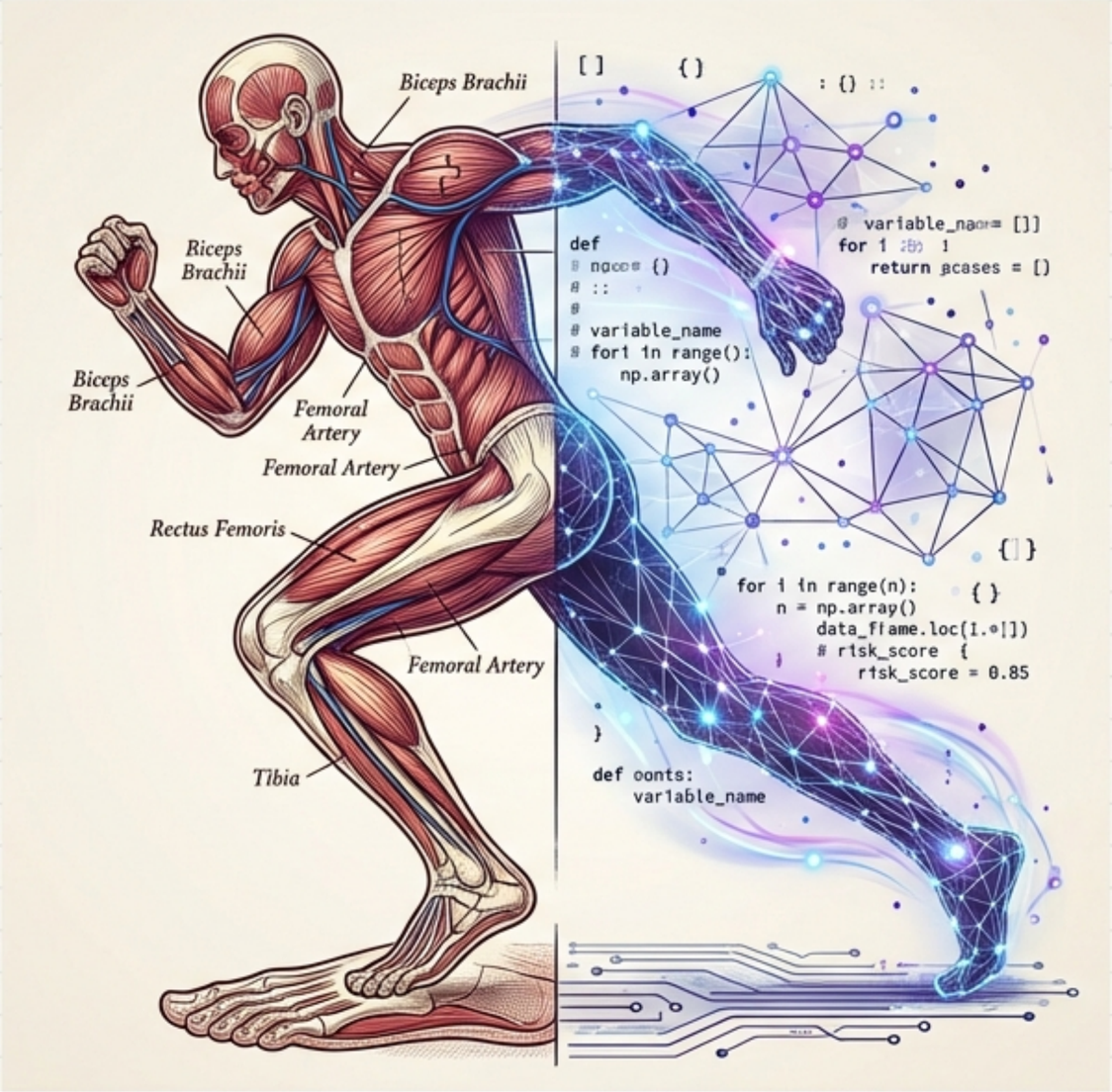
# SynthRun: Bridging Biological Complexity and Synthetic Reality

A code-aligned longitudinal dataset generator for predicting running-related injury and illness risk



# Executive Summary: Synthetic Data Grounded in Epidemiological Truth

## Core Value Proposition

SynthRun is a Python-based generator that creates multi-table datasets (Users, Daily Logs, Activities) to overcome the scarcity of high-resolution longitudinal health data.

**1. Input: Clinical Evidence + Python Rules**

**2. Process: Stochastic Simulation**


**3. Output: Risk-Ready Datasets**

## Key Differentiators

### Glass Box Methodology

Unlike “black box” GANs, SynthRun uses explicit, evidence-based rules derived from peer-reviewed etiology (e.g., Load/Capacity models).

### Specific Use Case


Designed to train “Short-Horizon” (Next 7 Days) prediction models for injury and illness.

### Engineered Imperfection

Explicitly models data “messiness” (missing wearables, dead batteries) to ensure training realism.

# The Problem: The ‘Data Desert’ in Sports Science

Developing robust ‘Early Warning’ systems requires dense, longitudinal data, but three structural barriers exist:

### 1. Scarcity

High-resolution datasets that jointly capture training exposure, wearable physiology, and contextual factors are rare.

### 2. Privacy

GDPR and governance constraints make sharing individual-level health and activity traces difficult or impossible.

### 3. Cost

Assembling cohorts large enough to capture rare injury events takes years and massive funding.

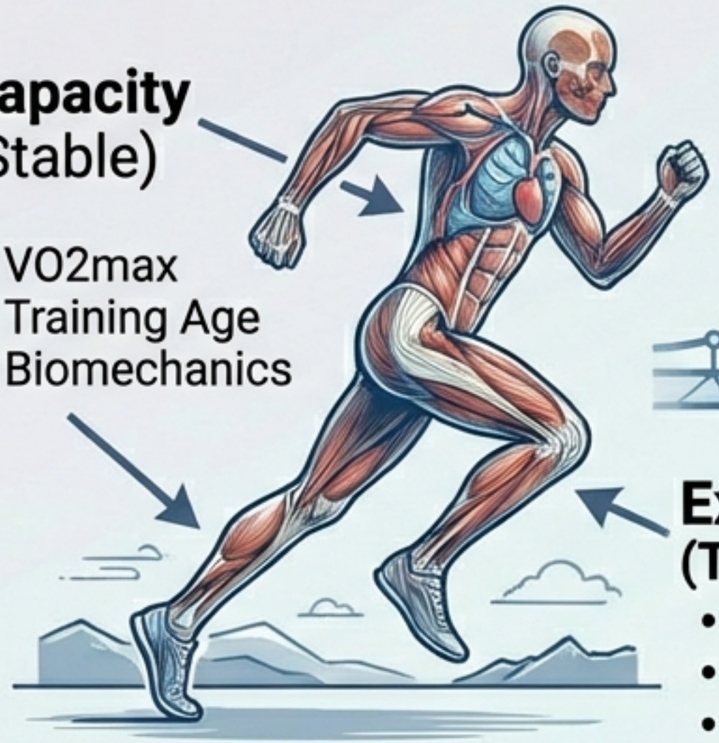
**The SynthRun Solution:** Accelerates model iteration by simulating realistic cohorts for hypothesis testing before final validation on real-world data.

# Architecture Mirrors Etiology: The Exposure-Capacity Model

## **The Science** (Biology)

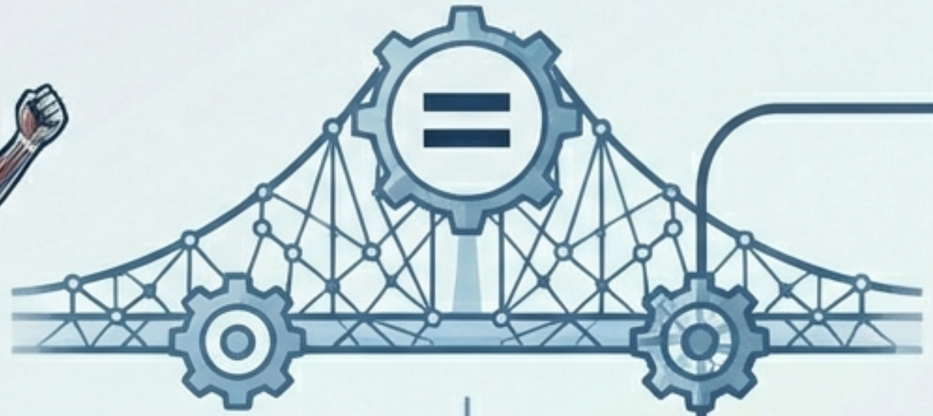
**Capacity**
(Stable)

- VO2max
- Training Age
- Biomechanics



**Exposure**
**(Time-Varying)**

- Distance
- Intensity
- Terrain



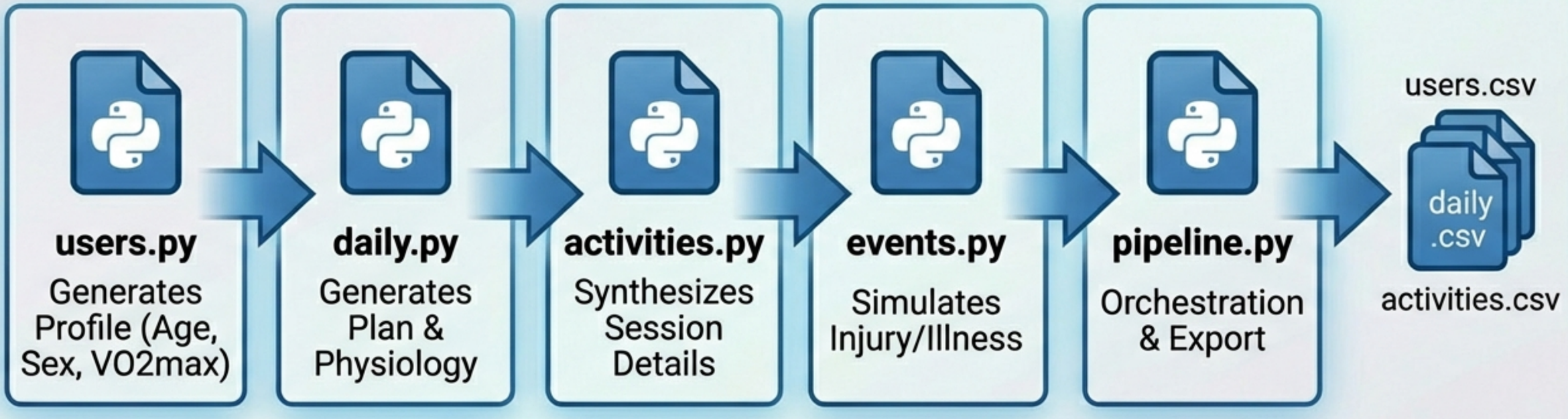
## **The Code** (Schema)

| Users Table | | |
|---|---|---|
| [redacted] | [redacted] | Static profiles |
| | ... | **vo2max** |
| | ... | **injury_proneness** |
| | ... | ... |

| Daily/Activity Tables | | |
|---|---|---|
| [redacted] | [redacted] | Dynamic logs |
| | ... | **km_total** |
| | ... | **elev_gain_m** |
| | ... | ... |

**Scientific Basis:** Injuries are emergent outcomes of the interaction between a runner's Capacity and Exposure (Bertelsen et al., 2017).

# The Generation Pipeline



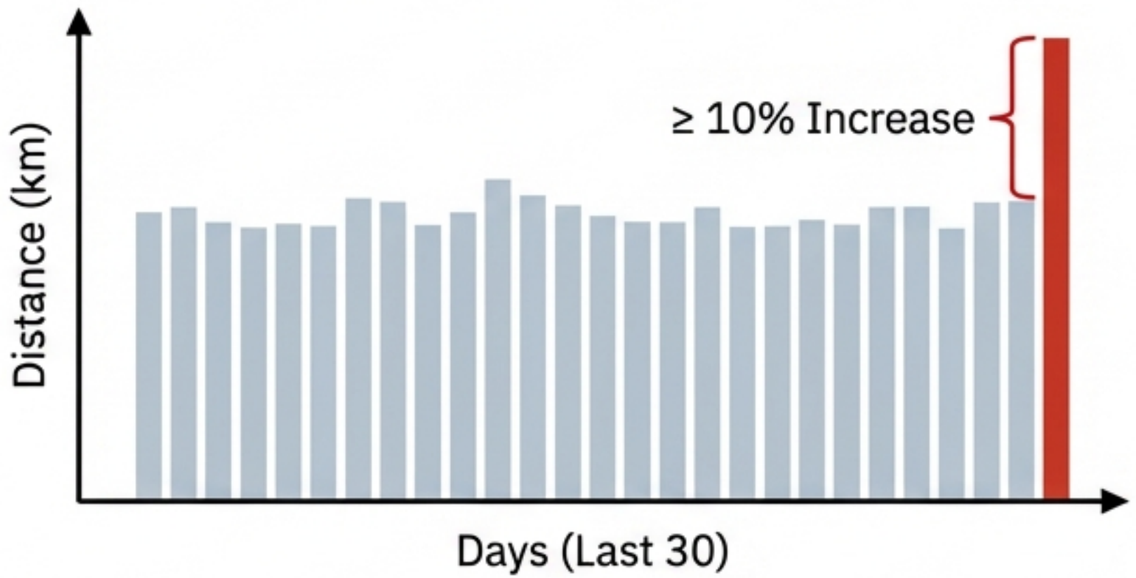
# Feature Focus: Modeling External Load (The ‘Spike’)

## The Evidence

**Source:** Garmin-RUNSAFE Cohort (2025, N=5200)

**Finding:** Running sessions exceeding the runner’s prior 30-day maximum distance by ≥10% are associated with materially elevated injury risk.

**Insight:** Risk is relative to personal history, not arbitrary thresholds.




## The Implementation

**Logic:** Calculate rolling 30-day max distance for every simulated day.

```
def check_spike_risk(today_km, history_30d):
    max_30 = max(history_30d)
    ratio = today_km / max_30
    if ratio >= 1.10:
        return 'HIGH_RISK_SPIKE'
    return 'NORMAL'
```

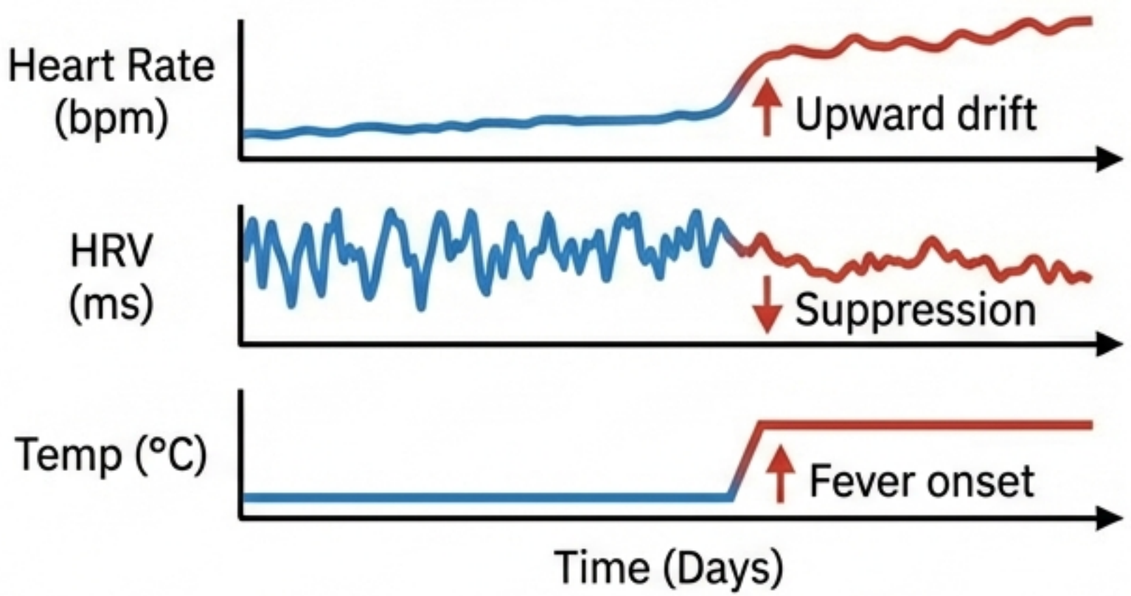
**Result:** Generates ‘long_run_spike_risk’ feature.

# Feature Focus: Internal Load & Illness

## The Evidence

**Consensus:** Acute illness (infection) risk correlates with high load + low recovery.

**Detection:** Relies on deviations from personal baselines in multi-modal signals (HR, HRV, Temp).




## The Implementation

**Logic:** Simulates stable baseline vitals, then injects specific deviations prior to 'illness_onset'.

```
if days_until_illness < 3:
    current_rhr += random.uniform(2, 5)
    current_rhr += 0.85    # Upward drift
    current_hrv *= 0.85    # Suppression
    skin_temp += 0.4       # Fever onset
```

**Goal:** Train multi-modal algorithms to detect pre-symptomatic patterns.

# Secondary Risk Factors: Monotony and Terrain

## 1. Monotony

**Science:** Foster (1998) links high load with low variability to injury incidence.

**Code:** Calculates rolling monotony indices to flag ‘stale’ training blocks.

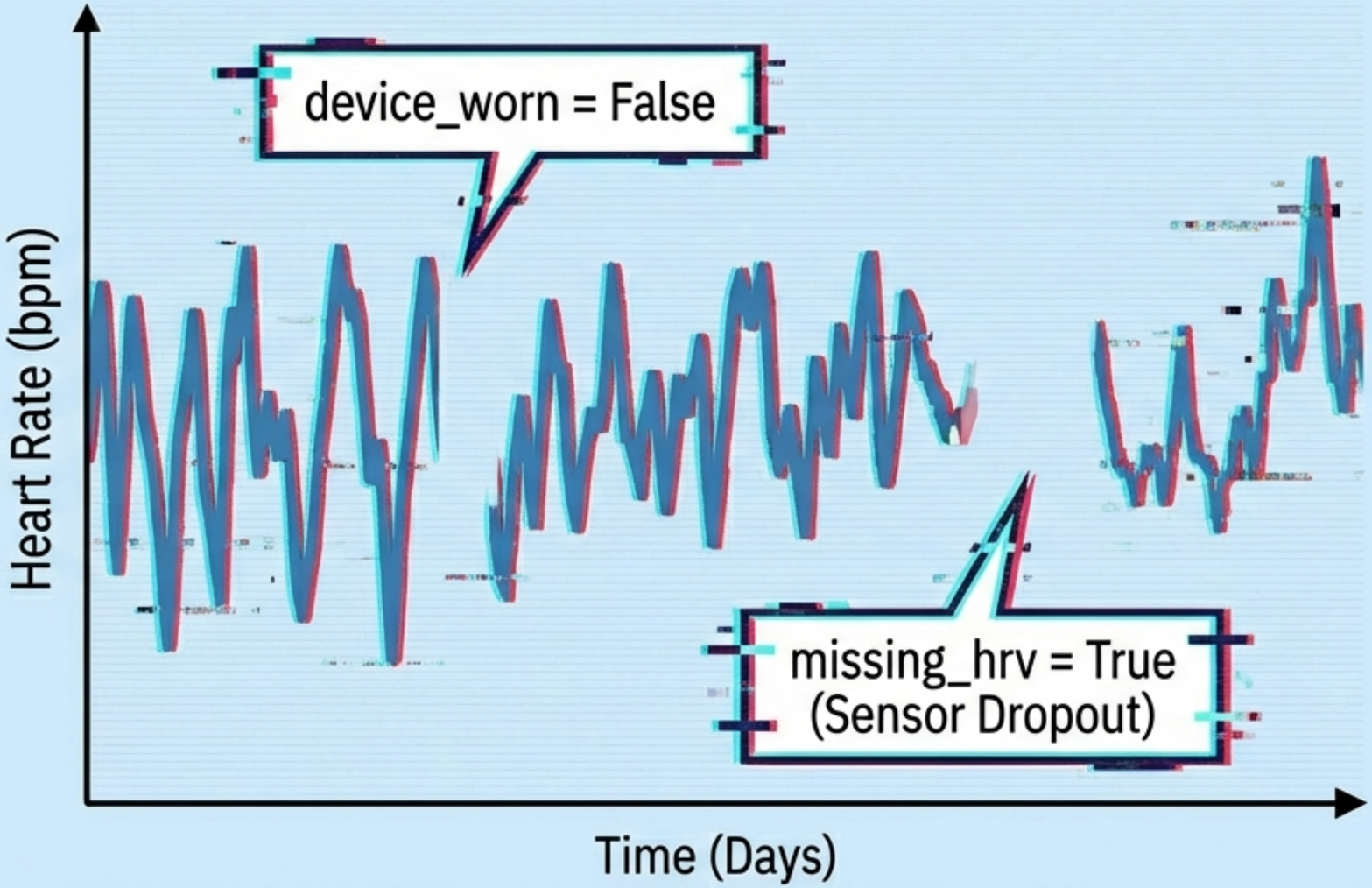

## 2. Terrain

**Science:** Gradient changes biomechanical demand. Hilly runs impose higher internal load.

**Code:** Uses ‘elev_gain_m’ as a proxy for grade exposure.

# **Engineering Realism:** The Necessity of ‘Messy’ Data

Perfect data is useless for training robust models. Real consumers forget to charge watches.




**The Mechanism:**

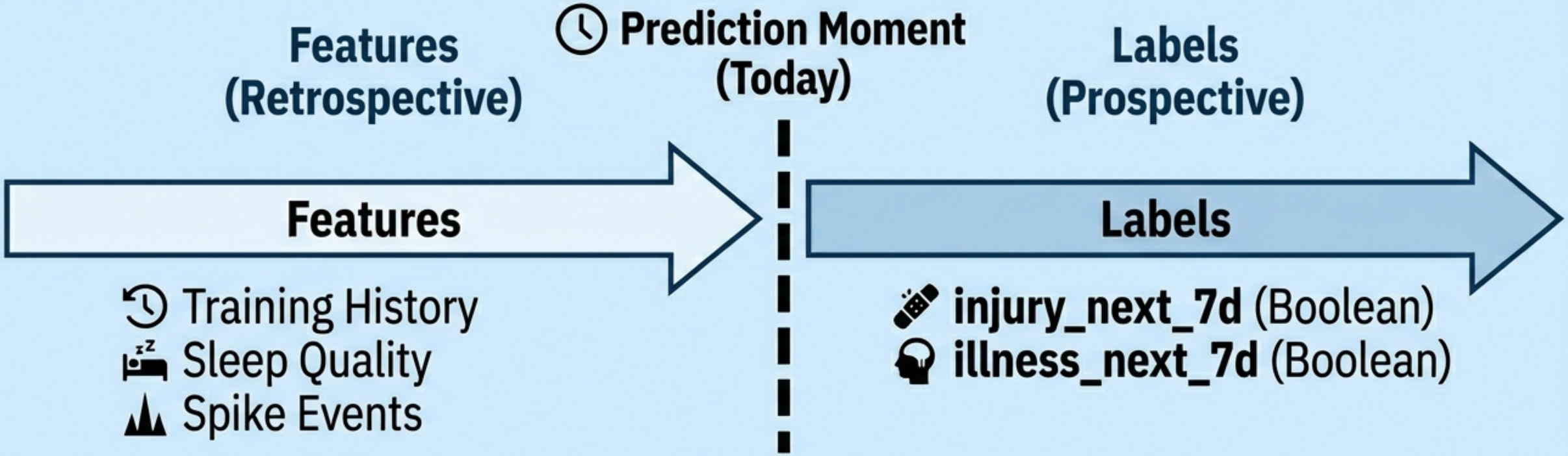
1. **Device Adherence:** Simulates boolean wear status (did they wear it?).
2. **Sensor Dropouts:** Independent missingness probabilities for specific signals.

**Benefit:**

Enables training of models resilient to missing inputs.

# **The Event Engine:** Forward-Looking Labels

Task: ‘Risk-of-Next-Week’ Prediction (Not Diagnosis)




**Leakage Control:** Feature windows are strictly retrospective. Labels are strictly prospective. This simulates a live app environment.

# **Dataset Output:** The Schema

**Users (users.csv)**
- user_id
- vo2max
- injury_proneness
- rest_day_frequency

**Daily (daily.csv)**
- sleep_quality
- readiness_score
- acwr
- spike_absolute_risk
- device_worn

**Activities (activities.csv)**
- pace
- avg_hr
- cadence
- vertical_oscillation
- elev_gain_m

# **Evaluation Framework:** Fidelity, Utility, Privacy

## 1. Fidelity

### Statistical Realism

Do distributions (pace, HRV) and multivariate relationships (load vs. sleep) match real cohorts?

## 2. Utility

### Task Realism

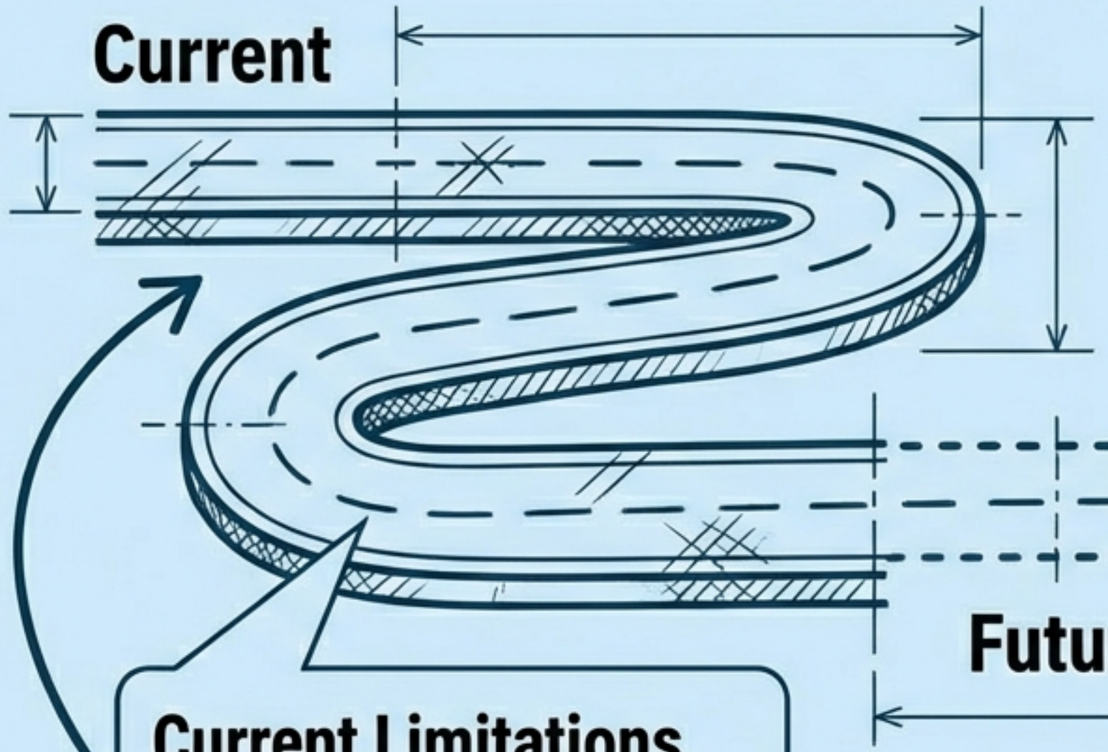
Can models trained on SynthRun accurately predict outcomes when transferred to real-world datasets?

## 3. Privacy

### Risk Assessment

Audit for overfitting. Ensure no ‘Membership Inference’ is possible (synthetic users cannot be traced back to seed data).

# **Limitations & Future** Roadmap




**Current Limitations**

- Terrain is elevation only (no surface type).
- Biomechanics are simplified proxies.

**1. Richer Grade Exposure**
Uphill/Downhill time bins.

**2. Contextual Factors**
Footwear history & strength training.

**3. Interaction Effects**
Life-stress interactions with tissue capacity.

# Accelerating the Path to Safer Running

**Summary:**

SynthRun is a foundational step for consumer health risk management, solving the “cold start” problem for data-hungry algorithms.

**Impact:**

Enables rapid prototyping of “Risk-of-Next-Week” alerts without compromising privacy.


**Better data—even when synthetic—leads to better models, and safer runners.**

# Key Scientific References

**Etiology**
Bertelsen et al. (2017) - A framework for the etiology of running-related injuries. *Scand J Med Sci Sports*.
**The Spike**
Schuster Brandt Frandsen et al. (Garmin-RUNSAFE, 2025) - Identifying high-risk running sessions. *Br J Sports Med*.
**Load**
Soligard et al. (2016) - IOC Consensus statement on load in sport and risk of injury/illness. *Br J Sports Med*.
**Monotony**
Foster (1998) - Monitoring training in athletes with reference to overtraining syndrome. *Med Sci Sports Exerc*.
**Physiology**
Radin et al. (2020) - Harnessing wearable device data to improve influenza-like illness surveillance. *Lancet Digit Health*.