EVER MEDICAL TECHNOLOGIES · VISION & RESEARCH

From Sites to Patients: A Patient-Contributed, Inference-Scaled Trajectory Foundation Model on a Decentralized Health-Data Commons

Working paper — draft v0.3 · Ever Medical Technologies. Successor to and superseding *Sovereign Health Trajectory Prediction* (draft v0.2). Companion to the implementation in `ever-sovereign-wallet` and the integration design in `medOS-ultra/docs/architecture/central-trajectory-engine.md`, `ethos-ng-modern-trajectory-model.md`, and `multilingual-clinical-ai-plan.md`.

What changed from v0.2. Three substantive shifts. (1) **Unit of contribution.** v0.2 framed data acquisition as a set of per-HIE-site ETL pipelines ("each site is an adapter"). This version inverts that: the unit of contribution is the **patient**, who assembles a cross-institutional longitudinal record inside the sovereign wallet and grants us a permissioned key exchange to use it for training. Institutional records still flow in — but *through the patient*, who is the point of convergence, not through a per-institution integration contract. (2) **Emphasis.** v0.2 argued the critical path was data engineering, not modeling, and treated the model as a contract awaiting fill. The data-engineering point still holds (Section 10), but this paper is **about the model**: what 2026 trajectory-model architecture we are building, why patient-contributed scale changes the modeling calculus, and how inference-time compute unifies prediction, abstention, and on-device delivery. The wallet is the substrate; it appears here as context, not as the subject. (3) **Geography.** We describe an initial contributing population in **Asia** and a cross-border expansion path; we do not name a single launch country, and we treat each new region as both a generalization test and a calibration target.

Status & honesty note. This remains a systems + architecture paper. The sovereign-wallet substrate and the patient-contribution / key-exchange mechanism are **shipped** (we cite the files). The multimodal ingestion-to-MEDS layer is **partially shipped**. The trajectory model itself is **designed but not yet trained**: `population.ts` is still a null implementation that honestly abstains, and `forecast.ts` still ships only the classical linear floor. Every predictive number in Section 11 is a **target**, not a measured result. The multi-million-record asset that patient contribution is intended to produce is a *data flywheel under construction*, not a trained model. We make no clinical-performance claim we have not measured.

Abstract

Generative clinical foundation models now simulate a patient's medical future token-by-token and read predictions off the simulated trajectory, zero-shot across dozens of tasks. The 2025–2026 frontier (CoMET on Epic Cosmos — 118M patients, 115B events; the first large-scale EHR scaling laws; multimodal structured-plus-unstructured models such as GDP) has established two things: **these models obey power-law scaling**, so data access is the binding constraint, and **simulation-based inference is itself a form of test-time compute**, so the same sampling that produces a forecast can be made to produce a calibrated *abstention*. Yet essentially all of this work is hospital-centric: the model is trained on institutional data, served from the cloud, and reasons over records the patient neither holds nor controls — and most of it validates on the same institution it trained on.

We describe a different acquisition substrate and a different delivery substrate, joined by one model. **Acquisition:** instead of integrating *sites*, we let *patients* contribute. Each patient assembles a longitudinal, cross-institutional record in a self-sovereign wallet (encrypted under patient-custodied keys, content-ad-

dressed, identity-anchored by a biological DID with zero-knowledge proofs) and grants a permissioned key exchange for de-identified training use. Because a patient's record already spans every provider they have seen, the contributed corpus is **multi-institutional and longitudinal by construction** — it solves the single-center problem at the source rather than by federating sites after the fact. Because the patient, not Ever, is the data owner, the trust barrier to contribution falls, which is precisely the condition for a data network effect: the substrate that can reach the most patients is the substrate that reaches the scale where these models become good. **Delivery:** an offline-first, on-device student model, distilled and quantized from the large teacher, runs the trajectory simulation locally; an abstain-by-default population prior conditions an honest personal forecast; and a deterministic guardrail redacts diagnosis and certainty claims from every generated string. The product is recommender-only — it tells the patient what to discuss with a clinician, never a diagnosis.

We contribute: (1) the **patient as the unit of contribution**, with an analysis of why decentralized, non-custodial data ownership is an *acquisition strategy*, not only a privacy posture; (2) an **engineering treatment of self-reported data** — multimodal normalization to the Medical Event Data Standard (MEDS), provenance-conditioned modeling that distinguishes clinician-attested from self-asserted events, robustness to noise and missingness, and an explicit defense against model collapse from AI-generated content; (3) a **2026 trajectory-model design** that combines long-context / state-space backbones, per-code numeric and continuous-time encoders, a generative-plus-time-to-event objective, conformal calibration, and **inference-time scaling as the unifying mechanism** for forecasting and abstention; (4) a **decentralized training posture** (cross-device federation where data residency forbids egress, consented de-identified pooling where it does not) with differential privacy, secure aggregation, and the wallet's existing ZK/HE primitives, targeting the membership-inference risk that long clinical sequences uniquely pose; and (5) **held-out-region validation** as the acceptance gate — train on contributing populations, validate on an unseen region with no fine-tuning — which is the only honest basis for "a generalizable model that began in Asia" rather than "an Asia-only model."

---

1. Introduction

Roughly 3.5 billion people have unreliable internet access. When a parent's child is feverish at 2 a.m., today's consumer health AI is a cloud service that requires connectivity, transmits the symptom to a server, and answers questions but does not predict. The clinical-AI frontier, meanwhile, has produced generative trajectory models that *do* predict — forecasting future health events token-by-token from a patient's history, zero-shot across many tasks. But those models live in the hospital, served from the cloud, reasoning over records the patient does not hold, and — for the most part — validated on the very institution they were trained on.

v0.2 of this work asked how to deliver that predictive power to a patient on their own device, under a sovereignty and safety envelope, and how to make the underlying model multi-site rather than single-center. It answered with a shipped sovereign-wallet substrate and a designed multi-site model in which **each HIE site was an ETL adapter**.

This version revises the central premise. The hardest part of a multi-site model is not the architecture; it is getting longitudinal, cross-institutional sequences at all. v0.2 proposed to solve this by integrating sites one at a time. But there already exists a structure in which cross-institutional longitudinal data converges naturally: **the patient**. A person accumulates records from every clinic, hospital, lab, and pharmacy they touch. If the patient can hold all of those records in one place and choose to contribute them, the corpus

is multi-institutional and longitudinal *before any federation* — and onboarding a new country becomes onboarding *patients*, not negotiating HIE-by-HIE data-use agreements.

So this paper asks two sharper questions:

1. **If scale is the lever** — and the 2025–2026 scaling-law results say it is [CoMET; Zhang et al.] — **what is the most scalable way to acquire longitudinal clinical sequences?** We argue it is patient contribution through a decentralized, non-custodial wallet, because non-ownership lowers the trust barrier and produces a data network effect that per-site integration cannot.
2. **Given that patient-contributed data is self-reported and self-selected, how do we build a trajectory model that is robust to it, honest about it, and safe to ship to non-clinicians?** This is the modeling problem we take up directly: multimodal normalization, provenance conditioning, inference-time abstention, conformal calibration, and on-device delivery.

The wallet — `ever-sovereign-wallet`, an Electron desktop application with React-Native mobile and web siblings sharing one TypeScript core — is the substrate that makes patient contribution possible. We summarize it in Section 7 and refer to v0.2 for the full custody design. The rest of this paper is about the model.

The result is not "a smaller CoMET." It is a different point in the design space — **patient-contributed, sovereign, edge-served, population-calibrated, inference-scaled, abstaining** — that complements the hospital model rather than competing with it.

---

2. Background and the 2026 landscape

2.1 Generative trajectory foundation models and the scaling-law turn

The canonical model is ETHOS (Renc et al., *npj Digital Medicine* 2024): a decoder-only GPT over tokenized Patient Health Timelines (PHTs) in which continuous values are quantized into deciles, time gaps into interval tokens, and future PHTs are generated and read off as risk. Its successor `ethos-ares` (Renc et al., *GigaScience* 2025) rebuilds it on the MEDS event standard and adds the ARES per-patient attribution mechanism. Delphi-2M (Shmatko et al., *Nature* 2025) adds continuous-age encoding and an exponential waiting-time head over 1,000+ diseases on a 20-year horizon, externally validated from UK Biobank to a Danish registry; MOTOR (Steinberg/Shah, ICLR 2024) is a piecewise-exponential time-to-event foundation model; CEHR-GPT, Foresight (Kraljevic et al., *Lancet Digital Health* 2024), TransformEHR, and EHRMamba round out the family.

What is genuinely new in 2025–2026, and what reframes our work, is **scale and its laws**:

- **CoMET** (Cosmos Medical Event Transformer; Epic Systems, Microsoft Research, Yale, *Generative Medical Event Models Improve with Scale*, 2025) is, to date, the largest scaling-law study on real-world patient journeys. Pretrained on Epic Cosmos — de-identified longitudinal records covering 300M unique patients and 16.3B encounters across 310 health systems — CoMET is a family of decoder-only transformers (to ~1B parameters) trained on **118M patients, 115B discrete medical events, 151B tokens**. Conditioned on a patient's history it autoregressively generates the next event and reads predictions off simulated trajectories. Across **78 real-world tasks** it generally matched or beat task-specific supervised models *with no fine-tuning*, and — the load-bearing finding — its predictive power improves **predictably with model and data scale**, following power laws analogous to the text domain but with a higher optimal token-to-parameter ratio.

- The first dedicated **EHR scaling-law study** (Zhang et al., *Exploring Scaling Laws for EHR Foundation Models*, 2025) reached the same qualitative conclusion at smaller scale on MIMIC-IV: power-law relationships among compute, model size, and token count, with a distinctively high token-to-parameter ratio for medical events.

The implication for us is direct and is the spine of this paper: **if performance scales with data, the binding constraint is data access, and the right strategic question is "what is the most scalable acquisition path?"** Hospital integration scales linearly in negotiated contracts. Patient contribution, if trust permits it, scales with adoption and crosses borders for free. CoMET shows what scale buys; we argue patient contribution is how a sovereign, in-region program can reach it.

2.2 Heterogeneous and multimodal ingestion

Patient-contributed data is messy and multimodal — scanned PDFs, lab photographs, wearable streams, free-text notes, structured exports. The relevant frontier model is **GDP (Generative Deep Patient**, Sivarajkumar et al., 2025): a multimodal foundation model that natively encodes structured EHR time-series via a CNN-Transformer encoder and fuses unstructured records through cross-modal attention into a LLaMA-based decoder, trained in two stages (generative pretraining with masked-feature and next-time-step prediction, then multi-task fine-tuning). On MIMIC-IV it reports heart-failure AUROC 0.923, type-2-diabetes 0.817, and 30-day readmission 0.627. The lesson we take is *not to serialize numeric EHR into text* (which discards temporal and quantitative structure) but to encode each modality natively and fuse — exactly the discipline a self-reported, multi-format corpus demands.

2.3 Inference efficiency for generative prediction

Simulation-based inference is expensive: a forecast requires sampling many futures. Two 2026 estimators make this affordable, which matters acutely on-device. **SCOPE and REACH** (*Efficient Generative Prediction for EHR Foundation Models*, 2026) reuse a single sampled pool across an arbitrary number of outcomes at no marginal generation cost (SCOPE) and provide a per-task variance guarantee (REACH), together cutting the inference budget required for generative EHR models — particularly for **rare, high-impact outcomes**, the ones a screening tool most needs and most easily gets wrong. We adopt this family directly (Section 5.3): the variance estimate that REACH provides is also the signal our prior uses to decide whether to abstain.

2.4 On-device language models in 2026

On-device deployment converged in 2026 on a clear playbook (Edge AI & Vision Alliance, *On-Device LLMs in 2026*; Chandra, *On-Device LLMs: State of the Union, 2026*). The major labs ship sub-billion-to-few-billion models — Llama 3.2 (1B/3B), Gemma 3 (down to 270M), Phi-4-mini (3.8B), SmolLM2 (135M–1.7B), Qwen2.5 (0.5B–1.5B) — and three facts shape any edge clinical model:

- **Below ~1B parameters, architecture beats size**: deeper, thinner networks outperform wide, shallow ones, and capability at small scale is driven by data quality and distillation from larger teachers more than by parameter count.
- **Decode is memory-bandwidth bound, not compute bound**: generating each token streams the full weights, so for long-context use **KV-cache compression is often more impactful than weight quantization** (StreamingLLM's attention sinks, DuoAttention's retrieval-vs-streaming head split, ChunkKV's semantic-chunk compression; KV cache quantizable to ~3 bits with negligible loss).
- **Test-time compute lets a small model punch up**: a Llama-3.2-1B with search strategies can outperform the 8B base model. 4-bit quantization has moved from compromise to strategic lever for

clinical LLMs on consumer hardware (e.g., the Gemma-3 family on medical benchmarks).

These facts are why an on-device trajectory student is feasible at all, and they connect to our central design idea: the generative model already *is* a test-time-compute machine, so the edge budget should be spent on **sampling more futures for uncertain cases** rather than on a bigger backbone.

2.5 Decentralized and privacy-preserving training

Federated learning began as a **cross-device** method — McMahan et al.'s FedAvg (2017) trained on mobile phones — before the healthcare literature redirected it to **cross-silo** settings (hospitals as clients; FED-EHR 2025; numerous DP+HE frameworks 2025–2026). Our setting returns FL to its cross-device roots: the federation units are **patient-held wallets**, not institutions. The privacy toolkit is mature: differential privacy (DP-SGD with per-sample gradient clipping and Gaussian noise, accounted by a privacy budget ε), secure aggregation, and homomorphic encryption defend against the two attacks that matter here — **membership inference** and **gradient inversion** — with the well-documented caveat that DP noise trades off utility and can disproportionately degrade underrepresented subgroups, which makes the fairness audit (Section 8) non-optional. Parameter-efficient federated variants (e.g., FedSA-LoRA-DP, 2025) combine selective low-rank aggregation with DP, which fits our per-country LoRA-spoke design.

2.6 The specific hazards of self-reported and AI-contaminated data

Three results define the risk surface for a patient-contributed corpus:

- **Model collapse from self-referential training.** A 2026 study of clinical-text models showed that training across generations on AI-generated clinical notes collapses the model: vocabulary fell ~98.9% (from ~12,000 to ~200 unique words), unique medical terms fell ~66%, rare pathologies vanished, and the model acquired *dangerous false confidence*. As LLM-drafted notes saturate real EHRs, a naive corpus will ingest synthetic text. Mitigations the authors identify — real/synthetic mixing limits and quality-aware filtering — become first-class requirements for us (Section 4.4).
- **Patient-report vs. EHR concordance.** Self-reported clinical facts agree with the chart only partially (e.g., O'Brien et al., ADAPTABLE trial, *JAMA Cardiology* 2022; older validations of self-reported chronic conditions). Patient report adds signal the chart lacks (cross-provider continuity, symptoms, social context) but cannot be treated as ground truth.
- **Access-to-care bias in the records themselves.** A 2026 *Nature Health* analysis on All-of-Us data showed that EHR completeness is itself a function of healthcare access, biasing AI predictions. Patient contribution can *reduce* one bias (fragmentation across providers) while *introducing* another (who chooses to contribute). We treat contribution self-selection as a calibration problem, not a footnote (Section 4.5).

2.7 Where we sit

Every model above is hospital/cloud-served and trains on data the patient does not hold; most report same-site validation. Our contribution is orthogonal on both axes: we **acquire** through patient contribution to a decentralized commons, and we **deliver** an abstaining prior on the patient's device under a sovereignty and safety envelope — and we hold ourselves to **held-out-region** validation.

---

3. The patient-contributed data commons

This is the section that most changes from v0.2.

3.1 The patient is the unit of contribution

In v0.2, the data graph was *site → pipeline → corpus*. Here it is **provider → patient's wallet → (permissioned key exchange) → de-identified commons**. The patient's wallet is the aggregation point. Records enter it the way a person actually accumulates them: a discharge summary imported from one hospital, lab results from a reference lab, a pharmacy fill history, a wearable export, a self-entered allergy. The patient then makes one decision — to contribute, under a granted key exchange — and that single decision yields a **multi-institutional, multi-year sequence**, the exact object a trajectory model needs and the exact object single-center datasets lack.

Concretely, the wallet already namespaces records as `biomarkers, medications, allergies, conditions, immunizations, familyHistory, vitals, wgs, gutMicrobiome, encounters, careplan, timeline`, each tagged with CCDA/LOINC + HL7v2 + FHIR (`CLINICAL_SECTION_MAP`), and seals them under NaCl SecretBox keyed by PBKDF2 over the patient passphrase and wallet address. Contribution is therefore not "send us your data" but "**grant us a re-keyed, de-identified, in-region copy of an already-assembled timeline**," with the patient retaining custody of the original. No party writes into another party's store; sharing is by an immutable signed folder (link) or a re-keyed copy (sync) — the property that preserves sovereignty under collaboration.

3.2 Why non-custodial ownership is an *acquisition strategy*, not only a privacy posture

The standard framing of self-sovereign data is privacy and portability. We add a second, more strategically important claim: **decentralized, non-custodial ownership lowers the trust barrier to contribution, and the trust barrier is the real bottleneck on clinical-data scale.**

The logic is straightforward. The 2025–2026 scaling laws say model quality is data-bound. Hospital data sits behind data-use agreements, IRBs, and licensing constraints (the published ETHOS/ARES weights are a MIMIC-IV derivative and are *non-redistributable* under the PhysioNet DUA — the `ethos-ares` authors decline to publish their weights for exactly this reason). The one party that can lawfully aggregate a person's cross-institutional record and consent to its use is **the person**. A patient is far likelier to contribute when (a) they hold the keys, (b) they can revoke (right-to-erasure crosswalk), (c) the operator is *not* the ultimate owner, and (d) contribution proves a property without revealing the payload (W3C Verifiable Credentials with Groth16 ZK proofs let a contributor prove "over 18" or "identity matches" without disclosing the data). Each of these is a trust reduction, and each trust reduction widens the contributing population — which, by the scaling laws, is the thing that improves the model. **Non-ownership is how you get the data.**

This produces a **flywheel**: more contributors → a larger, more longitudinal commons → a better trajectory prior → a more useful on-device product (better forecasts, better abstention) → more reason to contribute. The asset compounds and is in-region and ours to distill to the edge (retrain-not-redistribute). We are honest that the flywheel is *under construction*: today the commons is a designed pipeline plus a shipped wallet, not a multi-million-patient trained model.

3.3 Cross-border expansion without per-country integration

Because the model reasons over **codes** (ICD, ATC, LOINC, SNOMED), which are language-independent, and because the contribution mechanism is the same wallet everywhere, **adding a country is adding patients, not building a pipeline**. Geographic expansion in v0.2 meant standing up a new HIE adapter; here it means the app reaching new users who already hold their own records. Each new region is simultaneously (a) a source of fresh contributors, (b) a **calibration target** (local base rates, formularies, pharmacogenomics; Section 8), and (c) a **held-out-region generalization test** (Section 11). We start with an

initial contributing population in Asia and expand outward; we name no single launch country in this document.

3.4 What does *not* get easier

Honesty requires the converse. Patient contribution does **not** eliminate — and in places worsens — three problems, taken up in Section 4: data is **self-reported** (concordance, not ground truth), **self-selected** (contribution bias), and arrives in **heterogeneous formats** (normalization burden moves from a few institutional schemas to an open-ended set of patient-supplied formats). The commons trades a contracting problem for an engineering problem. We think that is the right trade — engineering scales, contracting does not — but it is a trade, not a free lunch.

---

4. Engineering for self-reported data

The user-facing question this paper must answer: *does it change anything that the data is self-admitted?* Yes — in five specific, addressable ways.

4.1 Multimodal normalization to MEDS

A contributed timeline may include a photographed lab report, a scanned discharge PDF, a FHIR export, a wearable stream, and free-typed text. We make **MEDS the single internal representation** between ingestion and the tokenizer, and we treat normalization as a **multimodal extraction problem** in the GDP lineage (Section 2.2): native encoders per modality (OCR + layout for documents, signal encoders for wearables, structured parsers for FHIR/HL7), each emitting time-stamped MEDS events with units and codes, rather than flattening everything to text and losing magnitude and time. Per-source format becomes an **adapter**; the tokenizer, trainer, and evaluation code run unchanged across formats — the same decoupling v0.2 used for sites, now applied to the open set of patient-supplied formats. Vocabulary harmonization (ICD-9→ICD-10, local formularies→ATC, lab catalogs→LOINC, procedures→SNOMED/PCS) is the real engineering and is where quality is usually lost; we budget for it and audit crosswalk accuracy on a held-out set (Section 10).

4.2 Provenance as a first-class, conditioned signal

The decisive move for self-reported data is to **stop pretending all events are equally trustworthy and instead model the difference**. Every MEDS event carries a provenance tag: **clinician-attested** (imported from an institutional record, optionally backed by a Verifiable Credential), **device-measured** (a wearable or connected meter), or **self-asserted** (patient-entered). Provenance is then (a) a covariate the model conditions on, so it can learn that a self-asserted "diabetes" carries different predictive weight than a clinician-coded E11 with corroborating HbA1c labs, and (b) a filter for training-set construction and for evaluation strata. This is exactly the discipline the patient-report-vs-EHR concordance literature implies: patient report is *signal with a known error model*, not ground truth, and the right response is to give the model the provenance and let it calibrate.

4.3 Robustness to noise, missingness, and irregular sampling

Self-assembled records are sparser and more irregular than a single institution's. We adopt: **per-code continuous value embeddings with per-source normalization** (so a creatinine value keeps its magnitude and is comparable across assays — the multi-site failure mode v0.2 identified, now also a multi-*source* failure mode); **continuous-time encodings plus an explicit waiting-time / time-to-event head** (so ir-

regular gaps are modeled, not discretized away); and **missingness-aware tokenization** (absence is informative and is represented, not imputed silently). The generative objective is itself somewhat robust to label noise because it learns a distribution over futures rather than fitting a single hard label.

4.4 Defending against model collapse

Because LLM-drafted notes increasingly populate real records, a patient-contributed corpus *will* ingest AI-generated content, and the 2026 collapse result (Section 2.6) is a direct threat: train on synthetic text and the model loses rare pathologies and gains false confidence. Our defenses:

1. **Never train the prior on AI-generated free text.** The trajectory model trains on **coded MEDS events**, not on narrative prose, which sidesteps the text-collapse pathway by construction; the conversational shell stays retrieval-grounded against a physician-cited knowledge base rather than fine-tuned on contributed prose.
2. **Provenance-gate synthetic content.** Where an imported document is detectably machine-generated, it is tagged and excluded from (or down-weighted in) the training mix per the real/synthetic-mixing thresholds the collapse literature recommends.
3. **Quality-aware filtering and rare-event preservation.** Long-tail sequences are protected, not pruned, in tension with the de-identification suppression of Section 6 — a tension we manage explicitly rather than resolve by default.

4.5 Contribution self-selection and calibration

Whoever chooses to contribute is not a random sample of the population. This biases base rates and therefore calibration. We treat it as a measurable, correctable quantity: we estimate the contributing cohort's composition against regional reference statistics, **reweight** during training where feasible, **calibrate** with conformal methods on a representative held-out set, and **report** stratified performance (region, ethnicity, age, sex, and *provenance mix*) so the bias is visible rather than hidden. Calibration is a design axis (Section 8), and contribution self-selection is one of its inputs.

4.6 Biosignal and acoustic modalities — voice, EEG, and wearable streams

A patient-held wallet sees signals a hospital chart never records. The wallet already normalizes **EEG band powers** (delta/theta/alpha/beta/gamma, a derived focus index), **voice acoustic biomarkers** (jitter, shimmer, MFCC-derived indices via a DisVoice-style adapter), and **continuous wearable vitals** (heart rate, HRV, SpO_2, respiration, temperature, sleep stages) into one flat `SignalObservation` stream (`src/shared/wellness/twin/signals.ts`; ingest adapters `fromBrainflowBandPowers` / `fromNeuroKit2` / `fromDisVoice` in `ingest.ts`). These are **device-measured** under the Section 4.2 provenance scheme — a third, distinct trust class between clinician-attested and self-asserted — and they are exactly the high-frequency, longitudinal channels that institutional EHRs lack and that a *patient*-centric corpus uniquely contributes.

Three modeling consequences:

1. **Encode natively, do not serialize to text.** Following the GDP discipline (Section 2.2), each biosignal modality keeps its native shape — band-power vectors, acoustic feature frames, and irregularly-sampled vital series are encoded by signal encoders into MEDS events with units and (where one exists) a LOINC/SNOMED code, falling back to a stable Ever channel vocabulary (`eeg_alpha`, `voice_jitter`, `hrv`, …) when no standard code applies. This reuses the §4.3 per-code continuous value embeddings and continuous-time encoders; a biosignal is just another channel on the trajectory, sampled far more densely.

2. **Asynchronous, optional, and abstaining.** EEG and voice are *linkable* but rarely present (a headband, a microphone session); the model conditions on them when available and is unharmed when absent (missingness-aware tokenization, §4.3). They never gate a prediction — they enrich it.
3. **Acoustic and neural data are sensitive and stay on-device.** Raw audio and EEG never leave custody; only derived, de-identified feature events enter the commons under the same consent, k-anonymity, and DP envelope as coded events (Sections 6, 9). Voice in particular can carry identity, so only abstracted acoustic indices — never waveforms — are contributable.

In short: the patient-contribution thesis is not only "more complete *clinical* records"; it is **modalities the clinical record does not contain at all** — neural, vocal, and continuous-physiological — fused into the same trajectory the coded events ride on.

---

5. The trajectory model (the AI core)

We fill `population.ts` with a generative trajectory model — call it **ETHOS-NG** — retrained on the contributed commons, not on published MIMIC-derived weights. The ETHOS code is MIT, so we reuse the architecture and train our own weights; training on the deployed populations' real data directly answers the Western-bias problem (Section 8), keeps data in-region for data-protection law, and yields weights we own and may distill to the edge.

5.1 The four upgrades over 2024 ETHOS, updated for 2026

Dimension	ETHOS (2024)	ETHOS-NG (this design)	Why it matters
Context / backbone	learned absolute positions, 2,048-token window, O(n²) attention	RoPE + FlashAttention-style long context; optional Mamba/SSM hybrid (EHRMamba lineage) for lifelong PHTs	Dense multi-decade, multi-institution patient-assembled histories truncate at 2,048 tokens — the original's stated weakness. Long context is the single biggest fidelity win, and is *more* pressing for cross-provider records that are longer than any one institution's.
Numeric values	10 global decile tokens	per-code continuous value embeddings (or per-code adaptive binning) with **per-source** normalization	A "Q8 creatinine" loses magnitude and differs by assay/units across sources — the exact self-reported failure mode. Real value embeddings plus per-source normalization fix both.
Time	13 discrete interval tokens; next-token only	continuous time embeddings + explicit time-to-event head (MOTOR / Delphi-2M style)	Clinicians want calibrated risk over the next 30/90/365 days — a survival quantity, not a sampling byproduct — and irregular patient-contributed sampling demands explicit time modeling.
Calibration / trust	raw sampled probabilities	conformal prediction intervals + ARES-style per-patient attribution + **provenance conditioning**	Distribution-free intervals, "which past events drove this," and "how much of this rests on self-asserted vs. clinician-attested data" are what a clinician, a regulator, *and a self-reporting patient* will accept.

Two things stay unchanged on purpose: the **generative, zero-shot future-timeline mechanism** (sample N futures, read off empirical risk), because it is the whole advantage over single-task classifiers and the

thing the 2026 scaling laws reward; and **interpretable, code-level tokens**, because they keep the model auditable and the edge distillation honest.

5.2 Pretraining objective and tokenization

Combine **next-token (generative)** with a **time-to-event** loss, so the model both simulates trajectories and emits calibrated horizons. Tokenization is hierarchical over ontologies (decompose ICD-10 / ATC into structural tokens, as ETHOS and CoMET do), with static covariate tokens prepended — demographics, a coarse region/health-system token, and a **provenance-mix token** — so the model learns region- and provenance-specific baselines that can be swapped counterfactually to probe bias.

5.3 Inference-time scaling as the unifying mechanism

This is the conceptual core of the AI design, and the cleanest place the 2026 frontier maps onto our safety requirements.

A generative trajectory model produces a prediction by **sampling many futures and reading off the empirical rate** of the target event. That is *already* test-time compute: more samples buy a more precise estimate. Three consequences:

1. **Spend compute where uncertainty is.** Following SCOPE/REACH (Section 2.3), we reuse one sampled pool across many outcomes and obtain a **per-task variance estimate** essentially for free. Easy, common outcomes converge with few samples; rare, high-impact outcomes get more. On-device, where the binding constraint is memory bandwidth and battery, this lets a small student match a larger model on the cases that matter (the same "small model + search beats bigger model" effect seen generally in 2026).
2. **Abstention is variance, not a heuristic.** v0.2's "abstain-by-default population prior" becomes principled: if the sampled futures **disagree** — if REACH's variance estimate for a target exceeds a calibrated threshold — the prior **abstains**, contributing nothing, and the forecast falls back to the honest linear floor in `forecast.ts`. Confidence is read directly off the simulation, not asserted.
3. **Conformal calibration on top.** Conformal risk control wraps the sampled estimate in a distribution-free interval with a coverage guarantee, computed against a representative calibration set, before any number reaches the patient.

So a single mechanism — sample futures, measure their spread — delivers the forecast, the abstention decision, and the uncertainty interval. This is what makes a generative model *safe* to put in front of a non-clinician: it can say "I don't know" in a way that is grounded in its own disagreement with itself.

5.4 Teacher–student for the edge

Train a large multi-site teacher on the commons; **distill and quantize** to a small on-device student (GGUF via `node-llama-cpp`) in the 2026 edge band (architecture over raw size; deeper-thinner; 4-bit weights; **KV-cache compression** for lifelong context). Per-country **LoRA spokes** adapt formulary, guidelines, and disease mix without re-training the backbone or re-translating strings. The student is necessarily weaker than the in-region teacher (Section 12); abstention and the deterministic guardrail are how we make that gap safe rather than dangerous.

---

6. Decentralized and privacy-preserving training

The data is patient-held and encrypted; the training method should match.

- **Cross-device where residency forbids egress; pooled where it does not.** Where de-identified MEDS may lawfully be centralized in-region, pool it (simpler, stronger gradients). Where data-residency law or patient preference forbids egress, train across **patient-held / regional nodes** in the FedAvg cross-device tradition, moving only weights or distilled students. Expect a hybrid: a pooled in-region core plus federated spokes — the same hybrid v0.2 anticipated, now with the *patient* (not the site) as the potential federation unit.
- **Differential privacy + secure aggregation.** DP-SGD bounds what any single contributor's record can do to the model, defending against **membership inference**; secure aggregation prevents reconstruction of individual updates; the wallet's existing ZK/HE primitives extend this to verifiable, payload-free contribution. We adopt the documented caveat as a hard constraint: DP noise can disproportionately degrade underrepresented subgroups, so the subgroup-fairness audit (Section 8) gates release.
- **Membership inference is the dominant threat for long clinical sequences.** v0.2 already noted that a long ICD/drug sequence is strongly re-identifying; "safe by construction" does not hold for clinical PHTs. We enforce **k-anonymity / suppression on rare long-tail trajectories** and a **fixed cross-site salt** for linkage and right-to-erasure — in explicit tension with the rare-event preservation of Section 4.4, a tension we tune rather than wish away.
- **Parameter-efficient + private.** Per-country LoRA spokes compose with DP (the FedSA-LoRA-DP pattern): share and privatize only low-rank adapters, keeping communication and privacy cost down.

The invariant: **the patient's raw record never leaves their custody on the default path**, the prior either runs on-device as a distilled student or receives only a de-identified, code-only prefix when it runs as an in-region API, and the training corpus is confined to consented, de-identified, in-region data.

---

7. On-device delivery and the sovereign-wallet substrate (in brief)

Per the user's framing, the wallet is context here, not the subject; full detail is in v0.2. The essentials the model relies on:

- **Three data layers** — local custody (NaCl SecretBox over `electron-store`, offline-first), IPFS portability (encrypted CIDv1 envelopes via an embedded Helia node, never server-resident in plaintext), and real-time liveness (WSS P2P shipping pointers, never payloads).
- **On-device-first reasoning** — `AIService` runs local inference via `node-llama-cpp` / `ModelManager` (a ~380 MB free model up to a ~4 GB premium clinical model); the cloud is a metered, authenticated, in-region *fallback*, not the default.
- **The prediction seam** — the shared TypeScript core (`src/shared/wellness/twin/`): `signals.ts` (signal fusion), `patterns.ts` (self-relative drift/anomaly), `forecast.ts` (the gated linear floor with a pluggable `ForecastBackend`), `population.ts` (the abstaining L5-prior contract — where ETHOS-NG lands), and `guardrail.ts` (the deterministic redactor on every generated string).
- **Self-sovereign identity** — `did:bio` (a Poseidon-committed biological identity attestation over genomic markers plus a device-held salt), biometric liveness as a step-up factor, and W3C Verifiable Credentials with Groth16 ZK proofs.

The keystone is unchanged: the seam to a hospital-grade trajectory model already exists, abstains by default, and is bounded by a deterministic guardrail. This paper is about what fills it and where its training data comes from.

---

8. Population calibration as a first-class axis

A model trained on Western cohorts miscalibrates for others: Framingham/QRISK/KFRE were built on Western data; diabetes-screening BMI cutoffs use European thresholds (≥25) rather than Asian-appropriate ones (≥23); and pharmacogenomic variant prevalence differs sharply by ancestry. The clinically load-bearing examples for an Asia-first program:

- **HLA-B*5801** (allopurinol hypersensitivity): roughly **6–8% in Han Chinese and several other East/Southeast Asian populations**, versus under 1% in Europeans.
- **CYP2C19 poor metabolism** (clopidogrel, PPIs): roughly **12–23% in East Asians**.
- **G6PD deficiency** (antimalarials, sulfonamides): roughly **5–25% across Southeast Asia**.

We treat calibration as design, not post-hoc correction:

1. **Train the prior on the deployed populations** (the contributed commons) so its base rates are local.
2. A **pharmacogenomic overlay** modulates *narration and warnings*, not trajectory tokens: with optional, local-only ethnic/genetic context, guidance is population-aware; without it, guidance degrades to general framing ("higher prevalence in some Asian populations") — less precise, never less safe, and never deterministic ("because you are [ethnicity]…").
3. **Per-country LoRA spokes and bioregional overlays** adapt formulary, guidelines, and disease mix without re-translating strings — "adapt the medicine, don't translate the words."

Because the model reasons over codes, a new language adds nothing to the reasoning layer; only the narration shell localizes (clinical content is English plus the initial deployment-region language where supplied, English-fallback elsewhere).

---

9. Governance and safety invariants

Carried forward from v0.2, with the prior now trained on a patient-contributed commons:

1. **Recommender-only.** The system surfaces what to discuss with a clinician; it never diagnoses, prescribes, or acts. Accept/edit/reject (and thumbs up/down) is the supervision signal.
2. **Abstain by default** — now operationalized as sampled-future variance (Section 5.3). Absence is never papered over; callers never special-case a null or low-confidence prior.
3. **Deterministic guardrail on every string.** Diagnosis/certainty claims are redacted before display, in code, independent of the model.
4. **Offline-first, zero-transmission default.** The on-device path is default; any network path carries ciphertext or a de-identified, code-only prefix.
5. **Retrain, don't redistribute.** Ship weights trained on consented in-region contributed data; never ship a MIMIC-derived checkpoint.

6. **Consent, de-identification, erasure, provenance.** Contribution is opt-in (default off), de-identified, in-region, with k-anonymity over re-identifying trajectories, a right-to-erasure crosswalk, a fixed cross-site salt, and a **provenance record per event**.
7. **Calibrate before trust.** Conformal calibration, subgroup-fairness audit (with explicit attention to DP's disparate utility cost), and held-out-region validation precede any patient-facing risk.
8. **No model collapse.** The trajectory prior trains on coded events, not AI-generated prose; machine-generated imports are detected, tagged, and gated.
9. **Information, not device — by design.** The conversational product sits on the information side of the medical-device line (general guidance, escalation flags, a non-removable disclaimer); any individual named-patient risk output is treated as Software-as-a-Medical-Device and routed to a separately-cleared track.

---

10. Critical path — acquisition and normalization first, model second

Ordered. Do not reorder. (Revised for the patient-contribution model.)

- **P0 — Contribution flywheel + longitudinal reconstitution.** A trajectory model is only as good as its sequences. The gating prerequisite is a working **patient-contribution path** (wallet → permissioned key exchange → de-identified in-region commons) that yields **patient-centric, multi-year, multi-institution** timelines — not visit-level samples, which teach only "admit then discharge."
- **P0 — Multimodal ingestion → MEDS** across the open set of patient-supplied formats (documents, FHIR/HL7 exports, wearables, free text), emitting normalized MEDS events with provenance tags.
- **P1 — Vocabulary harmonization + crosswalk QA** (ICD-9→ICD-10, drugs→ATC, labs→LOINC, procedures→SNOMED/PCS), with a held-out audit of crosswalk accuracy.
- **P1 — Sequence k-anonymity + cross-site salt + synthetic-content gating** (suppress rare long-tail sequences; fixed salt for linkage; detect and tag machine-generated imports).
- **P2 — Tokenizer and value/time/provenance encoders** (per-code value embeddings, continuous time, region and provenance tokens).
- **P2 — Pretrain (next-token + time-to-event) on a single pooled in-region slice** as a smoke test.
- **P3 — Scale-up across contributing populations, then held-out-region validation** as the release gate.
- **P4 — Conformal calibration, SCOPE/REACH inference, ARES-style attribution.**
- **P5 — Distillation to an edge student + per-country LoRA**, wired into `population.ts` and `twin_metric_*`.

The honest first move is still data — but now "data" means **building the contribution flywheel and the multimodal normalizer**, not negotiating institutional pipelines.

---

11. Evaluation plan

We evaluate against the open **Ever Medical Safety Benchmark (EMSB)** rather than knowledge leaderboards (USMLE/MedQA are saturated and irrelevant to safety, grounding, and equity).

Part 1 — safety and comprehension (conversational layer). Emergency-detection recall (target 100% on a physician-curated set), harmful-advice rate (zero critical failures), RAG grounding / hallucination (under 5% out-of-context), pharmacogenomic safety gates (100% on critical variants), lay comprehension (over 90%), and cross-ethnicity / cross-language equity (under 5% variance).

Part 2 — predictive accuracy (trajectory layer). Per-condition AUROC/AUPRC; calibration (Brier + conformal coverage); NRI versus the incumbent clinical score (KFRE for CKD, Framingham/QRISK for CVD); C-index for time-to-event; stratified fairness by **region / ethnicity / age / sex / provenance mix**; and — the multi-region gate — **held-out-region AUC and calibration** (train on contributing populations, validate on an unseen region with no fine-tuning). First target: a **diabetes-to-CKD eGFR-decline model intended to beat KFRE on the local population**.

Part 3 — robustness to self-reported data (new). Performance as a function of **provenance mix** (clinician-attested vs. device-measured vs. self-asserted); degradation under injected label noise and missingness; sensitivity of calibration to **contribution self-selection**; and a **collapse stress test** (confirm the prior does not degrade when synthetic-content fraction rises, given the gating of Section 4.4).

We will open-source EMSB and report the on-device student against frontier baselines (including, where licensing permits, ETHOS-NG-teacher and published-model references). The testable claim: a small, sovereign, population-calibrated, patient-contributed model wins on the safety / grounding / equity / pharmacogenomics axes a generic cloud model is not optimized for, is honest about self-reported provenance, and **generalizes to a region it never saw**.

---

12. Limitations and honest gaps

- **The trajectory model is a contract, not a shipped model.** `population.ts` is a null implementation; `forecast.ts` ships the classical linear floor. Section 11 metrics are targets. The commons is a data flywheel under construction, not a trained model.
- **The flywheel is unproven at scale.** The strategic claim — that non-custodial ownership lowers the trust barrier enough to reach scaling-law-relevant data volumes — is a *hypothesis* about adoption, not a measured result. It may be that contribution rates stay low, in which case the acquisition advantage does not materialize and the model stays small.
- **Self-reported data is signal with an error model, not ground truth.** Provenance conditioning and reweighting reduce but do not remove concordance and self-selection bias.
- **De-identification of trajectories is harder than of telemetry.** A long ordered sequence is strongly re-identifying; active de-identification, k-anonymity, and DP are load-bearing assumptions, not formalities — and they trade off against rare-event fidelity and (for DP) against subgroup utility.
- **Model collapse is a live risk** as AI-drafted records proliferate; our coded-event training and synthetic-content gating are mitigations, not proofs.
- **On-device fidelity gap.** A distilled, quantized edge student is necessarily weaker than the in-region teacher; abstention and the guardrail manage the gap but the trade-off is real, and memory-bandwidth limits bound how much inference-time compute the edge can spend.
- **Held-out-region is non-negotiable for the multi-region claim.** Without it we can defend only "the regions in the training set," not "a model that began in Asia and generalizes."
- **Multilingual i18n is partial.** String externalization across the renderer is in flight; clinical content is bilingual where supplied, English-fallback elsewhere.

13. Ethics and data governance

Pharmacogenomic and ethnic data are local-only, encrypted, optional, and self-reported; population framing replaces deterministic ethnic attribution. Training data for the population prior is opt-in (default off), de-identified, in-region, with a named data owner per tenant, a per-inference audit log, a right-to-erasure crosswalk, and a **per-event provenance record**. Contribution is revocable; the operator is explicitly *not* the ultimate owner of the patient's record. The product is positioned conservatively on the medical-device line, with per-region regulatory review across the relevant national health authorities in each deployment region, plus US FDA clinical-decision-support guidance and the EU AI Act / MDR framework where applicable. No raw patient identifier reaches any model prompt.

14. Conclusion

The 2025–2026 frontier settled two questions. Scaling laws (CoMET; the first EHR scaling studies) showed that generative medical-event models **improve predictably with data**, which makes data access the binding constraint. And simulation-based inference showed that the same sampling that produces a forecast can be made to produce a **calibrated abstention** — the property that makes such a model safe in front of a non-clinician.

This paper takes both seriously. If data is the constraint, the most scalable way to acquire longitudinal, cross-institutional clinical sequences is to let the **patient** — the one party who lawfully holds a record spanning every provider — contribute it, through a decentralized wallet whose **non-custodial ownership is itself the mechanism that lowers the trust barrier and turns contribution into a network effect**. And if simulation is the model, then **inference-time scaling** is the single mechanism that delivers the forecast, the abstention, and the uncertainty interval, and that makes a distilled student viable on the patient's own device.

The constraints that make this hard each become a design principle: retrain-not-redistribute, provenance-conditioned modeling of self-reported data, no-collapse coded-event training, offline-first delivery, active de-identification with DP, MEDS-standardized multimodal ingestion, inference-scaled abstain-and-guardrail, and held-out-region validation. The result is a **sovereign, patient-contributed complement** to the hospital trajectory model: same codes, same evidence base, opposite custody — and a credible path to a population-calibrated, multi-region, provably private, on-device health-forecasting model that **begins in Asia and generalizes beyond it**.

References (selected)

1. Renc P. et al. *Zero-shot health trajectory prediction using transformer (ETHOS)*. npj Digital Medicine, 2024. DOI 10.1038/s41746-024-01235-0. Code MIT (`ipolharvard/ethos-paper`).
2. Renc P. et al. *Foundation model of EMRs for adaptive risk estimation (ARES / `ethos-ares`)*. GigaScience, 2025. arXiv:2502.06124. Built on MEDS.
3. *Generative Medical Event Models Improve with Scale (CoMET / Cosmos Medical Event Transformer)*. Epic Systems, Microsoft Research, Yale School of Medicine, and the Cosmos Governing Council, 2025.

arXiv:2508.12104. (118M patients; 115B events; 151B tokens; 310 health systems; power-law scaling; 78 tasks.)

4. Zhang et al. *Exploring Scaling Laws for EHR Foundation Models.* 2025. arXiv:2505.22964.
5. Sivarajkumar S. et al. *Generative Foundation Model for Structured and Unstructured Electronic Health Records (Generative Deep Patient, GDP).* 2025. arXiv:2508.16054.
6. *Efficient Generative Prediction for EHR Foundation Models: The SCOPE and REACH Estimators.* 2026. arXiv:2602.03730.
7. Shmatko A. et al. *Learning the natural history of human disease with generative transformers (Delphi-2M).* Nature, 2025. (UK Biobank train; Danish-registry held-out validation.)
8. Steinberg E., Shah N. et al. *MOTOR: a time-to-event foundation model for structured medical records.* ICLR 2024. arXiv:2301.03150.
9. Oufattole N. et al. *Medical Event Data Standard (MEDS).* 2024. (github.com/Medical-Event-Data-Standard.)
10. Kraljevic Z. et al. *Foresight — generative pretrained transformer for clinical timelines.* Lancet Digital Health, 2024. arXiv:2212.08072.
11. Yang Z. et al. *TransformEHR.* Nature Communications, 2023.
12. Fallahpour A. et al. *EHRMamba.* 2024. arXiv:2405.14567.
13. McDermott M. et al. *Event Stream GPT.* NeurIPS 2023. arXiv:2306.11547.
14. Gorishniy Y. et al. *On Embeddings for Numerical Features in Tabular Deep Learning.* NeurIPS 2022. arXiv:2203.05556. (Labrador lab-value model: arXiv:2312.11502.)
15. Hu E. et al. *LoRA.* 2021. Dettmers T. et al. *QLoRA.* 2023.
16. Angelopoulos A. et al. *Conformal Risk Control.* ICLR 2024.
17. McMahan B. et al. *Communication-efficient learning of deep networks from decentralized data (FedAvg).* AISTATS 2017.
18. *On-Device LLMs in 2026: What Changed, What Matters, What's Next.* Edge AI and Vision Alliance, 2026. (Model convergence; KV-cache compression; test-time compute on edge.)
19. Chandra V. et al. *On-Device LLMs: State of the Union, 2026.* (StreamingLLM attention sinks; DuoAttention; ChunkKV; KV-cache to ∼3 bits.)
20. *Self-referential training of clinical language models causes collapse* (vocabulary loss, rare-pathology loss, false confidence). medRxiv, 2026.
21. O'Brien E. C. et al. *Concordance between patient-reported health data and electronic health data in the ADAPTABLE trial.* JAMA Cardiology, 2022.
22. *Access to care affects electronic health record reliability and AI-driven disease prediction.* Nature Health, 2026. (All of Us.)
23. *Differential privacy for medical deep learning: methods, tradeoffs, and deployment implications.* 2025–2026. (DP-SGD; membership inference; disparate subgroup utility.)
24. *FedSA-LoRA-DP: Enhancing Privacy and Communication Efficiency in Federated Learning Through Selective Low-Rank Adaptation and Differential Privacy.* 2025.
25. W3C *Decentralized Identifiers (DIDs)* and *Verifiable Credentials Data Model.*
26. Ever Medical Technologies. *Central Trajectory Engine; ETHOS-NG; Multilingual Clinical-AI Plan; Medical Intelligence PRD.* 2026 (`medOS-ultra/docs/architecture/`, `ever-sovereign-wallet/docs/`).