BLUE GOAT CYBER
Medical Device Cybersecurity · FDA Specialists

AI/ML PCCP & CYBERSECURITY GUIDE

AI/ML Medical Device PCCP & Cybersecurity Guide

Build a PCCP that survives FDA review and a threat model that covers the ML pipeline - poisoning, inversion, adversarial, prompt injection.

bluegoatcyber.com · 100% FDA submission success rate

AI/ML SaMD

AI/ML medical device PCCP & cybersecurity guide.

AI/ML-enabled SaMD presents unique cybersecurity threats - the model itself is an asset that can be attacked. The FDA's 2026 cybersecurity guidance specifically calls out poisoning, inversion, adversarial inputs, and prompt injection. At the same time, the FDA's **PCCP** framework lets you pre-authorize model updates without a new submission - but only if the plan is precise, bounded, and verifiable.

Key takeaway: A PCCP without a cybersecurity dimension is incomplete. Every modification protocol step needs a security check - not just performance metrics.

1. PCCP - the three required components

Description of Modifications	Exactly what can change: weights, thresholds, features, architecture, training data sources. Be specific - vague plans get rejected.
Modification Protocol	Verification, validation, performance, **and security** checks per change type. Include rollback criteria.
Impact Assessment	Benefit/risk analysis for each modification, tied to ISO 14971. Document residual risk per modification class.

2. ML-specific threats to model

Data poisoning	Adversary corrupts training or fine-tuning data. Mitigations: dataset provenance, anomaly detection, federated training controls.
Model inversion	Extract training data from model outputs. Mitigations: differential privacy, output rate limits.
Membership inference	Determine if a record was in the training set. Mitigations: regularization, DP-SGD.
Adversarial inputs	Crafted inputs that flip decisions. Mitigations: adversarial training, input validation, ensemble verification.
Model theft / extraction	Query-based reconstruction. Mitigations: rate limiting, query monitoring, watermarking.

Prompt injection	For LLM-backed SaMD, override system instructions. Mitigations: input/output filtering, structured prompts, principle of least privilege for tool use.
Supply chain	Compromised pre-trained models, datasets, or libraries. Mitigations: ML-BOM, signed model artifacts, vetted sources.

3. ML-BOM - the bill of materials for your model

- Use CycloneDX 1.5 ML-BOM extension.
- Include training datasets, pre-trained base models, and fine-tuning data.
- Document model architecture, training framework, and hyperparameters.
- Hash and version every model artifact (weights, tokenizers, configs).
- Track model provenance through the deployment pipeline.
- Include licenses for datasets and base models - not just code.

4. Monitoring - drift and cybersecurity together

Performance drift	Clinical metrics - already required by GMLP. Trigger retraining within bounded plan.
Input distribution drift	Statistical drift can be early signal of adversarial probing or data quality issues.
Output anomaly detection	Unusual confidence patterns, repeated boundary cases - catch evasion + inversion.
Inference rate-limiting	Per-user, per-API-key. Defense against extraction and DoS.
Forensic logging	Inputs, outputs, model version, decision logged with retention sufficient for incident review.

5. Aligning PCCP modifications with security checks

- **Threshold tuning** - re-run adversarial robustness suite against new thresholds.
- **Re-training on new data** - revalidate dataset provenance + run poisoning detectors.
- **Architecture change** - re-threat-model the change; update STRIDE-per-element analysis.
- **Adding LLM/agent capabilities** - prompt injection test suite; tool-use authorization review.
- **Deployment change (edge -> cloud)** - revisit data flow, encryption, key management.

6. GMLP × 2026 cybersecurity guidance crosswalk

Multi-disciplinary expertise	Add security + ML safety reviewer to the team.
Good software / security engineering	SPDF + IEC 62443-4-1 patterns apply to model code AND training pipelines.

Clinical evaluation independence	Hold-out test sets that adversaries can't influence; security test sets distinct from clinical.
Training data representativeness	Provenance documentation feeds both bias review AND poisoning defenses.
Model performance monitoring	Combine clinical drift with security drift in one monitoring stack.

7. Common review-letter findings for AI/ML SaMD

- PCCP too vague - reviewer can't tell what's in/out of scope.
- Threat model missing AI-specific threats (poisoning, prompt injection).
- No ML-BOM - just a code SBOM.
- Monitoring plan covers performance but not adversarial behavior.
- No human-in-the-loop or rollback criteria for autonomous modifications.
- LLM features without prompt injection testing or output filtering.

Key takeaway: The FDA is explicit in 2026: AI/ML threats belong in your threat model. If your existing model lists STRIDE for the application but says nothing about the ML pipeline, expect a deficiency.

READY FOR THE NEXT STEP?

Talk to a senior medical device security engineer.

We've contributed cybersecurity documentation, threat models, SBOMs, and pen test reports to **250+ FDA submissions** across 510(k), De Novo, and PMA pathways - with a **100% submission success rate**. If you're building a cyber device and want to go in with airtight artifacts on the first try, we should talk.

Book an AI/ML SaMD working session
30 minutes. No slide deck. Bring your device profile and your top three questions.

bluegoatcyber.com/contact

WHAT WE DO

- **FDA premarket cybersecurity** - SPDF, threat model, SBOM, pen test, security risk assessment, eSTAR-ready package
- **FDA cybersecurity deficiency response** - rapid remediation when reviewers send a letter
- **Postmarket programs** - vulnerability monitoring, VEX, coordinated disclosure, patch validation
- **Legacy device remediation** - bring older devices up to current FDA expectations