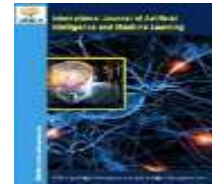




# International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

## Adversarial Resilience in Machine Learning Models Deployed on Connected Medical Devices

Venkata Sai Abhinav Piratla

Independent Researcher, India. E-mail: [piratla.abhinav1@gmail.com](mailto:piratla.abhinav1@gmail.com)

### Abstract

Machine learning controllers now dose insulin inside closed-loop artificial pancreas systems. The adversarial machine learning literature has established that such models can be manipulated, and the author's own prior work catalogued this device class's attacks and candidate defences. This paper argues the missing artefact is a benchmark. Five structural gaps follow, each stated as what was not found rather than what does not exist: no empirical adversarial testing on real devices, no link from robustness to clinical harm, no validation of defences on constrained hardware, no regulator-aligned adversarial robustness testing standard, and almost no adversarial benchmarks for physiological time series. No systematic screening protocol was run, so no screening count exists. The gaps generate requirements forming a specification stratified along five axes: device tier, signal modality, patient stratum, threat model and clinical harm severity, with perturbation budgets in clinical units, resource cost alongside every robustness figure, and per-patient rather than cohort-mean results. Existing standards on neural network robustness, adversarial machine learning terminology and medical device safety and security risk management are engaged; none supplies the missing test method. The paper reports no adversarial robustness measurements, because none were performed. Its only empirical figures are anomaly-detection results from the author's separately published on-device detector, included strictly to illustrate honest on-device evaluation, not as evidence of adversarial robustness. The benchmark is specified, not instantiated: a minimum viable starting configuration is proposed so it can be built, and carries no results. Building it requires researchers, manufacturers and regulators to agree on the axes before arguing about the numbers.

**Keyword:** adversarial machine learning; medical device security; artificial pancreas; benchmark specification; robustness evaluation; resource-constrained inference; continuous glucose monitoring; regulatory science

### 1. Introduction

An insulin pump driven by a machine learning controller is a safety-critical system whose input arrives over a wireless link that its manufacturer does not own. That single sentence contains the whole problem. The controller inherits every weakness of the model class it belongs to, and it inherits them in a setting where the consequence of a wrong output is a dose of a drug with a narrow therapeutic margin, delivered to a person who is asleep.

The security community has established that the connectivity layer of these devices is exploitable in practice, not in principle. A vulnerability in the wireless radio-frequency protocol used by a family of Medtronic MiniMed insulin pumps was documented publicly, with the advisory stating that the protocol does not properly implement authentication or authorisation and that an attacker with adjacent access can inject, replay, modify or intercept data, and could change pump settings and control insulin delivery [19]. Three severity scores are on record for that vulnerability across the advisory and the national vulnerability record, under different scoring-system versions, so none is quoted here. Two years earlier, a regulator-approved firmware update was issued as a corrective action across approximately 465,000 implanted cardiac pacemakers in the United States, after confirmation that the vulnerabilities involved could allow an unauthorised user to access a device with commercially available equipment and modify programming commands, with potential patient harm from rapid battery depletion or inappropriate pacing [20]. Neither of those was an attack on a model. Both established that the channels a model depends on are reachable.

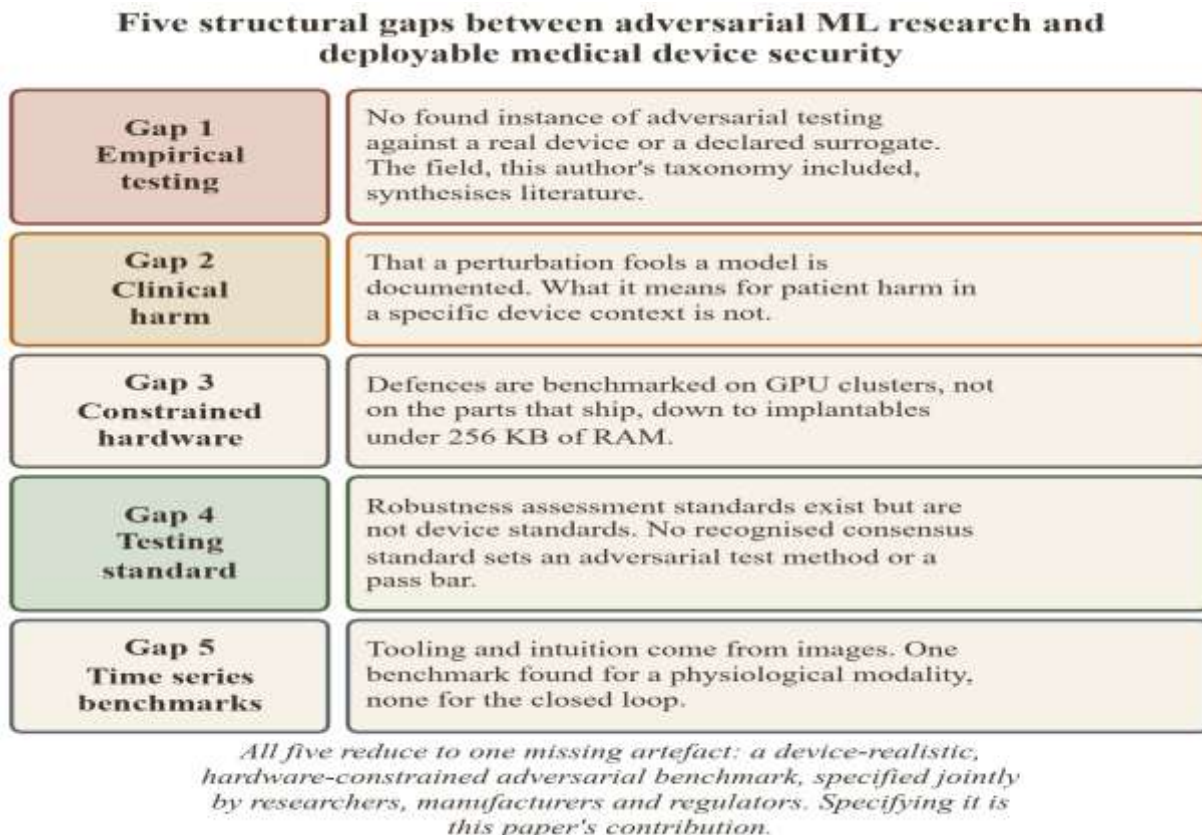
The author's own prior work has mapped what happens when adversarial machine learning is layered onto that reachable substrate. A systematic review of the artificial pancreas security and reliability literature established the field's shape and its blind spots [2]. A subsequent paper established a structured taxonomy of adversarial attacks against medical device AI controllers spanning training-phase, inference-phase, privacy and model-integrity categories, evaluated a catalogue of candidate defence mechanisms, mapped both to the MITRE ATLAS technique vocabulary, aligned the result with regulatory cybersecurity and AI lifecycle guidance, and derived tiered defence recommendations from device resource profiles [1]. A companion paper translated threat coverage into an actionable control framework for AI-enabled continuous glucose monitoring, mapping threats

to specific regulatory requirements and to established control catalogues [4]. Separately, an unsupervised on-device anomaly detector for artificial pancreas telemetry was built and evaluated [3].

None of that taxonomy, defence catalogue, framework mapping or tiering is restated as a contribution here. It is established prior work, it is cited as such, and this paper stands on it rather than repeating it. What this paper adds is the thing that catalogue work cannot supply on its own.

The problem is this. A taxonomy tells a manufacturer which attacks exist and which defences have been proposed against them. It cannot tell that manufacturer whether a given defence, compiled and quantised down onto the microcontroller inside a specific implantable, still resists a specific attack, at what cost in memory and latency, for which patients, and with what consequence when it fails. Answering that requires measurement under a fixed protocol, and no such protocol exists for this device class. The literature contains robustness numbers, but they were produced on image classifiers running in float32 on graphics processors, self-evaluated by the authors proposing the defences in a practice that independent re-evaluation has repeatedly found to overstate robustness [8], [10], reported as population averages, and denominated in units with no clinical meaning. Transferring them to an implantable is not conservative extrapolation. It is a category error.

This paper therefore takes a position rather than reporting an experiment. The position is that the field's central unmet need in this area is not another attack taxonomy or another defence proposal, but a device-realistic, hardware-constrained adversarial robustness benchmark, specified jointly by researchers, device manufacturers and regulators, so that the word robust stops meaning something different to each of those three groups. The paper's contribution is a requirements specification for that benchmark: the axes it must stratify along, the tuple it must publish for every cell, the protocol by which cells are filled, the reasons each requirement is non-optional, and a first proposed configuration concrete enough to build. Section 3 states how the gap claims were arrived at and what that permits them to say. Sections 4 through 8 set out five structural gaps that motivate the specification, Section 9 states the specification itself and closes with a minimum viable starting configuration, and Section 10 gives a worked illustration of what honest evaluation of an on-device model looks like when adversarial claims are withheld.



**Figure 1.** The five structural gaps this paper is organised around, and the single artefact all five reduce to. The gaps are stated in the author's own terms, including the acknowledgement that his published taxonomy is itself literature synthesis rather than a red-team result. Three of the five were raised in the discussion of the author's published taxonomy [1]; Table 1 states which, and states what this paper adds.

The layout was developed by the author for unpublished internal work and is used here for the first time in print.

## 2. Background and Prior Contribution

This section fixes the boundary between what is already published and what this paper claims, because the distinction carries the paper's argument as well as its integrity.

### 2.1 What the prior work established

Four prior publications form the base. Table 1 states what each established and, explicitly, what this paper does not restate from it.

One element of that boundary belongs in prose as well as in the table, because it concerns this paper's organising structure rather than its content. Three of the five gaps set out in Sections 4 through 8 were raised in the discussion section of the published taxonomy [1], and a fourth appears there as a stated limitation. The enumeration is therefore not this paper's discovery and is not claimed as one. What [1] did with those gaps was name them and call for adversarially annotated datasets. This paper's contribution begins where that call stops, at the question of what an artefact capable of closing them would have to contain, which is the question a call for datasets leaves open.

**Table 1. The prior-work boundary. The left column is established; the right column states this paper's relationship to it. Nothing in the middle column is presented as a contribution of this paper.**

| Prior work                                                                                                                                                                                                                 | What it established                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | What this paper does with it                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Systematic review of artificial pancreas security and reliability, covering the closed-loop literature [2]                                                                                                                 | The field's coverage and its blind spots, in particular that cyberattacks on these systems were treated as simulated or theoretical rather than empirically exercised                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Cited as the source of the empirical-testing gap. The review is not re-summarised.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Taxonomy of adversarial attacks against medical device AI controllers, with a catalogue of defence mechanisms, MITRE ATLAS mapping, regulatory alignment and tiered defence recommendations by device resource profile [1] | Which attacks apply to which closed-loop component; which defences have been proposed; how both map to established technique and terminology vocabularies; which defences are feasible at which hardware tier. Its discussion also raised three research gaps: no standardised adversarial robustness benchmark for medical time-series data, no empirical validation of server-grade defences on the embedded platforms these devices use, and adversarial attacks on reinforcement-learning controllers being understudied. Its limitations section recorded that its clinical impact severity ratings derive from simulation studies. | Cited as the premise. The attack taxonomy, the defence catalogue, the ATLAS mapping and the tiering are <b>not</b> reproduced, re-derived or extended here. The tiering is reused only as a benchmark stratification axis, with attribution. <b>On the gap enumeration: Gaps 5, 3 and 1 here correspond to the three gaps raised in [1]'s discussion, and Gap 2 develops a point [1] recorded as a limitation. The enumeration is not claimed as this paper's contribution. [1] called for adversarially annotated datasets; this paper specifies what such a benchmark would have to satisfy, namely the axes, the report tuple, the protocol ordering, the harm-classification structure, the governance condition and a starting configuration. Gap 4 is not among [1]'s gaps: [1] presents regulatory alignment as one of its own contributions, and Section 7 argues that alignment is not sufficient, which is disagreement with the prior work rather than restatement of it.</b> |
| Security framework for AI-enabled continuous glucose monitoring, translating regulatory guidance into actionable controls and mapping threats to requirement and control catalogues [4]                                    | That threat coverage for this device class can be expressed as a traceable set of controls against named regulatory requirements                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Cited to draw a boundary. This paper makes no control-mapping claim. Its regulatory argument is about the absence of a <i>testing standard</i> , which is a different object from a control map. Section 7 states that distinction directly.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |

|                                                                                            |                                                                                                                                    |                                                                                                                                                                               |
|--------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AP-GUARD, an unsupervised on-device anomaly detector for artificial pancreas telemetry [3] | That a lightweight detector can run on-device as a parallel monitor and be evaluated honestly under patient-based cross-validation | Cited once, in Section 10, as a worked illustration of disciplined on-device evaluation. Its metrics are anomaly-detection metrics and are labelled as such at every mention. |
|--------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 2.2 The setting a benchmark has to cover

The specification in Section 9 is grounded in the closed-loop artificial pancreas because that system is concrete enough to argue about. It has five components: the continuous glucose monitoring sensor, the insulin pump, the onboard AI controller, the companion mobile application, and the cloud backend. Three trust boundaries in that chain sit outside the manufacturer's control: the sensor-to-controller link, the controller-to-pump link, and the device-to-cloud link. Each is a point at which data can be intercepted or manipulated by a party the manufacturer neither authenticates nor patches. This decomposition is taken from the prior taxonomy [1] and appears here as the setting a benchmark must cover, not as a fresh attack-surface analysis.

The prior taxonomy also organises devices into four hardware constraint classes: severely constrained implantables, on ARM Cortex-M class parts with under 256KB of RAM; moderately constrained wearables; lighter controllers and gateways; and cloud infrastructure. Those four tiers reappear in Section 9 as a stratification axis. The reason is that they are the axis along which the transferability of published robustness evidence actually breaks.

One placement matters for reading the rest of the paper. A commercial insulin pump and a continuous glucose monitor transmitter are Tier 2 wearables, not Tier 1 implantables, so the closed-loop system this paper is grounded in sits at Tier 2. Tier 1, the severely constrained implantable under 256KB of RAM, is the case the argument extrapolates to rather than the case it analyses. Where the sub-256KB envelope appears below it marks the tightest point on the axis, not the grounding device, and nothing here depends on the grounding device sitting at that point: the on-device model in Section 10 is a 95KB artefact that fits inside even the Tier 1 envelope, which is precisely why the missing quantity is a robustness figure at a declared tier rather than a tighter resource claim.

## 2.3 What "robust" currently means, and why that is the problem

In the adversarial machine learning literature, a robustness claim is a number: accuracy retained under a perturbation of bounded size in some norm. Three properties of that convention matter here. It is defined on a static input, usually an image. It is denominated in a unit with no external meaning, a pixel-space norm. And it is commonly produced by the same group proposing the defence being evaluated, a practice whose consequences are documented rather than presumed [5], [8], [10]. Each property is defensible inside the research community that produced it, where the goal is comparative progress on a shared task, and none survives contact with a regulated device. A clinician cannot act on a pixel norm. A glucose trace is not a static input; it is a stream with temporal structure, physiological limits on rate of change, and a control loop reacting to it in real time. And the route by which a manufacturer's test evidence becomes reviewable by a regulator runs, in other areas of device engineering, through a standard the regulator has recognised: a manufacturer may elect to declare conformity with a consensus standard from a list the regulator maintains and periodically modifies [21]. Nothing on that list makes an adversarial robustness claim reviewable on those terms, which is Section 7's argument in a sentence.

## 3. How the Gap Claims Were Arrived At, and Why Current Robustness Evidence Does Not Transfer

### 3.1 Search and screening: what was done, what was not, and what that permits

Each of the five gaps in Sections 4 through 8 is a claim that something is absent from a published literature, and a claim of that form is worth exactly as much as the search behind it. This paper states its search before its gaps, including where there was no search.

No systematic search was conducted for this paper. No review protocol was registered, no database query strings were executed to count studies against a gap criterion, and no screening was performed against inclusion and exclusion criteria. There is therefore no records-screened count, and none is reported: inventing a number in a paper arguing that unmeasured claims should not be trusted would be self-refuting. What the gap claims rest on is set out instead.

Gaps 1, 2, 3 and 5 rest on two prior literature exercises by the author, neither designed for this purpose. The systematic review of artificial pancreas security and reliability surveyed that domain's published work and reported that cyberattacks on these systems were treated as simulated or theoretical rather than empirically exercised [2], but it was scoped to security and reliability rather than to adversarial robustness benchmarking, so its coverage of adversarial evaluation practice is incidental to its design. The published taxonomy [1] was

assembled from published sources and inherits their coverage, a limitation it records about itself and which Section 4 restates as a self-implication rather than a defence. Neither set out to enumerate empirical adversarial evaluations on device hardware, per-patient adversarial reporting, or time-series adversarial benchmarks, so neither can be cited as a count of them.

Gap 4 is the one claim for which a targeted search was run and can be described. The artefact sought was any recognised consensus standard, or any regulatory guidance document, specifying a test method or a pass threshold for the adversarial robustness of a machine learning component in a medical device. What was checked, on 10 and 11 August 2026: the current final premarket cybersecurity guidance for medical devices [16] and the guidance addressing artificial-intelligence-enabled device software functions [17]; the regulator's list of recognised consensus standards as most recently modified [21]; the catalogue records for the ISO/IEC 24029 series on assessment of the robustness of neural networks [22], [23]; the national adversarial machine learning taxonomy and terminology reference [24]; and the consensus standards for medical device risk management and for device security risk management [25], [26]. Section 7 reports what each contains and why none supplies the missing method. This is a targeted check of the documents a reader would reach for first, not a systematic review, and it is reported at that strength.

Every gap claim below is therefore worded as a claim about what was not found, not about what does not exist. That is the weaker claim and the defensible one, and a single published counter-example refutes it, which is the correct standing for an argument of this kind. Such a counter-example would close the specific gap without touching the specification in Section 9, whose value rests on the artefact not yet existing in a form anyone can run.

### **3.2 Four mismatches between published robustness evidence and a device**

Four mismatches separate published robustness evidence from anything a device manufacturer could rely on. They are stated compactly here because Sections 4 through 8 develop their consequences and Section 9 answers them one at a time; the point of naming them first is that the five gaps are not five unrelated shortfalls but four category errors with five visible symptoms.

**Threat-model mismatch.** The most widely used threat model bounds an adversary by a norm ball around a fixed input and lets it move freely inside that ball [6]. On a continuous glucose stream that admits perturbations no physiological process could produce and no plausible adversary could inject: instantaneous jumps, oscillations faster than the sensor samples, trajectories inconsistent with glucose-insulin dynamics. A threat model that admits impossible attacks optimises defences against the wrong adversary. Sequence classifiers have been shown to need their own attack formulation rather than a reused image-domain one [13], with the clinical case demonstrated on electrocardiogram models [12].

**Evaluation-integrity mismatch.** Robustness claims in this field have a poor track record when re-evaluated by third parties. Defensive distillation, which appears in the prior taxonomy's defence catalogue [1], was substantially undermined by a stronger attack construction shortly after publication, in work that also defeated a set of detection-based defences evaluated in the same paper [5]. The pattern recurred at scale: a standardised parameter-free attack ensemble applied to more than fifty previously published models found reported robust accuracy overstated for the large majority of them, in many cases by more than ten percentage points [8], and subsequent work turned to raising confidence in evaluations rather than proposing defences, on the grounds that the evaluation step is where the failures concentrate [10]. The image domain's response was a shared, versioned, third-party-run benchmark, so that self-evaluation stopped being the unit of evidence [9]. Medical devices have no equivalent, so a manufacturer can make an adversarial robustness claim about a controller that nobody can check.

**Platform mismatch.** The published robustness numbers this paper has examined were measured on full-precision models running on graphics processors. Deployed device models are quantised, sometimes pruned, compiled for a microcontroller, and held inside a memory envelope smaller than the activation footprint of many research checkpoints. Reduced numerical precision changes adversarial robustness rather than leaving it intact [15], so the deployed model is a different object from the evaluated one in exactly the property under evaluation.

**Outcome mismatch.** The unit of account in the literature is a change in model accuracy; the unit of account for a device is patient harm, and the mapping between them is neither monotonic nor uniform. A misclassification an existing interlock absorbs is not a clinical event. A small, sustained, undetected bias toward over-delivery is one, even though it registers as a smaller accuracy change. The stakes were raised early and prominently for medical machine learning generally [11], but raising stakes is not quantifying them.

## **4. Gap 1: No Found Instance of Empirical Testing on Real Devices**

The work behind this paper found no published instance of an adversarial robustness evaluation executed against a marketed medical device, or against a declared hardware surrogate of one. The work it did find,

including the author's own published taxonomy, is built from literature synthesis rather than from red-team results on production systems [1], [2].

That sentence is deliberately self-implicating, and the self-implication is the point. A taxonomy assembled from published sources inherits the coverage of those sources. If the underlying literature contains no adversarial test executed against a real insulin pump, then no amount of careful classification will produce one, and a reader who mistakes a well-organised catalogue for an evidence base has been misled by the organisation rather than by any individual claim. The honest description of the current state is that the field knows, in structured detail, what could be done to these devices, and that nothing measured on the devices themselves was found describing what happens when it is.

The reasons are structural rather than a matter of insufficient interest. Production devices are regulated, so testing one in clinical use is not a research decision. Manufacturer firmware is closed, so an external group cannot instrument the controller under test. Devices safe to attack, being bench units or engineering samples, are not distributed to researchers. And the datasets that do exist are patient telemetry, governed by access controls that do not contemplate adversarial experimentation. Every obstacle is legitimate, and together they make the most needed class of experiment the least available.

A benchmark does not remove those obstacles, but it changes what they block. If a specification defines a device profile, a resource envelope, a signal modality and an attack budget precisely enough, then a researcher without device access can still test a faithful surrogate at the correct tier, and a manufacturer with device access can run the identical protocol internally and publish comparable numbers. The obstacle stops being a wall and becomes a division of labour. That only works if both parties are running the same protocol, which is the requirement Section 9 states.

Requirement it generates. The benchmark must be runnable on a declared surrogate at a declared tier, with the device profile and resource envelope specified precisely enough that a manufacturer's internal run and an external run are comparable results rather than two unrelated studies.

## 5. Gap 2: Robustness Is Not Connected to Clinical Harm

It is well established that a perturbation can fool a model, including in the clinical setting, where the stakes were raised for medical machine learning generally [11] and small perturbations were shown to change the diagnostic output of an electrocardiogram model [12]. The work behind this paper found no published study that quantifies, for a specific closed-loop device, what a successful perturbation means for patient harm. The nearest thing found was severity rating by expert judgement or by simulation, including in the author's own taxonomy, whose limitations section records that its clinical impact severity ratings derive from simulation studies [1].

The gap is not a failure of concern. It is a missing translation layer. Reported robustness is a property of a classifier; harm is a property of a patient and a delivery mechanism over time. Getting from one to the other requires knowing which model failures the surrounding system absorbs, which it propagates, how quickly a propagated failure reaches a clinical consequence, and whether that consequence is reversible. None of that information is in an accuracy delta.

The artificial pancreas is an unusually tractable place to build that translation, which is part of why the specification is grounded there. Glycaemic control already has accepted outcome measures, including time in range and time below range, and the clinical consequences of dosing error are directional and asymmetric in a way clinicians describe consistently: over-delivery and under-delivery are not symmetric risks, and their time courses differ. A benchmark for this device class can therefore classify an induced failure by its clinical consequence rather than only by its statistical form. Table 2 proposes a structure for doing so.

That structure has to be positioned carefully, because medical device engineering already has a consensus framework for harm and it is not this one. Risk management for medical devices is governed by a published consensus standard under which harm is characterised by severity and by probability of occurrence across the device lifecycle [25], and security risk management for devices is specified as a process performed inside that safety process rather than beside it [26]. Table 2 is not a competing severity scale and must not be read as one. Its four classes capture what the severity dimension does not: whether an induced model failure was contained, by what, and whether anything in the declared configuration would have surfaced it. Those are containment and observability properties of the system around the model, and recording them is what makes a severity estimate under [25] possible at all. Severity classes without a containment record hand the risk file a number it cannot trace; a containment record that never maps to severity produces results no risk file can use. Section 9 requires both, in that order.

**Table 2. A proposed structure for recording how an induced failure was contained and whether it was observable, so that the result can be carried into a severity and probability estimate under the applicable risk management standard [25] rather than substituting for one.**

| Severity class               | Definition                                                                                                                                                     | What the benchmark must record                                                                                                                                                         |
|------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| S0, absorbed                 | The induced model failure is contained by an existing safety interlock or constraint outside the model, and no change in delivered therapy results.            | That the interlock, not the model, provided the containment, and which interlock it was. A robustness figure that silently credits the model for an interlock's work is misattributed. |
| S1, transient                | Delivered therapy deviates, but the deviation is self-correcting within the loop's own response and does not leave the accepted control range.                 | Magnitude and duration of deviation, and whether correction depended on patient action.                                                                                                |
| S2, sustained and detectable | Delivered therapy deviates persistently in a direction that leaves the accepted range, and the deviation is observable to the patient or to a monitoring path. | Direction, magnitude, time to leave range, and which monitoring path surfaced it.                                                                                                      |
| S3, sustained and silent     | Delivered therapy deviates persistently and the deviation is not surfaced by any monitoring path in the declared configuration.                                | Direction, magnitude, time to consequence, and the reason no path surfaced it. This class is where a robustness figure and a harm figure diverge most sharply.                         |

The classes are structural; instantiating them for a given device requires clinical input that is outside this paper's scope, and no failure is assigned to any class here because no adversarial test was run.

Two features of that structure matter. First, class S3 is not the class with the largest accuracy change. A silent, sustained, small bias can be more dangerous than a large, obvious, immediately detected failure, and any evaluation that ranks defences by accuracy retained will systematically under-weight it. Second, class S0 exposes a misattribution that current reporting invites. When a system-level interlock absorbs a model failure, the model looks robust and is not. Recording which mechanism did the containing is cheap and prevents an entire category of false confidence.

**Requirement it generates.** Every induced failure must be recorded with its containment mechanism and its observability in the declared configuration, and mapped from there onto a clinical harm severity expressed in the terms the applicable risk management standard already uses [25], so that defences are comparable on harm as well as on retained accuracy.

### 6. Gap 3: No Found Validation of Defences on Constrained Hardware

The work behind this paper found no published instance of a proposed adversarial defence whose robustness was measured on the constrained hardware these devices actually run. An implantable with under 256KB of RAM cannot simply inherit defences benchmarked on GPU clusters.

The prior taxonomy already contains the cost half of this analysis. Heavier defences, including full adversarial training and model ensembles, carry compute overhead that makes them realistically deployable only at the cloud tier. Lighter options, including feature squeezing and detection-based approaches, suit severely constrained implantables because they add minimal latency and memory footprint [1], [7]. That assessment is established and is not re-derived here.

What is missing is the other half of the same table: for each defence at each tier there should be a robustness column measured on that tier, and there is not one. The resulting reporting asymmetry has a predictable direction of error. Cost is assessed from documented resource requirements, which port across platforms reasonably well. Robustness is assessed from published experiments, which do not port at all, because reduced numerical precision alters adversarial robustness [15] and the deployed model is quantised. The constrained tiers therefore carry a well-characterised cost beside a robustness figure imported from a platform they do not resemble, and a manufacturer reading that table would conclude, reasonably and wrongly, that a cheap defence's published robustness number applies to the part they are shipping.

This is why the specification in Section 9 treats hardware tier as a stratification axis rather than a caveat. A caveat is a sentence appended to a number. An axis means the number does not exist except at a specified tier. Figure 2 shows the tier envelopes and the tuple that must accompany every robustness figure inside them.

**Requirement: every robustness number is reported inside a declared resource envelope**

*Tier envelopes are the axis a benchmark must stratify along.*

|                               |                                                                                    |
|-------------------------------|------------------------------------------------------------------------------------|
| <b>Tier 1<br/>Implantable</b> | ARM Cortex-M class, under 256 KB RAM, battery-bound, years in situ, no field patch |
| <b>Tier 2<br/>Wearable</b>    | more headroom than an implantable, still no GPU; duty-cycled radio, daily recharge |
| <b>Tier 3<br/>Gateway</b>     | companion phone or bridge; general-purpose CPU, intermittent connectivity          |
| <b>Tier 4<br/>Cloud</b>       | effectively unbounded compute; the only tier where heavy defences are deployable   |

**for every (defence, attack, tier) cell, publish this tuple**

|                           |                                        |
|---------------------------|----------------------------------------|
| <b>clean performance</b>  | <b>robust performance, fixed suite</b> |
| to be measured            | to be measured                         |
| <b>peak RAM (KB)</b>      | <b>flash footprint (KB)</b>            |
| to be measured            | to be measured                         |
| <b>added latency (ms)</b> | <b>energy per inference</b>            |
| to be measured            | to be measured                         |
| <b>fits the envelope?</b> |                                        |
| to be measured            |                                        |

*The report slots are empty by design: they are fields to be measured, and none was measured here. The cost asymmetry the tiering encodes is documented in the author's prior taxonomy. Missing is the paired robustness column, at that tier, on that silicon.*

**Figure 2. Hardware tier envelopes as a benchmark stratification axis, with the report tuple required for every cell. The report slots are empty deliberately: they are fields to be measured, and none was measured here.**

The tier definitions and the cost asymmetry they encode are from the author's prior taxonomy [1]; the requirement that a robustness figure be reported inside a declared envelope is this paper's.

There is a second-order consequence worth naming. Because the deployed artefact differs from the research checkpoint in precision, size and sometimes architecture, robustness must be measured after the compression step, not before it and inherited. That single sequencing requirement invalidates most of the way robustness evidence is currently produced and reported for embedded targets, and it is not expensive to satisfy. It requires only that the benchmark test the artefact that ships.

Requirement it generates. Robustness must be measured on the artefact as compressed and loaded at a declared tier, after the deployment transformation rather than before it, and published in the same cell as the resource cost of the defence at that tier, so that a defence's cost and its robustness are established on the same silicon rather than one being documented and the other imported.

## 7. Gap 4: No Testing Standard Aligned to Regulatory Expectation

The targeted search described in Section 3.1 found no regulator-aligned testing standard for adversarial robustness of the kind that exists for functional safety testing, so manufacturers have no consistent bar to test against.

This claim needs care, because a related claim is already published and this paper must not re-make it. The author's prior work on continuous glucose monitoring security mapped threats to specific regulatory requirements and to established control and technique catalogues [4], and the prior taxonomy aligned its defence recommendations with regulatory cybersecurity and AI lifecycle guidance [1]. That is control mapping. Control mapping answers the question of what a manufacturer should have in place. It does not answer how the manufacturer demonstrates that what they have in place works, or what bar the demonstration must clear. Those are the questions a testing standard answers, and they are separate questions. A device can be fully mapped to every applicable control and still ship a controller whose adversarial robustness has never been measured, because no control in any current catalogue specifies an adversarial robustness test method or a threshold.

The regulatory record supports the distinction, and it has recently moved. The current final premarket cybersecurity guidance for medical devices was reissued on 3 February 2026, superseding the version of 27 June 2025 [16]. The reissue is a quality-system realignment rather than a change of technical expectation: references to subsections of the device quality system regulation were replaced by clauses of the quality management system standard that the regulation now incorporates by reference, and the language in the security controls section that pointed at the superseded design-control subsections was removed. The security testing expectations carried across unchanged, and they are the established security-engineering ones: a secure product development framework, threat modelling, a software bill of materials, vulnerability scanning, fuzz testing and penetration testing across the total product lifecycle. Those methods target implementation flaws in software and protocols. They are not adversarial robustness evaluation, and they do not become adversarial robustness evaluation when the device contains a model, because the failure mode is different in kind: the model is behaving exactly as implemented and has been steered by a valid-looking input.

Gap 4's premise therefore survives the reissue, and the reissue is what makes the premise worth restating rather than what undermines it. A premarket cybersecurity guidance was revised in 2026, cybersecurity became a quality system element in the same month, and no adversarial robustness test method or threshold entered the document in the process. Separately, the guidance addressing artificial-intelligence-enabled device software functions across the lifecycle remains a draft as of the date of writing [17], and its focus is lifecycle management and performance monitoring rather than a prescribed adversarial robustness test method. That status should be re-checked by any reader relying on this section, because a draft can finalise and this argument is time-stamped.

Four further documents have to be addressed, because they are what a reader reaches for before accepting a negative claim about standards. Each is real, each is relevant, and none of them closes this gap, for reasons that are specific rather than dismissive.

Robustness assessment for neural networks. A two-part consensus deliverable on assessment of the robustness of neural networks exists. Part 1 is a technical report whose stated purpose is to provide background about existing methods for assessing the robustness of neural networks [22]. Part 2 is a standard whose stated purpose is to provide methodology for the use of formal methods to assess robustness properties, focused on how to select, apply and manage formal methods to prove such properties [23]. This is the closest existing artefact to what Gap 4 asks for, and the claim has to be narrowed against it rather than routed around it. Three things separate them. Both deliverables are domain-general: not written for, scoped to, or recognised for a medical device or a clinical signal, so they supply no clinically denominated perturbation budget and no patient stratification. Part 2's instrument is formal proof over a network's input space, a different instrument from empirical evaluation under a fixed adaptive attack suite and the one least likely to stay tractable on a quantised artefact at the constrained tiers. And neither states a pass condition for a device, which is the object a submission needs. What they do establish, and this paper concedes it plainly, is that robustness assessment is already a subject of consensus standardisation. What is missing is not the idea of a robustness standard but a device-realistic instantiation of one.

Adversarial machine learning terminology. A government adversarial machine learning taxonomy and terminology reference for attacks and mitigations exists in final form [24], and it is already used in the author's published taxonomy for what it is built for, a common vocabulary [1]. Its own framing is that it is intended to inform other standards and future practice guides; it defines no conformance test and sets no pass threshold. The same holds for MITRE ATLAS, which supplies a shared vocabulary of adversarial tactics and techniques against AI systems and is genuinely useful for that [18]. A vocabulary is not a test standard: it lets two parties name the same attack, not agree on whether a defence passed. That is not a criticism of either document but a statement of which class of artefact they belong to.

Risk management, and security risk management. Medical device risk management is governed by a consensus standard under which harm is characterised and controlled across the device lifecycle [25], and security risk management for devices is specified as a process performed inside that safety process rather than parallel to it [26]. Section 5 states how Table 2 sits relative to [25], and the position is deliberate: this paper adds nothing to the risk management framework and asks for nothing in it to change. What neither standard supplies, because neither is a test-method standard, is the measurement a risk file is waiting for. A security risk management process can require that adversarial manipulation of a model be treated as a hazard, and that a control be identified against it, and still leave entirely open by what method and against what bar that control is shown to work. That open question is Gap 4, stated in the vocabulary of the standards surrounding it.

The contrast with functional safety is instructive, and the mechanism can now be named rather than gestured at. The regulator maintains a list of recognised consensus standards and modifies it periodically, most recently in February 2026, expressly to assist manufacturers who elect to declare conformity with a consensus standard to meet certain requirements for medical devices [21]. That mechanism gives a functional safety claim a named standard, a defined method and a defined pass condition, and makes a declaration of conformity a reviewable object rather than an assertion. The device security risk management standard above has been recognised through it [26]. No adversarial robustness test method has. So the position, stated to this author's knowledge as of the date of writing and on the basis of the check described in Section 3.1, is narrower and firmer than an unqualified negative: robustness assessment standards for neural networks exist and are not device standards; a device security risk management standard exists and is recognised, and is not a test method; and nothing found on the recognised list specifies an adversarial robustness test method or threshold for a machine learning component in a medical device.

The absence has an under-appreciated cost. Without a bar, the rational strategy for a manufacturer is to say as little as possible about adversarial robustness, because a specific claim invites scrutiny it cannot resolve while silence does not. A standard converts adversarial robustness from a reputational liability into a compliance item, and that is the change which makes manufacturers willing to measure and publish. It is also how a benchmark, a technical artefact, becomes a regulatory one.

Requirement it generates. The benchmark must be specified as a referenceable test plan with a versioned protocol and defined reporting, so that a premarket submission can cite a protocol version and a regulator can read the result against a fixed method rather than a self-designed one.

## **8. Gap 5: Almost No Time-Series Benchmarks, and None for the Closed Loop**

The work behind this paper found one adversarial robustness benchmark for a physiological time-series modality [14], and none for continuous glucose readings or electrocardiogram traces in a closed-loop control setting. Set against the image domain, where a shared versioned benchmark has been maintained for years [9], that is the imbalance in a sentence, and it matters because a field's tooling and intuition follow its benchmarks. The imbalance is not neutral, because tooling encodes assumptions and the assumptions in image-domain adversarial tooling are wrong for physiological streams in four respects. Inputs are treated as independent rather than as windows over a continuing signal, so an attack persisting across windows is not naturally expressible. Budgets are pixel norms, with no clinical interpretation and no physiological constraint. There is no rate-of-change limit, so nothing prevents a nominally in-budget perturbation from being physically impossible. And evaluation is per-sample, which obscures whether a defence fails on particular patients rather than uniformly. Foundational work established that time-series classifiers need their own attack treatment [13], and the clinical case was demonstrated on electrocardiogram models where the perturbations needed to change a diagnostic output were small [12]. What follows is not only that the attacks exist but that the evaluation apparatus needs rebuilding for the modality.

The one benchmark found is also the strongest argument that this paper's ask is achievable rather than aspirational, and it belongs in the foreground. An adversarial robustness benchmark has been constructed for electroencephalogram-based brain-computer interfaces, covering a physiological time-series modality with cross-subject evaluation [14]. That work settles the feasibility question: a physiological signal domain can support a standardised robustness benchmark with subject-level evaluation, and someone has already built one. The remaining question is not whether it can be done but why it has not been done for the modality where the model actuates rather than only classifies, and what such a benchmark would have to add.

The closed-loop case adds a requirement the classification benchmarks do not have. In a classification setting, a successful attack changes a label. In a closed-loop setting, it changes a therapy delivered over time, and the loop responds to its own output. An adversarial evaluation that stops at the model boundary will therefore miss the loop dynamics that determine whether a perturbation becomes a clinical event. The benchmark must evaluate the model in the loop, which is why Figure 3 specifies the closed-loop chain as the setting under test rather than the controller alone.

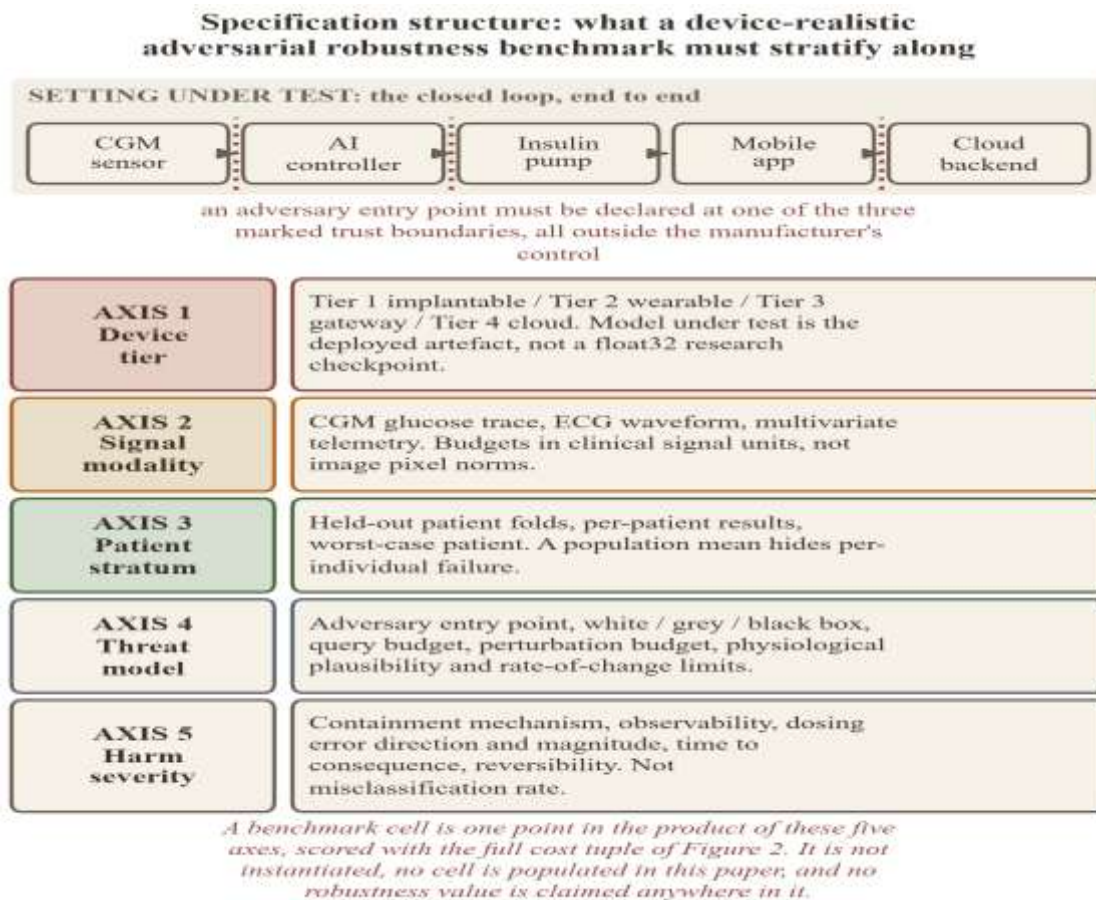
Requirement it generates. Perturbation budgets must be expressed in clinical signal units with physiological plausibility and rate-of-change constraints, and the evaluation must span the closed loop rather than terminate at the model's output.

## 9. Requirements for a Device-Realistic Benchmark

This section is the paper's contribution. It states what a device-realistic, hardware-constrained adversarial robustness benchmark for AI-enabled medical devices would have to contain, and why each element is not optional. It is a specification. It is not instantiated, no cell of it is populated, and no robustness value appears anywhere in it or in this paper.

### 9.1 Structure: five stratification axes, seven requirements

Two structures are in play and the relationship between them should be stated before either is used. Five axes define what a benchmark cell is: a cell is one point in the product of device tier, signal modality, patient stratum, threat model and clinical harm severity. Seven requirements define what the benchmark must do, and they do not stand in one-to-one correspondence with the axes. R1 to R4 and R6 specify how the five axes are populated and reported. R5 specifies what accompanies every cell regardless of where it sits. R7 specifies who has to write the thing for it to be used. Figure 3 shows the axes and the setting they apply to; the requirements follow it.



**Figure 3.** The proposed specification structure. The upper band is the setting under test, the closed-loop chain with the three trust boundaries outside the manufacturer's control, taken from the prior taxonomy [1]. The five axes below are this paper's specification. A benchmark cell is one point in their product, scored with the full cost tuple of Figure 2. No cell is populated here.

**R1.** Device tier, and the artefact under test. Every result is reported at a declared tier: severely constrained implantable, moderately constrained wearable, controller or gateway, or cloud. The model evaluated must be the deployed artefact, quantised and loaded for that tier, not a full-precision research checkpoint. This is R1 rather than a footnote because reduced precision changes the property being measured [15], so a pre-deployment number is not a conservative estimate of the deployed one. It is a number about a different model.

Because most researchers will never hold a production device, R1 stands or falls on what counts as an acceptable substitute, and Section 4's division-of-labour argument rests on that term being defined rather than invoked. A surrogate is faithful here when four conditions are declared and met: same model architecture and parameter count as the target; same numerical representation from the same quantisation procedure, verified by bit-exact agreement of the quantised weights; peak RAM, flash footprint and added latency within a declared tolerance of the target's, measured on the surrogate and stated in the cell; and clean performance on the pinned split within a declared tolerance of the target's. The tolerances are part of the declaration rather than fixed here, because they are what a manufacturer and an external group would have to negotiate first. A surrogate failing any of the four is reported as a distinct tier entry rather than a stand-in, which keeps the failure visible instead of averaging it into the target's row.

**R2.** Signal modality, with clinically denominated budgets. Perturbation budgets are expressed in the signal's own clinical units, with physiological plausibility constraints and rate-of-change limits, rather than as pixel-space norms carried over from image work [13]. The threat model additionally declares an adversary entry point at a named trust boundary, sensor-to-controller, controller-to-pump, or device-to-cloud, together with the access class assumed at that point, white, grey or black box, and a query budget. Without a declared entry point and query budget, a robustness figure describes an adversary nobody has specified.

**R3.** Patient stratification. Results are reported per patient and at the worst-case patient, defined as the minimum over the per-patient results with the cohort size stated alongside it, never only as a cohort mean, using patient-disjoint splits fixed and published before evaluation. Independent precedent for stratifying by subject rather than pooling exists in the adversarial literature itself: the one adversarial robustness benchmark found for a physiological time-series modality evaluates across subjects [14], so subject-level reporting is already established practice where such a benchmark has been built. The closest supporting precedent available to the author is by analogy rather than by direct evidence, and Section 10 gives it: in developing an on-device anomaly detector, a single global detection threshold performed inconsistently across patients because glucose-insulin dynamics vary meaningfully by individual, which forced a move to patient-based cross-validation and per-patient hyperparameter tuning [3]. That was an anomaly-detection and generalisation finding rather than an adversarial one, and it is presented as such. Its implication for benchmark design is nonetheless direct and does not depend on it being adversarial: if a population-level average can hide per-individual failure for a benign generalisation problem on this signal, it can hide per-individual failure for an adversarial one, and the reporting convention that permits the hiding is the same convention. A cohort mean is the wrong summary statistic for a device that is worn by one person at a time.

**R4.** Threat model, versioned and externally fixed. The attack suite is standardised, adaptive, parameter-free where possible, and fixed by the benchmark rather than chosen by the party reporting the result. The rationale is the evaluation-integrity record in Section 3: self-designed evaluations have repeatedly overstated robustness, at scale and by wide margins [5], [8], [10], and the image domain's remedy was a shared versioned benchmark rather than better individual practice [9]. Adopting a fixed suite is the single highest-leverage requirement in this list, because it is the one that makes every other number comparable.

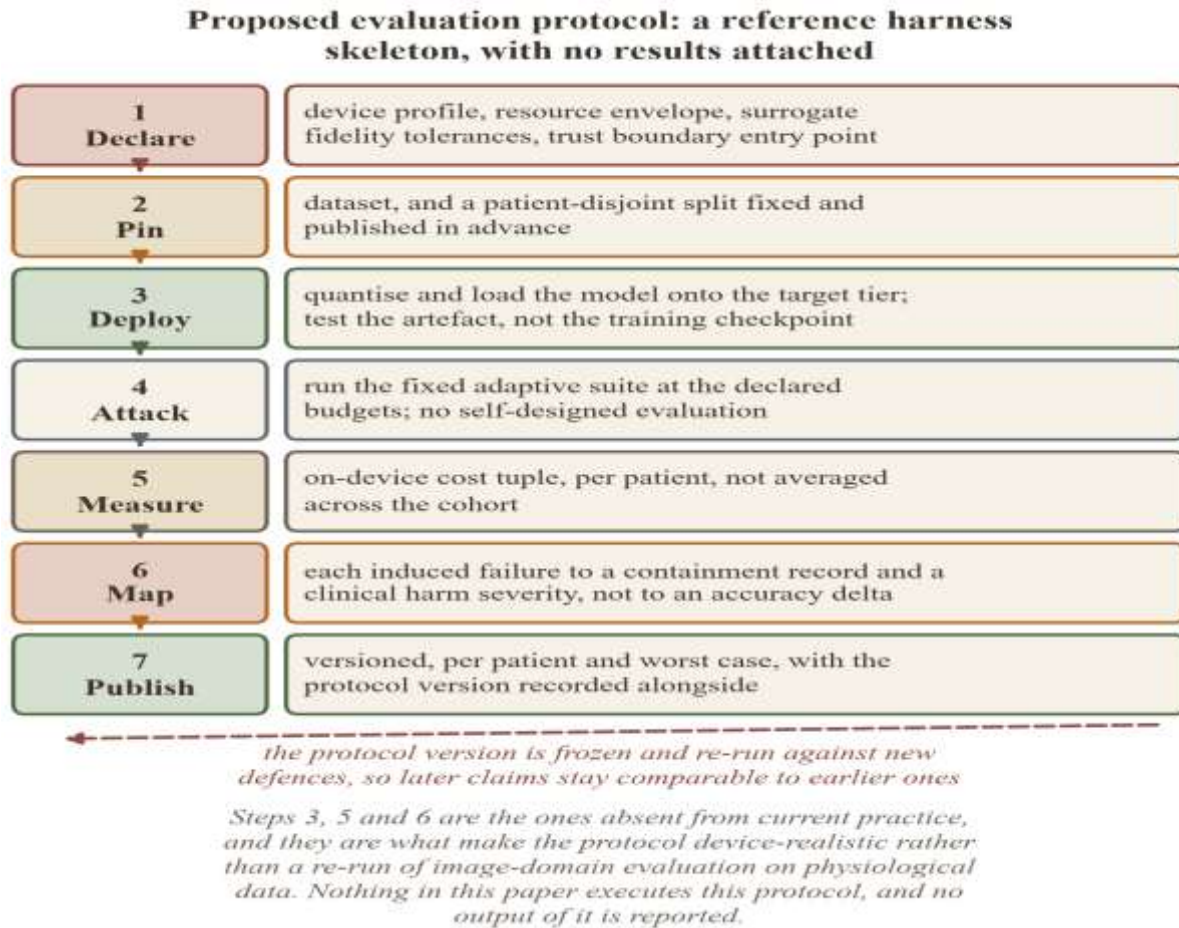
**R5.** Cost reported alongside robustness. Every cell publishes the full tuple of Figure 2: clean performance, robust performance under the fixed suite, peak RAM, flash footprint, added inference latency, energy per inference, and an explicit determination of whether the defence fits the declared envelope. A robustness figure without a cost figure is not actionable at the constrained tiers, because the deployment question there is not which defence is strongest but which defences fit at all. Reporting the pair converts the tiering from a recommendation into a measurement.

**R6.** Clinical harm severity mapping. Each induced failure is recorded with its containment mechanism and its observability using a structure such as Table 2, so that a defence credited with robustness is not in fact being credited for an interlock's work, and is then mapped onto a severity expressed in the terms the applicable device risk management standard already uses [25] rather than onto a scale invented for the benchmark. Defences are compared on harm as well as on retained accuracy, because the two orderings differ and the harm ordering is the decision-relevant one.

**R7.** Joint governance and regulatory referenceability. The specification is developed jointly by researchers, manufacturers and the regulator, published as a versioned test plan, and structured so that a premarket submission can cite a protocol version. This requirement is organisational rather than technical, and it is the one most likely to be treated as optional. It is not. A benchmark written by researchers alone will assume device access that manufacturers will not grant. One written by manufacturers alone will not be trusted. One written without regulatory involvement will not be cited in a submission, and a benchmark nobody cites in a submission does not change what gets built. Section 7's argument applies here directly: the specification must be distinguishable from, and complementary to, existing control mapping and the existing robustness assessment and risk management standards [4], [16], [18], [22], [23], [26].

## 9.2 Protocol: a reference harness skeleton

The seven requirements imply an ordering, because several of them constrain steps that must happen before others. Figure 4 states that ordering as a protocol.



**Figure 4. The proposed evaluation protocol. Steps 3, 5 and 6, deploying before attacking, measuring on-device cost per patient, and mapping failures to clinical harm, are the steps absent from current practice and the ones that make the protocol device-realistic. Nothing in this paper executes it.**

Three ordering constraints in that flow are load-bearing. Deployment precedes attack, because the artefact under test must be the one that ships (R1). The patient-disjoint split is pinned and published before evaluation, because a split chosen after seeing results is not a held-out split (R3). And the protocol version is frozen and re-run against later defences, because a benchmark whose method drifts produces numbers that cannot be compared across years, which defeats its purpose (R4, R7).

## 9.3 Which requirement closes which gap

Table 3 maps the requirements back to the gaps, so that the specification can be audited against its motivation rather than taken on trust.

| Requirement                                                                                            | Gap it closes                                                                                                                                                   | How it is evaluated                                                                                                                                                                   |
|--------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| R1. Declared tier and declared faithful surrogate; test the deployed, quantised artefact               | Gap 1, no found empirical testing on real devices, which the surrogate declaration is what makes tractable; Gap 3, defences unvalidated on constrained hardware | Device profile, resource envelope and surrogate fidelity tolerances declared in advance; model loaded at target tier before any attack is run; envelope conformance reported per cell |
| R2. Clinically denominated, physiologically constrained budgets; declared entry point and query budget | Gap 5, rarity of time-series benchmarks; contributes to Gap 2                                                                                                   | Budget stated in signal units with rate-of-change limits; adversary entry point named at a trust                                                                                      |

|                                                                                        |                                                                                                     |                                                                                                                                                                                                          |
|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                        |                                                                                                     | boundary; access class and query budget fixed per cell                                                                                                                                                   |
| R3. Per-patient and worst-case reporting on pinned patient-disjoint splits             | Gap 2, robustness disconnected from harm at the individual level; contributes to Gap 1              | Split published before evaluation; per-patient results and worst-case patient reported; cohort mean permitted only as a secondary figure                                                                 |
| R4. Fixed, adaptive, externally specified attack suite                                 | Gap 4, absence of a bar to test against; contributes to Gap 1 by making independent runs comparable | Suite and its version supplied by the benchmark, not the reporting party; protocol version recorded with every result                                                                                    |
| R5. Full cost tuple published with every robustness figure                             | Gap 3, unvalidated cost-robustness pairing on constrained tiers                                     | Peak RAM, flash, added latency, energy per inference measured on the target tier; explicit fit determination against the envelope                                                                        |
| R6. Containment and observability recorded, then mapped to clinical harm severity [25] | Gap 2, robustness not connected to clinical harm                                                    | Each induced failure recorded with its containing mechanism and whether any monitoring path surfaced it, then mapped to a severity in risk-management terms; defences ranked on harm as well as accuracy |
| R7. Joint governance; versioned, citable test plan                                     | Gap 4, no regulator-aligned testing standard                                                        | Specification co-developed and published with version identifiers; premarket submissions cite a protocol version rather than a bespoke method                                                            |

**Table 3.** Benchmark requirement against the gap it closes and the means of evaluation. The primary attribution in each row is the gap whose section generates that requirement, so the table can be read against Sections 4 through 8 without reconciliation. Secondary contributions are stated where they exist, because the mapping is many-to-many: no single requirement closes a gap alone, which is the reason the specification has to be adopted as a set rather than piecemeal.

#### 9.4 What the specification deliberately does not do

Three exclusions are worth stating, because a specification is defined as much by its limits as by its contents. It does not propose new defences. The candidate defence space for this device class is already catalogued [1], and the argument of this paper is that the catalogue's problem is unmeasured entries rather than missing ones. It does not propose new attacks. The attack taxonomy is likewise established [1], and the requirement in R4 is that a suite be fixed and externally specified, not that it be novel. Standardisation, not novelty, is the contribution being asked for.

It does not map controls to regulatory requirements. That work exists [4], and repeating it would confuse a control map with a testing standard, which is precisely the confusion Section 7 argues against.

#### 9.5 A minimum viable v0

A requirements list nobody can start from is advocacy. This subsection therefore proposes one concrete configuration, enough to fill a single cell, so the specification can be argued with at the level of its parameters rather than only its principles. No result is attached to any of it. Nothing in this configuration was run, no attack was executed, and no number here is a measurement. The values are first cuts chosen to be criticised and replaced; the point of stating them is that a first cut can be improved and an unspecified requirement cannot. Dataset and cohort. The OhioT1DM dataset, which contains eight weeks of continuous glucose monitoring, insulin, physiological sensor and self-reported life-event data for each of twelve people with type 1 diabetes, with a glucose value recorded every five minutes [27]. It is proposed as the v0 corpus for three reasons: it is publicly available for research, it is real patient telemetry rather than simulator output, and six of its twelve

contributors are the cohort behind the author's own prior on-device work [3], so a v0 built on it inherits a documented baseline of ordinary, non-adversarial behaviour. Patient-disjoint folds over all twelve contributors, pinned and published before any attack is run.

Tier and artefact. Tier 2, the moderately constrained wearable, on an ARM Cortex-M class part, declared with its RAM, flash and clock figures in the cell. The artefact under test is the model after post-training integer quantisation of weights and activations, loaded on the target part or on a surrogate meeting the four fidelity conditions in R1. Tier 2 rather than Tier 1 is proposed for v0 because the grounding device class is a wearable pump and monitor, and because a v0 whose first cell is the hardest cell will not be filled.

Perturbation budget, in clinical units. The budget is expressed on the glucose signal in mg/dL, not as a pixel-space norm, and it is constrained twice. The rate constraint comes first: perturbations may not imply a rate of change above 3 mg/dL per minute, the boundary at which the trend-arrow conventions in clinical continuous glucose monitoring practice report change as rapid [28]. The per-sample constraint follows, because a five-minute sampling interval [27] and a 3 mg/dL per minute ceiling together bound a single sample's admissible displacement at 15 mg/dL. The v0 budget is therefore an infinity-norm bound of 15 mg/dL per sample with the rate ceiling enforced across the whole attack window, which excludes by construction the instantaneous jumps and impossible trajectories a pixel-norm budget admits. Both numbers are proposals: the rate ceiling is the one with external grounding, the per-sample figure is derived from it, and a benchmark committee may well conclude that a tighter sustained bias is the more dangerous adversary and set a smaller number over a longer window.

Entry point and access class. The sensor-to-controller link as the declared entry point, black-box access, and a query budget declared per cell. That combination is proposed for v0 because it is the weakest set of adversary assumptions under which the documented radio-layer reachability of these devices [19] still permits an attack, which makes it the assumption a manufacturer is least able to dismiss.

Attack set. Assembled from techniques already established in the literature this paper cites, not newly invented, because R4 asks for standardisation rather than novelty. Three white-box attacks re-parameterised to the clinical budget above: projected gradient descent within the constrained norm ball [6], the Carlini and Wagner construction [5], and the time-series adaptations shown to be necessary for sequence classifiers rather than reusable from the image domain [13]. One query-limited black-box attack, matching the declared access class, of the kind carried in the parameter-free ensemble that exposed overstated robustness at scale [8]. The suite is versioned and supplied by the benchmark, and the reported figure is the worst case across the suite rather than the mean, following the practice that made the image-domain benchmark informative [8], [9].

Report. Exactly the tuple of Figure 2, at the declared tier, per patient and at the worst-case patient, with each induced failure recorded against Table 2 and carried into a severity estimate in risk management terms [25].

That is one cell. Filling it would produce the first adversarial robustness figure for a closed-loop medical device controller that is reported at a declared tier, in clinical units, per patient, against an externally fixed suite, with its resource cost attached. This paper has not filled it, and does not estimate what filling it would show.

## 10. A Worked Illustration of Honest On-Device Evaluation

The requirements in Section 9 are easier to accept if there is an example of what disciplined on-device evaluation looks like in this domain, including what it declines to claim. The author's separately published anomaly detector serves that purpose [3]. It must be read with one boundary held firmly in place, and the boundary is stated before the numbers rather than after them.

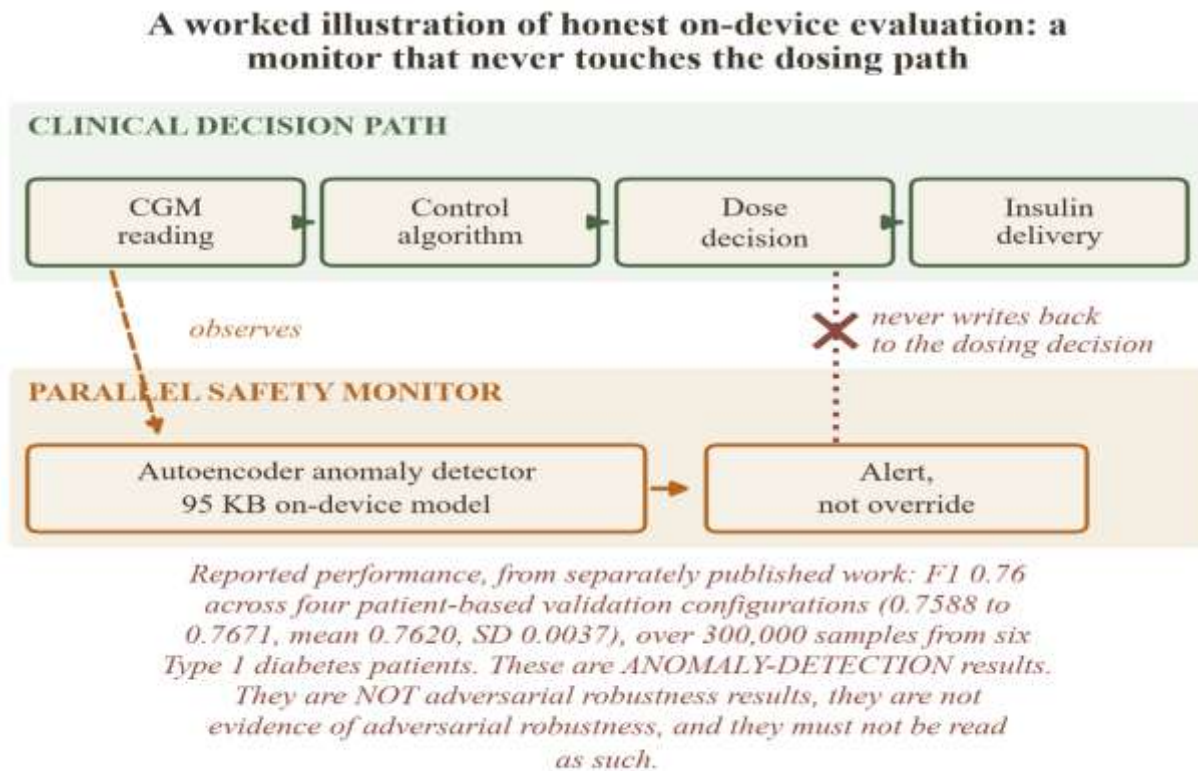
This section reports anomaly-detection performance. It does not report adversarial robustness, it is not evidence of adversarial robustness, and it must not be read as such. The detector was validated against device malfunction and anomaly patterns, not against adversarially crafted inputs. No attack was run against it. The distinction is not a formality: an anomaly detector's ability to flag naturally occurring irregularities says nothing about its behaviour under inputs designed specifically to evade it, and conflating the two is one of the more common ways adversarial claims get manufactured out of non-adversarial evidence.

With that established, three properties of the work illustrate requirements from Section 9.

On-device footprint, illustrating R1 and R5. The detector is a 95KB on-device autoencoder, evaluated on anomaly detection rather than under attack, that runs with negligible added latency. Those figures are reported for the artefact as deployed, not for a training-time checkpoint, and they are reported in the units R5 asks for: memory footprint and added latency, alongside detection performance rather than instead of it. What they constitute is a partially populated cost column, and the shortfall is worth naming precisely, because this is the one place where the author's own prior reporting can be held against the standard proposed two sections earlier. R5 asks for seven fields. This supplies model size, which is not the same quantity as peak RAM, and a latency claim of negligible, which is a characterisation rather than a value in milliseconds. Flash footprint, energy per inference and an explicit envelope-fit determination are absent, and robust performance is absent because no attack was run. Two of seven fields, one of them approximate. That is more than most of the

literature reports for a model it calls deployable, and it is still not what R5 requires: the standard proposed here is one the author's own published work does not yet meet.

Architectural placement, illustrating R6. The detector runs as a parallel safety monitor, in the anomaly-detection role for which it was built and evaluated. It observes the clinical path and raises an alert; it never writes back to the dosing decision. Figure 5 shows the placement. This matters for harm-severity reasoning in both directions. It bounds the harm the monitor itself can cause, since a false positive produces an alert rather than a dose change. It also means the monitor cannot be credited with containing a failure it merely observed, which is the misattribution R6 and severity class S0 in Table 2 are designed to catch. Nothing about the placement bears on how the detector would behave under adversarial input, because that was never tested.



**Figure 5. The detector's placement outside the dosing path. The layout was developed by the author for unpublished internal work and is used here for the first time in print. The performance figures shown are anomaly-detection results from separately published work [3]. They are not adversarial robustness results and are not evidence of adversarial robustness.**

Patient-based validation, illustrating R3. This is the property with the most direct bearing on the specification. During development, a single global detection threshold performed inconsistently across patients, because glucose-insulin dynamics vary meaningfully by individual. The response was to move to patient-based cross-validation and per-patient hyperparameter tuning rather than a single global configuration. The resulting figures, which are anomaly-detection scores and not robustness scores, were per-patient F1 between 0.7588 and 0.7671 across four patient-based validation configurations, with a mean of 0.7620 and a standard deviation of 0.0037, over more than 300,000 samples from six Type 1 diabetes patients in the public OhioT1DM continuous glucose monitoring dataset [27]. The overall anomaly-detection figure is an F1 of 0.76 [3]. No adversarial input was involved in producing any of those numbers.

Three observations follow. First, to repeat, this is a generalisation finding, not an adversarial one: the inconsistency arose from ordinary inter-patient variation, not from an attack. Second, the tight final spread is a consequence of the per-patient tuning, not a reason to skip it. A standard deviation of 0.0037 across configurations describes performance after the design was changed to accommodate patient heterogeneity; it is not evidence that heterogeneity did not matter, and reading it that way inverts the finding.

Third, and this is the argument the specification rests on: the reporting convention that would have concealed the original problem is the convention the adversarial literature uses by default. Had this work reported only a cohort aggregate under a single global threshold, the per-patient inconsistency would have been averaged into a respectable number and the design change would not have happened. The problem was found because the

evaluation was stratified by patient. An adversarial evaluation reporting only cohort means is exposed to exactly that failure, on the same signal, with a worse consequence, because the per-patient variation an attacker would exploit is the variation the average removes. R3 follows directly, and it is the requirement in Section 9 with the closest supporting precedent, by analogy. That precedent is not adversarial evidence and does not become adversarial evidence by being persuasive: what transfers is an argument about a reporting convention, not a measurement.

## **11. Limitations**

The limitations of this paper are substantial and specific, and stating them plainly is a condition of the argument rather than a concession at the end of it.

This is a position and research-agenda paper, not an empirical study. Its contribution is a specification and the argument for it. It contains no experiment, no measurement of any defence, and no comparative result.

No adversarial testing was performed. The author holds no attack success rates, no hardened-versus-unhardened comparisons, and no robustness test outputs, because the prior work this paper builds on was literature-based systematic analysis rather than empirical adversarial evaluation [1]. Nothing in this paper simulates, estimates or illustrates such a measurement, and the figures are constructed so that the places where a measurement would sit are visibly empty.

The gap claims are search-limited, and the search was not systematic. Section 3.1 states this in full rather than in summary: no review protocol was registered, no screening was performed, and no records-screened count exists for the four gaps that rest on prior literature work. Every gap is therefore worded as a claim about what was not found. Any one of them can be refuted by a single published counter-example, and a reader who holds one should treat the corresponding gap as closed and the specification as unaffected, since the specification's value rests on the artefact not existing yet rather than on any gap being total.

The benchmark is specified but not instantiated or validated, including the v0. No cell of it has been populated, no reference harness has been executed, and the specification's own usefulness is therefore argued rather than demonstrated. The minimum viable configuration in Section 9.5 is a proposal with no results attached: nothing in it was run, the perturbation budget and rate ceiling are first cuts rather than validated parameters, and the derived 15 mg/dL per-sample figure is an arithmetic consequence of a cited rate ceiling and a sampling interval, not a measured threshold of anything. It is entirely possible that instantiating the specification would reveal requirements that are impractical as stated, or that the seven requirements are not the right seven. A specification's first contact with implementation usually changes it.

The analysis is grounded in the artificial pancreas domain, and generalisation is argued rather than demonstrated. The closed-loop insulin delivery setting was chosen because the author's prior work is grounded there [1], [2], [3], [4] and because the domain is concrete enough to specify against. Whether the five axes transfer cleanly to implantable cardiac devices, to imaging systems, or to diagnostic software as a medical device is a claim this paper does not test. The tiering and trust-boundary structure appear likely to transfer; the clinical harm severity classes in Table 2 are shaped by the specific pharmacodynamics of insulin delivery and would need rebuilding for another therapy.

The only empirical figures in this paper are anomaly-detection results from prior published work, not adversarial robustness results. The 95KB footprint, the per-patient F1 range of 0.7588 to 0.7671, the mean of 0.7620, the standard deviation of 0.0037, the sample count above 300,000, and the six-patient cohort are all from separately published anomaly-detection work [3]. They characterise detection of naturally occurring anomalies. They say nothing about behaviour under adversarial input, and this paper makes no adversarial claim on their basis.

Two further boundaries. No security review board, clinical committee or regulator reviewed the prior work this paper builds on, and no regulatory submission has been made using any of it. The regulatory argument in Section 7 is an argument about the current state of published guidance and recognised standards, made from the outside. Separately, that argument is time-stamped and perishable. It rests on the premarket cybersecurity guidance as reissued in February 2026 [16], on the draft status of the guidance addressing artificial-intelligence-enabled device software functions [17], and on the recognised standards list as most recently modified at the time of writing [21]. Any of the three can move. A reader with access to a more current guidance record or recognition list should check the claim rather than accept it, and the specification in Section 9 survives the claim being narrowed, because a benchmark is still worth having in a world where a testing standard is on its way.

## **12. Conclusion**

The adversarial machine learning literature has told the medical device field what can go wrong. The author's own prior work contributed to that effort, cataloguing the attacks and candidate defences for AI controllers in

this device class, mapping them to established technique and terminology vocabularies, and tiering the defences by device resource profile [1], [2], [4]. That work is done and this paper does not repeat it.

What remains undone is the harder half. A manufacturer building an insulin pump controller today cannot find out whether a given defence survives on the microcontroller they are shipping, against an adversary anyone has specified, for the patients at the tails of their population, with a consequence anyone has classified clinically. Not because the answer is discouraging, but because the question has no standard form in which to be asked. Five structural gaps sustain that condition, each stated as what the work behind this paper failed to find rather than as what does not exist: no empirical adversarial testing on real devices or declared surrogates, no link from robustness to clinical harm for a specific device, no robustness validation inside the memory envelopes that actually apply, no regulator-aligned testing standard, and one time-series adversarial benchmark for a physiological modality with none for the closed loop.

The five reduce to one missing artefact. This paper has specified it: a benchmark stratified along device tier, signal modality, patient stratum, threat model and clinical harm severity; publishing memory, latency and energy cost alongside every robustness figure; reporting per patient and at the worst case rather than as a cohort mean; running a fixed external attack suite against the deployed quantised artefact rather than a research checkpoint; recording containment and observability so that harm can be estimated in the terms the risk management standard already uses [25]; and versioned so that a premarket submission can cite it. Section 9.5 proposes a first configuration, down to a named public dataset, a device tier, a perturbation budget in mg/dL and a rate ceiling, so that the next step is a build rather than another argument about principles.

The requirement that will decide whether this happens is the least technical one. Researchers, manufacturers and regulators currently operate with different working definitions of what robust means for a device like this, and each definition is coherent inside its own institution. A benchmark is simply the artefact that forces those three definitions into one. It has to be built jointly or it will not be used, and until it is built, adversarial robustness for AI-enabled medical devices will remain a property that is discussed, catalogued, and never measured where it matters.

## References

- [1] V. S. A. Piratla, "Adversarial machine learning threats to medical device AI controllers," *International Journal of Innovative Research and Creative Technology*, vol. 12, no. 2, 2026, doi: 10.62970/ijirct.v12.i2.2604012.
- [2] V. S. A. Piratla, S. Saxena, S. Bhatia, and N. Kumar, "Safeguarding the artificial pancreas: A review of security and reliability gaps and AI driven resilience," in *2025 5th Intelligent Cybersecurity Conference (ICSC)*, 2025, pp. 399-411, doi: 10.1109/icsc65596.2025.11140338.
- [3] V. S. A. Piratla, S. Bhatia, and S. Saxena, "AP-GUARD: Artificial pancreas generalized unsupervised anomaly recognition and detection," in *2025 IEEE 17th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2025, pp. 763-771, doi: 10.1109/cicn67655.2025.11368264.
- [4] V. S. A. Piratla, "A security framework for AI-enabled continuous glucose monitoring: Translating FDA guidance into actionable controls," *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, vol. 7, pp. 396-405, 2026, doi: 10.63282/3050-9262.ijaidsm-l-v7i1p163.
- [5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39-57, doi: 10.1109/sp.2017.49.
- [6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.
- [7] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proceedings 2018 Network and Distributed System Security Symposium (NDSS)*, 2018, doi: 10.14722/ndss.2018.23198.
- [8] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, PMLR vol. 119, 2020, pp. 2206-2216.
- [9] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: A standardized adversarial robustness benchmark," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, 2021.
- [10] R. S. Zimmermann, W. Brendel, F. Tramer, and N. Carlini, "Increasing confidence in adversarial robustness evaluations," in *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022, pp. 13174-13189, doi: 10.52202/068431-0957.

- [11] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287-1289, 2019, doi: 10.1126/science.aaw4399.
- [12] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, "Deep learning models for electrocardiograms are susceptible to adversarial attack," *Nature Medicine*, vol. 26, no. 3, pp. 360-363, 2020, doi: 10.1038/s41591-020-0791-x.
- [13] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Adversarial attacks on deep neural networks for time series classification," in 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1-8, doi: 10.1109/ijcnn.2019.8851936.
- [14] L. Meng, X. Jiang, and D. Wu, "Adversarial robustness benchmark for EEG-based brain-computer interfaces," *Future Generation Computer Systems*, vol. 143, pp. 231-247, 2023, doi: 10.1016/j.future.2023.01.028.
- [15] R. Bernhard, P.-A. Moellic, and J.-M. Dutertre, "Impact of low-bitwidth quantization on the adversarial robustness for embedded neural networks," in 2019 International Conference on Cyberworlds (CW), 2019, pp. 308-315, doi: 10.1109/cw.2019.00057.
- [16] U.S. Food and Drug Administration, "Cybersecurity in medical devices: Quality management system considerations and content of premarket submissions," guidance for industry and FDA staff, final guidance, issued 3 February 2026, superseding the guidance issued 27 June 2025. Docket FDA-2021-D-1158. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-management-system-considerations-and-content-premarket> (accessed 11 August 2026).
- [17] U.S. Food and Drug Administration, "Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations," draft guidance for industry and FDA staff, 2025. Availability notice: Federal Register, doc. 2024-31543, 7 January 2025. <https://www.federalregister.gov/documents/2025/01/07/2024-31543/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing> (accessed 11 August 2026).
- [18] MITRE, "MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems," knowledge base of adversary tactics and techniques against AI systems. <https://atlas.mitre.org/> (accessed 9 August 2026).
- [19] Cybersecurity and Infrastructure Security Agency, "ICS medical advisory ICSMA-19-178-01: Medtronic MiniMed 508 and Paradigm series insulin pumps," 27 June 2019, covering CVE-2019-10964. <https://www.cisa.gov/news-events/ics-medical-advisories/icsma-19-178-01> (accessed 9 August 2026).
- [20] U.S. Food and Drug Administration, "Firmware update to address cybersecurity vulnerabilities identified in Abbott's (formerly St. Jude Medical's) implantable cardiac pacemakers: FDA safety communication," 29 August 2017. <https://www.fda.gov/medical-devices/safety-communications/firmware-update-address-cybersecurity-vulnerabilities-identified-abbotts-formerly-st-jude-medicals> (accessed 9 August 2026).
- [21] U.S. Food and Drug Administration, "Food and Drug Administration Modernization Act of 1997: Modifications to the list of recognized standards, recognition list number: 065," notice, Federal Register, vol. 91, p. 7994, doc. 2026-03310, 19 February 2026, docket FDA-2004-N-0451. <https://www.federalregister.gov/documents/2026/02/19/2026-03310/food-and-drug-administration-modernization-act-of-1997-modifications-to-the-list-of-recognized> (accessed 11 August 2026).
- [22] International Organization for Standardization and International Electrotechnical Commission, ISO/IEC TR 24029-1:2021, Artificial intelligence (AI): Assessment of the robustness of neural networks, Part 1: Overview, technical report, edition 1, published 10 March 2021, ISO/IEC JTC 1/SC 42. <https://webstore.iec.ch/en/publication/68715> (accessed 11 August 2026).
- [23] International Organization for Standardization and International Electrotechnical Commission, ISO/IEC 24029-2:2023, Artificial intelligence (AI): Assessment of the robustness of neural networks, Part 2: Methodology for the use of formal methods, edition 1, 23 pp., published 1 August 2023, ISO/IEC JTC 1/SC 42. <https://webstore.iec.ch/en/publication/87683> (accessed 11 August 2026).
- [24] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," National Institute of Standards and Technology, NIST AI 100-2 E2025, March 2025, doi: 10.6028/NIST.AI.100-2e2025.
- [25] Association for the Advancement of Medical Instrumentation, ANSI/AAMI/ISO 14971:2019, Medical devices: Application of risk management to medical devices, 2019, doi: 10.2345/9781570207181.

- [26] Association for the Advancement of Medical Instrumentation, ANSI/AAMI SW96:2023, Standard for medical device security: Security risk management for device manufacturers, 2023, doi: 10.2345/9781570208621.
- [27] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data (KDH 2020), CEUR Workshop Proceedings vol. 2675, 2020, pp. 71-74, urn:nbn:de:0074-2675-4. <https://ceur-ws.org/Vol-2675/paper11.pdf> (accessed 11 August 2026).
- [28] G. Aleppo, L. M. Laffel, A. J. Ahmann, I. B. Hirsch, D. F. Kruger, A. Peters, R. S. Weinstock, and D. R. Harris, "A practical approach to using trend arrows on the Dexcom G5 CGM system for the management of adults with diabetes," *Journal of the Endocrine Society*, vol. 1, no. 12, pp. 1445-1460, 2017, doi: 10.1210/js.2017-00388.