identifiers, latency rules), automated quality checks + query rules, centralized triage/escalation, and closure criteria with audit trail.	DHT/Device; Data Handling/DM; RBM/Data Integrity roles & escalation documentation.	Data handling/derivations definitions + Section 7 monitoring outputs; capture review outcomes as traceability variables.
3. Validation (Fit-for-Purpose)	What "good enough" means + how you prove device data + AI outputs are reliable	Separates **device measurement validity** vs **AI output validity**, prespecifies reference standard, evaluation design (splits/stratification), key metrics with uncertainty (e.g., sensitivity, false alert rate, PPV/NPV, calibration), and re-validation triggers.	DHT/Device fit-for-purpose validation; Quality/System validation; RBM/Safety monitoring triggers for re-validation.	Section 7 metrics/analysis sets/CI methods + Appendix shells & acceptance criteria.
4. Documentation & Traceability	Make every signal and action **auditable and reproducible**	Prespecifies what is versioned (device/firmware/model/thresholds/features), what is logged (alert→review→action→disposition), the raw→derived→decision chain, and reproducibility controls (locked specs/code, rerun capability, archived outputs).	Quality/System validation; DHT/Device provenance; RBM/Data integrity decision-log requirements.	Dataset specs/derivations + Section 7 traceability variables/reporting.
5. Ongoing Monitoring	Routine surveillance for drift, data issues, and systematic failures	Prespecifies what is monitored (missingness/latency, implausibles, signal quality, alert rate/score distribution shifts, site/version heterogeneity), how abnormality is defined (thresholds/control limits), actions, cadence and escalation-*avoiding reactive threshold tuning*.	RBM/Central monitoring + Safety monitoring + Data handling sections.	Section 7 outputs + appendices (dashboards/definitions) and archiving expectations.
6. Change Control	Managing device/model updates **without breaking interpretability**	Prespecifies what may change, approvers/governance, evidence required (impact assessment; comparability checks; rollback criteria), and how comparability is preserved (version-stamped datasets, stratified reporting, prespecified sensitivity analyses).	Quality Mgmt/Change Control + DHT/Device + RBM.	Section 7 version handling + comparability/sensitivity analyses.

INTEGRATING AI/ML WITH STATISTICAL OVERSIGHT: A HYBRID EVIDENCE MODEL

In practice, the strongest operating model for reliability in the age of AI is hybrid: AI/ML augments signal extraction, while prespecified statistical methods provide the inferential backbone and regulatory defensibility. A typical pipeline starts with high-frequency device telemetry (e.g., force, temperature, pressure, error codes, connectivity), where AI/ML can learn multivariate patterns that precede performance degradation or data loss (e.g., clustering or anomaly detection on joint sensor streams). These AI-driven risk scores can then be integrated into classical statistical work in three transparent ways:

- as screening signals feeding governed operational monitoring (e.g., SPC alerts to trigger review and corrective action).
- as covariates in survival/reliability models to quantify how evolving device signals relate to time-to-failure, and.
- as contributors to missing-data prevention and data-quality assurance through automated monitoring and alerting..

FDA's Digital Health Technologies (DHT) guidance explicitly recommends fit-for-purpose validation and highlights operational specifications and alerts to minimize missing data and prevent data loss, which aligns naturally with SPC-style centralized monitoring. Contemporary trials illustrate this hybrid approach. The Apple Heart Study used a smartwatch algorithm to generate irregular pulse notifications, with subsequent confirmatory ECG patch monitoring as the clinical reference, demonstrating how an algorithmic signal can be operationalized and clinically verified. Similarly, implantable cardioverter-defibrillator (ICD) programs routinely use automated remote monitoring diagnostics and alerts to detect lead system events early, with adjudication and downstream clinical actions documented within a governed workflow. In respiratory home-monitoring programs, connected-device streams have been paired with statistical process control approaches (including CUSUM-style monitoring) to support early detection of exacerbation signals.

Survival Analysis for Time-to-Failure

A practical question in any medical device program-whether it is an implant, injector, pump, sensor, or software-enabled system-is: how long does it perform as intended before something goes wrong? In clinical trials, that "something" can be defined in trial-relevant terms (e.g., time to first device malfunction, time to surgically remove a device from the body after it has been implanted (also known as explant) or revision, time to loss of performance beyond a prespecified threshold, time to first clinically meaningful device related adverse event, or time to discontinuation due to device-related issues). These are all time-to-event endpoints, and that is where SA becomes a natural fit.

Survival methods were originally developed for clinical endpoints like time to death or disease recurrence, but the mechanics map directly to device reliability. The "event" is simply redefined as a device failure, malfunction, or performance degradation, and the same statistical framework handles key realities of trials:

- Censoring: many subjects will not experience a failure during follow-up (study ends, subject withdraws, device removed for non-device reasons).
- Unequal follow-up: subjects enter at different times and have variable observation windows.
- Comparisons: you can compare treatment arms, device versions, sites, training approaches, storage conditions, or lots/batches.

In this setting, a Kaplan–Meier curve provides an intuitive description of "survival" (i.e., remaining event-free) over time, while models such as Cox regression or parametric survival models allow formal estimation and covariate adjustment. Parametric models-especially Weibull models-are commonly used in reliability contexts because they can represent increasing, constant, or decreasing hazard patterns over time. That flexibility matters in devices where risk may rise with wear, repeated cycles, or cumulative exposure.

A trial-relevant example: imagine a connected insulin pump program where you follow subjects for 12 months and capture time to first pump malfunction requiring replacement. SA lets you estimate the probability the pump remains event-free through 3, 6, 9, and 12 months, and it naturally incorporates subjects who discontinue early or never experience a malfunction during the study. If the data suggest the hazard increases after, say, 8 months, that is actionable: it can inform labeling, replacement guidance, preventive maintenance intervals, and post-market surveillance priorities.

From a biostatistics perspective, the value is not just descriptive. Survival models give you estimable quantities that are decision-relevant and regulator-friendly, such as:

• event-free probability at clinically meaningful timepoints,
• median time-to-event (where estimable), with confidence intervals,
• hazard ratios (or differences) between device groups/versions,
• predicted failure risk in the next X months conditional on survival so far (a “remaining useful life” view).

This is also where statisticians help keep the endpoint trial-grade: precise event definitions, adjudication rules if needed, handling of replacements, and alignment with the estimand strategy (e.g., how to treat device replacement, explant for non-device reasons, or software updates as intercurrent events). The goal is that “time-to-failure” is not a bio-medical engineering idea loosely embedded in a trial-it is a well-defined clinical trial endpoint with a defensible analysis plan.

In summary, SA and statistical process control (SPC) serve distinct but complementary roles in medical device evidence generation. SA quantifies durability and clinical reliability by characterizing when device failures or other time-to-event outcomes occur over follow-up. SPC, in contrast, supports early detection of drift, shifts, and anomalous patterns in longitudinal device performance signals while the trial is ongoing (and, where relevant, during post-market surveillance). As many contemporary devices produce rich telemetry (e.g., injection force, delivered volume, pressure, temperature, calibration indices, error-code frequency, connectivity dropouts, and signal-quality metrics), control charts offer a structured and interpretable way to separate expected common-cause variability from special-cause changes that may warrant investigation or mitigation. Used together, SPC strengthens proactive oversight and risk management, while SA provides the endpoint-level quantification needed to support clinical and regulatory conclusions.

In device trials, survival analysis (SA) supports endpoint-level durability and comparative reliability over follow-up, while statistical process control (SPC) supports operational, in-study monitoring of telemetry quality and performance drift. Keeping both roles explicit-endpoint inference vs ongoing monitoring-helps maintain regulatory clarity and prevents 'alerting' logic from being conflated with confirmatory effectiveness analyses.

Survival Analysis vs Statistical Process Control (SPC) in Medical Device Clinical Trials: A Practical Comparison

Dimension	Survival analysis (time-to-event)	Statistical Process Control (SPC)
Primary purpose	Quantify **when** a clinically meaningful event occurs and compare time-to-event between arms/subgroups	Detect **drift, shifts, anomalies** in device/process performance signals over time for early warning
Core question	“What is the distribution of time until failure/event, and does it differ by group?”	“Is performance stable, or has the process moved out of control (non-random change)?”
Typical use in device trials	**Reliability/durability** and clinical outcome evaluation; often part of **primary/secondary endpoint** strategy	**Ongoing monitoring** for safety/performance oversight; often part of **quality/risk management** rather than endpoint testing
Data structure	One (or few) event times per subject (time to first event, recurrent events possible)	Many repeated measurements per unit/time (streaming or frequent signals)
Example endpoints/signals	Time-to-device-failure, time-to-reintervention/revision, time-to-first device-related SAE, time-to-healing	Injection force, temperature drift, battery metrics, sensor calibration indices, signal dropout rate, algorithm confidence score
Key statistical feature	**Censoring** is central (no event by end of follow-up, dropout, competing events)	Focuses on **process stability** and **change detection**; less about censoring, more about sequential behavior
Common methods	Kaplan–Meier, log-rank, Cox proportional hazards, parametric survival models; competing risks as	Shewhart control charts, **CUSUM**, **EWMA**, run rules, change-point detection

Dimension	Survival analysis (time-to-event)	Statistical Process Control (SPC)
	needed	
Output/interpretation	Survival/reliability curves, hazard ratio, median time-to-event (if estimable), cumulative incidence at fixed times	Control charts with control limits, out-of-control flags, estimated time of shift, trend/drift diagnostics
Decision cadence	Typically **scheduled analyses** (interim/final) per SAP and protocol	Often **near-real-time or frequent** review as data accrue
Strengths	Directly supports **clinical/reliability claims** about durability and risk over time	Provides **early warning** of emerging device issues before they manifest as clinical events
Limitations	Needs sufficient event accrual; assumptions (e.g., proportional hazards) may need checking	Alerts require governance to manage false alarms; control limits depend on base

Building on this distinction, the choice of SPC technique depends on whether the risk profile is dominated by abrupt excursions (e.g., sudden sensor malfunction) versus gradual drift (e.g., calibration decay), and on whether monitoring is conducted at the subject-, device-, site-, or lot-level.

In this context, Shewhart-type charts are typically used to detect abrupt excursions, while CUSUM (Cumulative Sum) and EWMA (Exponentially Weighted Moving Average) charts improve sensitivity to gradual drift or persistent small shifts that may not cross a simple threshold. Operationally, these tools support centralized monitoring by identifying emerging site-, lot-, or device-specific patterns-for example, a sustained increase in device alarms, rising rates of missing telemetry uploads, or deterioration in signal-quality indices at a particular site or device version-thereby triggering targeted follow-up such as retraining, device inspection/replacement, firmware investigation, or assessment of workflow and manufacturing factors. Importantly, this monitoring is not limited to 'device quality assurance'; it directly supports trial data integrity and participant safety, because device anomalies can increase missingness, compromise endpoint measurement, or introduce clinically meaningful risk.

For protocol and Statistical Analysis Plan (SAP) specification, it is typically appropriate to frame SPC outputs as risk indicators used for operational decision-making rather than confirmatory endpoints. Pre-specification should therefore focus on: (i) which device metrics are monitored and at what level (subject, site, lot/batch, device version), (ii) the baseline period and "in-control" assumptions used to set control limits, (iii) the alerting rules (e.g., points beyond control limits, sustained runs, or CUSUM/EWMA threshold crossings), and (iv) the governance pathway linking alerts to documented review and corrective and preventive actions (CAPA). Statisticians add value by selecting fit-for-purpose charting approaches and calibrating alerting rules to balance sensitivity against false alerts-minimizing operational noise while maintaining timely detection of true performance degradation.

Automated Data-Quality Monitoring to Prevent Missingness and Reduce Noise

Minimum fit-for-purpose validation package (recommended for SAP/appendix):

- Define the reference standard and adjudication rules for 'true' events and false alerts.
- Prespecify data partitioning to prevent leakage (e.g., subject-level split; stratify by site/device version where relevant).
- Prespecify operating characteristics and targets (e.g., sensitivity, false-alert rate, PPV/NPV, calibration) with confidence intervals.
- Define drift monitoring metrics and triggers for investigation and revalidation (including version-stamped model/device changes).

Automated data-quality monitoring is a practical statistician-led control layer that protects the integrity of AI signals and downstream analyses in device trials. In connected-device settings, missing data are rarely "random"-gaps often reflect connectivity issues, device non-use, sensor malfunction, or pipeline failures, all of which can bias endpoint estimation and inflate false AI alerts. Therefore, prespecified, automated checks should run continuously

to detect and correct data issues early, before they propagate into derived variables, model inputs, or analysis datasets. This aligns directly with this paper's emphasis that AI/ML is most defensible when embedded within prespecified statistical oversight and governed operational workflows.
Automated monitoring should cover four core domains

- **Completeness (missingness):** missing expected transmissions/records; gaps in daily uploads; incomplete device sessions.
- **Timeliness (latency):** delayed uploads beyond an acceptable window; sync failures; backlog spikes.
- **Plausibility / range & consistency:** out-of-range values; impossible sequences (e.g., negative durations); inconsistent device status (e.g., "battery depleted" but normal telemetry).
- **Signal quality (when applicable):** artifact/noise flags (e.g., uninterpretable physiologic waveforms) that can drive spurious AI outputs.

To keep monitoring "regulator-defensible," thresholds and responses should be prespecified (not tuned post hoc). A simple approach is to define triggers such as **Missingness threshold:** e.g., >X% expected data missing over 7 days at a site/device version would lead to review and query.

Within the proposed statistical framework, automated data-quality monitoring should use prespecified operational triggers to detect and correct upstream issues before they contaminate AI signals or analysis datasets. For example, a latency spike-such as the median upload delay doubling versus a prior baseline window-should prompt investigation of the data pipeline and connectivity workflows; a plausibility increase-such as out-of-range or otherwise implausible values rising to more than two times baseline-should trigger evaluation for sensor drift, device malfunction, or ETL/derivation defects; and signal quality degradation-such as a high proportion of artifact-laden or uninterpretable records-should be flagged to prevent noise-driven false AI alarms. To remain audit-ready and GCP-aligned, every flagged issue must be traceable from raw data → quality rule → query → resolution, with documentation of the trigger timestamp, impacted device/site/version, rule violated, query ID, resolution status/date, and whether the record was corrected or retained with an audit trail. This approach improves downstream analyses in three concrete ways: it reduces bias from informative missingness by identifying and managing device-related data loss transparently, stabilizes AI monitoring performance by reducing noise-driven false positives and improving feature integrity, and strengthens interpretability of survival analysis and SPC outputs by ensuring that time-to-failure and drift signals are not confounded by data transfer artifacts-consistent with the paper's hybrid model in which AI risk scores contribute to missing-data prevention and data-quality assurance through automated monitoring and alerting.

Proposed Protocol and SAP placement.

Protocol placement

Automated data-quality monitoring should be prespecified in the protocol in the same places regulators expect to find **fit-for-purpose system controls** and **data integrity safeguards**. Practically, this belongs in: (i) the **Digital Health Technology (DHT) / Device** section (define the device data streams, expected collection frequency/windows, upload/transfer expectations, and the minimum quality/validity requirements for usable telemetry), (ii) **Data Handling / Data Management** (describe automated checks for completeness, timeliness/latency, plausibility/out-of-range values, and signal quality; define query generation and resolution workflows), and (iii) **Risk-Based Monitoring / Data Integrity** (define escalation rules when thresholds are breached-e.g., site queries, participant contact if applicable, device evaluation, and when issues rise to a cross-functional review). In protocol language, emphasize that these controls are implemented to protect **GCP-aligned traceability and auditability** from raw device measurements through analysis datasets, and to minimize missingness/noise that could compromise trial interpretability.

SAP placement

In the SAP, automated data-quality monitoring should be captured in **Section 7 (Statistical Methods)** and/or a dedicated **Data Handling/Derivations** subsection to ensure statistical prespecification and consistent reporting. The SAP should: (i) define the operational data-quality endpoints/metrics (e.g., missingness rate per participant-week/device-day, upload latency distributions, out-of-range flag rates, artifact/uninterpretable signal rates), (ii) define prespecified thresholds (or control-limit logic) and the descriptive summaries that will be produced overall and stratified by time, site, device model/firmware, and other relevant factors, (iii) specify how quality flags and

query outcomes are captured as analysis variables (timestamps, rule violated, review outcome, action taken, and final disposition), and (iv) describe how data-quality outputs inform downstream analyses without introducing bias (e.g., documenting informative missingness patterns, supporting sensitivity summaries where needed, and ensuring SA/SPC outputs are interpreted in light of confirmed data integrity issues). This makes the quality layer reproducible, auditable, and aligned with the paper's lifecycle governance approach.

CONCLUSION

A regulator-ready operating model for AI-enabled medical device reliability is hybrid: AI/ML improves signal extraction from high-frequency telemetry, while prespecified statistical methods provide the inferential backbone and auditability expected in clinical research. By combining time-to-event reliability endpoints (via survival analysis) with centralized operational monitoring (via SPC and automated data-quality controls), sponsors can detect emerging issues early without compromising the integrity of endpoint analyses. The six-pillar framework presented here is intended to translate evolving FDA/ICH expectations into implementable protocol, SAP, and operational specifications that are transparent, traceable, and fit-for-purpose across the total product lifecycle.

IMPLEMENTATION CHECKLIST (PROTOCOL / SAP / OPERATIONS)

• Protocol: define context-of-use, governance, alert response workflow, and monitoring scope (including missingness and latency definitions).

• SAP: prespecify endpoints, estimands, survival/SPC methods, validation targets and operating characteristics, and sensitivity analyses for data-quality issues.

• Operations/QMS: ensure audit trails, version control, PCCP/change control, CAPA documentation, and periodic drift/performance reviews.

REFERENCES

• U.S. Food and Drug Administration (FDA). Digital Health Technologies for Remote Data Acquisition in Clinical Investigations. Final guidance. Issued December 22, 2023; web posting July 29, 2024.

• FDA. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence (AI)-Enabled Device Software Functions. Guidance. August 18, 2025.

• International Medical Device Regulators Forum (IMDRF). Good machine learning practice for medical device development: Guiding principles. IMDRF/AIML WG/N88 FINAL:2025. Published January 29, 2025.

• FDA. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. January 2021.

• FDA. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback. April 2019.

• International Council for Harmonisation (ICH). E6(R3) Guideline for Good Clinical Practice (GCP).

• ICH. E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials. 2019.

• FDA and European Medicines Agency (EMA). Guiding Principles of Good AI Practice in Drug Development. January 14, 2026.

• Montgomery, D.C. Introduction to Statistical Quality Control. 8th ed. John Wiley & Sons; 2020 (eBook available).

• Provost, L.P., & Murray, S.K. The Health Care Data Guide: Learning from Data for Improvement. John Wiley & Sons; 2011.

• Wheeler, D.J. Understanding Variation: The Key to Managing Chaos. 2nd ed. SPC Press; 2000.

• Page, E.S. Continuous Inspection Schemes. Biometrika. 1954;41(1/2):100–115. (CUSUM control chart).

• Lucas, J.M. Fast Initial Response for CUSUM Quality-Control Schemes. Technometrics. 2000;42(1):102–110.

• Buddavarapu, M.S., et al. Digital Health Technology in Clinical Trials: Opportunities, Challenges and Biometrics Future Re-imagined. PharmaSUG 2022. Paper DHT-194. (LexJansen).

• Kavalci, S., & Al-Jubouri, M. Risk-Based Monitoring in Clinical Trials with Machine Learning Powered Anomaly Detection. PHUSE 2023. Paper AR07. (LexJansen).

• Bradford, J., & Murray, J. The promise of wearable technology in clinical trials: a proof of concept to deliver it? PHUSE 2019. Paper RW02. (LexJansen).

• Qubeck, B., & Vijendra, S. Wearable Technology: Enhancing Patient Monitoring & Real-World Evidence. PHUSE US Connect 2025. Paper RE08. (LexJansen).

• Kumar, P., & Kumar, P. Regulatory Challenges and Strategies for AI/ML (SaMD) Enabled Medical Device Data Submission. PHUSE US Connect 2025. Paper SA07. (LexJansen).

• Perez, M.V., Mahaffey, K.W., Hedlin, H., et al. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. N Engl J Med. 2019;381:1909–1917. doi:10.1056/NEJMoa1901183.

• Apple Heart Study: Assessment of Wristwatch-Based Photoplethysmography to Identify Cardiac Arrhythmias. ClinicalTrials.gov Identifier: NCT03335800 (accessed 29-Jan-2026).

• Apple Heart Study Statistical Analysis Plan (Version 1.6, 09-Nov-2018). ClinicalTrials.gov large document (Prot_SAP_000.pdf) (accessed 29-Jan-2026).

• Stehlik, J., Schmalfuss, C., Bozkurt, B., et al. Continuous Wearable Monitoring Analytics Predict Heart Failure Hospitalization: The LINK-HF Multicenter Study. Circ Heart Fail. 2020;13(3):e006513. doi:10.1161/CIRCHEARTFAILURE.119.006513.

• Link-HF: Multisensor Non-invasive Telemonitoring System for Prediction of Heart Failure Exacerbation. ClinicalTrials.gov Identifier: NCT03037710 (accessed 29-Jan-2026).

• LINK-HF2 – Remote Monitoring Analytics in Heart Failure. ClinicalTrials.gov Identifier: NCT04502563 (accessed 29-Jan-2026).

• Ultra Long-Term Sleep Monitoring Using UNEEG™ Medical 24/7 EEG™ SubQ (deep learning-based automated sleep analysis). ClinicalTrials.gov Identifier: NCT04513743 (accessed 29-Jan-2026).

• Ahrens, A., et al. Automatic sleep staging – EEG SubQ data: Polysomnography agreement and algorithm performance. (Clinical study registered: NCT04513743). Published 2024. (PubMed).

• DELTA: Develop and Validate Machine-Learning Algorithm to Detect Atrial Fibrillation with Wearable Devices. ClinicalTrials.gov Identifier: NCT05795842 (accessed 29-Jan-2026).

• PROVISION-AF: Prospective Validation Study of AI-based Prediction Algorithm for Paroxysmal Atrial Fibrillation from Sinus Rhythm ECG. ClinicalTrials.gov Identifier: NCT05725187 (accessed 29-Jan-2026).

Contact Information

Your comments and questions are valued and encouraged.
Contact the author at:
Author Name- Shrutimita Pokhariyal
Principal Statistician
Symbio
Email-shruti.pokhariyal@symbioresearch.com
Number- +91 7528806366

Co-Author Name- Parveen Kumar
Director-Statistics
Geninvo
Email-parveen.k12@geninvo.com
Number- +91 9871819209