From research to regulated product

Where regulatory debt accumulates in medical-device AI, and which of it can *still be paid off*.

PURPOSE AND SCOPE

This guide is for research teams who have built something that works and now intend to sell it. It sets out what happens at the point where a funded research project becomes a commercial product, which of the decisions taken during research can be corrected later, and which cannot. It is written for AI and machine learning software intended for the UK, EU and US markets, and assumes the product will be regulated as a medical device in at least one of them.

01 THE SHAPE OF THE PROBLEM

Most engineering teams understand technical debt. You take a shortcut, you ship sooner, and you pay interest on the shortcut until somebody refactors it. Regulatory debt accrues the same way, and it accrues fastest during the months when a research project turns into a product development, because that is when decisions are made quickly by people who have not yet been asked to think about a regulatory submission.

There is one difference, and it is why this guide exists. Code can always be refactored; the debt is unpleasant but never permanent. Parts of regulatory debt cannot be repaid at any price. No amount of later diligence produces a design record that was never written at the time, and no budget widens the scope of a consent already given.

02 START WITH WHAT YOU ALREADY HAVE

The good news first. Teams coming out of research consistently underestimate how much of the regulatory evidence base they have already built. If the science was done properly, a good deal of the work is done.

A pre-specified evaluation with a defensible comparator and a sensible sample size is the same discipline a clinical evaluation asks for. The framing changes; the rigour transfers directly. Anything written to survive peer review sits close to verification evidence: methods sections, analysis plans and pre-registrations get reused rather than rewritten, and a literature review done for a paper covers much of the state-of-the-art analysis a clinical evaluation needs.

If you kept provenance records, which site, which scanner, which protocol, which period, you have answered one of the harder questions an assessor asks. Most teams have this and do not realise it counts. The difficulty is rarely that nothing exists. It is that what exists was assembled for a different reader, and nobody has mapped it across.

03 THE RESEARCH MODEL IS A DESIGN INPUT, NOT THE PRODUCT

This is the single most expensive misconception at the handover, and it is easy to hold without noticing.

A model that performs well in a research paper demonstrates that the approach works. It is not the thing you sell. It was optimised against a metric on a curated dataset, run by the person who built it, with failure cases discussed in a limitations section rather than handled in code. A regulated product has to accept input it was not designed for, fail in a defined and detectable way, produce the same output from the same input on a different machine a year later, run inside stated resource limits, and be verifiable against written requirements by an engineer who did not build it. Very few research models do any of these things.

The useful reframing is that the research is a **design input**. It establishes feasibility, informs your requirements, and gives you a performance target to hold the product against. The product itself is a re-implementation under design control. Teams who believe they already have the product, and need only someone to write the paperwork around it, are the teams whose budget and timeline come apart first.

TWO PIPELINES, TWO SETS OF OBLIGATIONS

The device software you place on the market is the **inference pipeline**: preprocessing, model, post-processing, error handling and recovery, and how the output is presented. That is what IEC 62304 governs and what the technical file describes.

The **training pipeline** is not part of the released device, which leads people to assume it is out of scope. It is not. It is the tooling that produced a design output, so it sits inside design and development controls, and under ISO 13485 clause 4.1.6 on validation of software used in the quality system. Data selection, labelling,

> splits, augmentation, parameter choices and the training run itself have to be reproducible by somebody other than the person who ran them. Re-running the notebook and getting a different model is an audit finding.

04 TRAPS AT THE HANDOVER

1. **The training pipeline is treated as private working.** Notebooks, ad hoc preprocessing and undocumented parameter sweeps form part of the design record. A notebook that has been overwritten cannot be brought under version control after the event, and the model it produced cannot be regenerated.
2. **Intended purpose drifts between documents.** A paper demonstrates feasibility. A website describes a benefit. A grant application projects an outcome. Each was written for a different reader and each is reasonable on its own terms, but only one form of words can be the regulatory intended purpose, and it determines classification, evidence burden and what you are permitted to claim. Setting them side by side takes an afternoon and usually clarifies more than it exposes.
3. **Retraining is a design change.** A model updated on new data is a modified device. The FDA expects this to be planned for rather than handled reactively, through a predetermined change control plan described in the marketing submission. That plan has to exist before authorisation, not after the first retraining run.
4. **Endpoints chosen after seeing the data.** Familiar scientific practice, and fatal to a pivotal study. Where the analysis plan post-dates the data, the result is exploratory, whatever the p-value says.
5. **Research use does not extend to deployment.** Software running under an approved investigation protocol sits differently from software an NHS trust has started to rely on. That transition often happens without a decision being taken, as a pilot becomes routine practice, and the obligations change at that moment whether or not anyone marked it.
6. **Class I is easier to reach than to keep.** Self-declaring Class I avoids a notified body, which is why it is attractive and why it is often the first assumption. Under MDR Rule 11 it is hard to sustain. Software providing information used to make diagnostic or therapeutic decisions is Class IIa at minimum, rising to IIb where a wrong decision could cause serious deterioration and III where it could be fatal. Very little clinical decision support stays at Class I. The classification is worth testing early rather than defending late, because it fixes the conformity route, the evidence burden and roughly two years of budget.

05 THE STANDARDS THAT WERE WRITTEN FOR MACHINE LEARNING

IEC 62304 predates modern machine learning and says nothing about training data or model behaviour. Guidance has since caught up, and it is worth knowing what exists before an assessor asks which of it you considered.

BS/AAMI 34971:2023	Applies ISO 14971 to machine learning. Covers hazards that ISO 14971 alone does not reach: data bias, distribution shift, silent degradation, and overtrust in automated output. Published as AAMI TIR34971:2023 in the US.
IEC 62366-1	Usability engineering. For an AI device the central question is how the output is presented and whether the interface encourages a clinician to check the underlying data or to accept the answer.
BS 30440:2023	UK validation framework for AI in healthcare, written as auditable clauses. Increasingly cited in NHS procurement, so it can matter commercially before it matters regulatorily.
ISO/IEC 42001	AI management systems. Relevant where the AI Act applies and the quality system needs extending rather than replacing.
GMLP principles	Ten guiding principles on good machine learning practice, issued jointly by the FDA, Health Canada and the MHRA. Not binding, and widely used as a reference point by all three.

Over-reliance is a design problem, not a labelling one

A clinician who accepts an incorrect output because the software produced it has made a use error, and it is one your design either encourages or discourages. The EU AI Act names this directly: Article 14(4)(b) requires the system to be provided so that the people overseeing it remain aware of the tendency to over-rely on its output. Deskilling over months of routine use is the same hazard on a longer timescale, and it belongs in post-market surveillance rather than in a footnote.

Addressing it with a warning in the instructions for use will not survive scrutiny. It is answered through formative usability work, through what the interface shows about confidence and provenance, and through whether the clinician can reach the underlying data quickly enough to be bothered to look.

Two competences, not one

Risk management for these products needs someone who can enumerate how the system fails, covering preprocessing edge cases, distribution shift, label noise and degradation that produces no error message. It equally needs someone who can judge what each failure does to a patient in the real care pathway: what a false negative means in a triage queue, whether the clinician would catch it, what a two-week delay costs. An engineer cannot make the second judgement and a clinician cannot make the first. Risk files written by one of them alone are the most common weakness I see, and they are usually thorough in one dimension and empty in the other.

06 WHAT CANNOT BE RECOVERED

Read this section carefully. Everything above can be corrected at a cost. These five cannot.

- **Consent scope.** Data collected under a research ethics approval is not automatically available for commercial product development. If the participant information sheet described a study and the model you intend to sell was trained on that data, this is not just a documentation problem. Related, and more often the actual constraint: collaboration agreements and data sharing agreements with universities and trusts routinely carry academic or non-commercial use clauses, which bind regardless of what the consent says.
- **A dataset that has been seen.** Cross-validation is normal practice for a paper. It is not independent validation for regulatory purposes. Once a dataset has driven development, feature selection and tuning, it cannot serve as independent evidence of performance, because the information has leaked into the model. Teams find this out while drafting a clinical evaluation, at which point the existing evidence becomes preliminary and a prospective study becomes pivotal. A separate limit applies to what the data covers rather than how it was used: a single-site convenience sample supports claims about people resembling the people it saw, and no reanalysis extends it to the intended population. The AI Act's bias examination requirements and the MDR's expectation of performance across the intended population are both answered with new data or not at all.
- **Design rationale, after the people leave.** Choices made during research had reasons behind them, usually good ones. If nobody wrote them down and the postdoc who made them has moved on, the reasoning has gone. A design history file assembled afterwards records what was decided but not why, and why is what gets asked about.
- **Claims already published.** What a website, a press release or a grant application says about the product forms part of its intended purpose, whatever the regulatory documentation later asserts. Wording can be corrected going forward, but the earlier version stays on the record.
- **Borrowed weights.** Fine-tuning a published foundation model is ordinary research practice and creates two problems that cannot be fixed afterwards. You cannot describe the provenance or composition of the base model's training data, which is what the AI Act's data governance requirements expect of you, and you cannot verify what was in it. Separately, the licence may prohibit commercial use, clinical use, or both, and licences on public model repositories are often neither read nor recorded at the moment the model is downloaded. Both questions cost nothing to answer before training and a great deal afterwards.

07 WHY THE BUDGET IS USUALLY WRONG

Teams budget for the work they can see: the model, the study, and perhaps a consultant to write the technical file. The estimate comes apart because a large part of the work is invisible until somebody lists it, and nobody lists it until the money has already been allocated elsewhere.

A partial inventory of what tends to be missing from the plan: implementing a quality system and passing the audit that follows it; reconstructing design history; building and maintaining a SOUP inventory; cybersecurity work under IEC 81001-5-1; formative and summative usability studies; risk management with clinical participation; clinical evaluation and the literature work underneath it; post-market surveillance and clinical follow-up plans, which have to exist before the product can be placed on the market; notified body queue time and two or three rounds of deficiency questions; and AI Act documentation layered onto the MDR file.

The result is a schedule problem before it is a money problem, which is what makes it dangerous. Runway is planned against a launch date. The date moves by three quarters. The round has to be raised earlier, on weaker evidence and worse terms, and the regulatory work then competes for money with the thing it was supposed to enable. A regulatory strategy at the front end costs a small fraction of a funding round, and most of its value is simply in making the invisible items visible while there is still time to budget for them.

08 THREE MARKETS, CONVERGING

Building for one market no longer means rebuilding for the others. The quality system underneath all three is now substantially the same, which is the best argument for starting properly rather than cheaply.

United States	21 CFR Part 820, now the Quality Management System Regulation, with marketing submission via 510(k), De Novo or PMA. Since 2 February 2026 Part 820 incorporates ISO 13485:2016 by reference, so a 13485 system is close to the US baseline rather than a separate exercise. FDA guidance on predetermined change control plans for AI-enabled devices has applied since August 2025.
European Union	MDR 2017/745, or IVDR 2017/746 where the software examines a specimen taken from the body, with the AI Act on top. An AI device needing notified body assessment is automatically a high-risk AI system, but the AI Act obligations fold into the MDR conformity assessment rather than forming a second one. The Digital Omnibus moved the date for AI embedded in regulated products to 2 August 2028; most SaMD sits on that track rather than the December 2027 one, which is widely misread.
United Kingdom	UK MDR 2002 and UKCA marking, with MHRA reform ongoing. Classification can differ from the EU outcome for the same product, so a single class statement covering both is usually wrong. Anything heading into the NHS will meet DTAC and clinical safety expectations under DCB0129 well before it meets a conformity marking.

09 THE PRACTICAL POINT

Almost none of this is hard if it is dealt with at the handover. Most of it is expensive at submission, and some of it is terminal. The teams who find the transition manageable are not the ones who worked harder afterwards, but the ones who spent a fortnight on it before writing production code.

An hour, at no charge

Engagements here begin *with a conversation*. If you want an unstructured hour to establish where you actually stand, I keep one free most weeks for early-stage teams. No charge, no commitment, and I will not follow up afterwards. Write to **peter.brady@ormer-health.ai**.

Peter Brady is co-founder and CEO of Ormer AI Ltd, with three decades in medical-device systems engineering, software development and regulatory science. He is an ISO 13485 lead auditor with EU notified body experience and holds an MSc in Computer Science specialising in machine learning and computer vision. He is currently lead statistician on a multi-centre clinical validation study and is developing machine-learning diagnostic models for respiratory disease.