SynthRun: a code-driven synthetic longitudinal runner cohort for injury and illness risk modelling and benchmarking

Authors: Simon Walsh
Affiliations: StrideWise Research & Development
Manuscript type: Original Paper (target: JMIR mHealth and uHealth)
Submitted: February 2026

Abstract

Background: Large, labelled longitudinal datasets that link running exposure, recovery and physiology are essential for developing and validating injury-risk models, yet are difficult to obtain at scale because consumer wearable records are personal data and injury labels are sensitive and inconsistently captured.

Objective: To introduce SynthRun, a synthetic cohort generator that simulates runners as individuals and produces multi-table smartwatch-like longitudinal data for injury/illness risk modelling and reproducible benchmarking.

Methods: SynthRun is a transparent, code-driven simulator. Training histories are generated using periodised plans (build/peak/recovery/base phases, down-weeks, and realistic session-type distributions). Physiological and recovery channels are generated as lagged responses to training and rest, with explicit missingness reflecting device non-wear. Injury/illness episodes arise from an auditable causation layer informed by load-capacity concepts and workload-change evidence. Labels are forward-looking (injury_next_7d, illness_next_7d). A deterministic representation bridge converts SynthRun into the published 7-day event-window format used by Lövdal et al., enabling indirect benchmarking when rich real-world smartwatch datasets are unavailable.

Results: In preliminary experiments, synthetic event-window data achieved 86.6% of the ROC AUC obtained on the real benchmark dataset while matching injury prevalence, supporting fidelity in the workload-driven signal space despite information loss during conversion.

Conclusions: SynthRun provides a reproducible substrate for developing and stress-testing injury-risk pipelines under realistic consumer constraints (missingness, heterogeneous training behaviour) while supporting indirect validation against a published benchmark. We propose an evaluation framework spanning fidelity, downstream utility, and privacy/disclosure risk.

Keywords

synthetic data; running; wearable sensors; training load; injury risk; time series; simulation; privacy; benchmarking

Introduction

Running-related injury and acute illness are common disruptions that reduce training consistency, impair performance progression and increase dropout from running programmes [1–4]. From a modelling perspective, the most informative predictors are often high-resolution, temporally ordered signals—training exposure (volume and intensity), workload change (spikes, ramps), recovery (sleep and subjective readiness), and wearable-derived physiology (resting heart rate, heart-rate variability, respiration, temperature proxies). However, assembling large cohorts that combine these signals with reliable onset labels is challenging: consumer wearable data are privacy-sensitive, platform-dependent, and typically governed by consent and retention constraints; injury outcomes are rare, definition-dependent, and under-reported.

SynthRun addresses this bottleneck by generating a realistic, labelled synthetic runner cohort designed for model development, stress-testing, and reproducible benchmarking. The core design choice is transparency. Rather than learning an opaque generative model from private records, SynthRun encodes plausible mechanisms from sports science and wearables research into a modular simulator: runners have heterogeneous traits, training is periodised, physiology responds to training and recovery with realistic lags, missingness reflects device wear behaviour, and injury/illness episodes are generated by a causation layer that can be audited and tuned.

This paper makes four contributions. First, we describe a code-aligned synthetic data generator that outputs a multi-table longitudinal dataset resembling consumer wearable ecosystems. Second, we document the specific feature families and coupling mechanisms implemented in code, with emphasis on variables commonly used in injury-risk and fatigue monitoring [6–9,13–15]. Third, we formalise forward-looking labels and episode-aware event construction suited to deployment. Fourth, we provide a representation bridge that converts SynthRun into the widely used event-window format released with the competitive runner injury dataset and modelling approach by Lövdal et al. [11,36], enabling indirect realism checks when rich real-world wearables are unavailable.

Scientific background and related work

Running injuries are commonly framed through a load–capacity lens: risk increases when training load rises faster than tissue adaptation, and is modified by recovery, previous injury, and individual susceptibility [5–9]. Systematic reviews identify workload history, rapid volume increases, and high mechanical stress exposures as recurrent correlates, while also emphasising heterogeneity and measurement limitations [1–4]. Large prospective cohorts such as Garmin-RUNSAFE provide evidence on session-level risk, including higher risk associated with unusually long sessions relative to an individual's recent maximum and with heavy-intensity sessions when combined with volume [10,34]. Machine-learning approaches have been applied to runner training logs, including the competitive runner dataset and "day-approach" representation released by Lövdal et al. [11,36].

Wearable sensors extend exposure logs by providing recovery and physiology signals. Heart-rate derived load and recovery markers (e.g., resting heart rate and HRV) are often used to monitor training status and adaptation, with evidence that HRV is suppressed after high load and may rebound during recovery [18,19]. Wearables have also been evaluated for detection of infectious illness, including studies using symptom-linked wearable features for COVID-19 detection and influenza-like illness surveillance [21–24]. These channels are attractive for consumer injury-risk systems because they are passively captured, but they are also incomplete and noisy in real-world use.

Synthetic data has become a practical strategy for development and evaluation when real data access is limited. In health contexts, guidance stresses that synthetic data is not automatically “safe” or “valid”: it must be evaluated for fidelity and downstream utility, and audited for disclosure risks such as identity disclosure and membership inference [27–33]. SynthRun adopts this perspective by (i) making its assumptions explicit, (ii) supporting multiple fidelity and utility checks, and (iii) providing an indirect validation pathway through a published benchmark representation.

Methods

Table 1. Primary SynthRun dataset outputs.

Table	Grain	Examples
users.csv	Runner (static)	Age, sex, anthropometrics, baseline fitness, injury_proneness/resilience, device wear propensity
activities.csv	Session	Distance, duration, pace, intensity-zone km, elevation gain, HR proxy
daily.csv	Runner-day	Aggregated load, rolling features (acute/chronic), physiology/recovery, missingness flags, labels (injury_next_7d)

Code-aligned generator architecture and dataset outputs

SynthRun is organised as a modular pipeline that generates runners, training plans, sessions, daily aggregates and event labels in a reproducible order. At a high level, the generator proceeds through:
(1) runner instantiation (stable traits, device wear propensity and risk traits);
(2) training plan generation (periodised weeks with down-weeks and session-type templates);
(3) session instantiation (distance, duration, intensity distribution, elevation and optional gait proxies);
(4) daily aggregation and rolling features (acute/chronic load, monotony, spike constructs);
(5) physiology and perception simulation (sleep, readiness, RHR, HRV, respiration, temperature proxies with lagged load/recovery coupling);
(6) injury and illness simulation (episode state machine and forward-looking labels).

The generator persists three primary tables—users.csv (runner profile), daily.csv (runner-day panel), and activities.csv (session detail)—linked by user_id and date, mirroring common wearable platform data organisation. The daily table is the primary modelling substrate and includes both raw and engineered features plus labels (injury_next_7d, illness_next_7d). Importantly, missingness is explicit: device_worn and per-signal missingness indicators are generated and propagated into downstream feature engineering.

Training plans are generated using a weekly periodisation structure (build, peak, recovery and base phases) with realistic modulation of volume and intensity and with athlete-to-athlete variability. Plan structure parameters are informed by distributions observed in the external competitive runner dataset (e.g., typical cycle length and down-week frequency), but remain configurable so users can stress-test alternative training cultures and risk regimes. Session templates include easy runs, long runs, tempo sessions, interval sessions and sprint exposure. Weekends preferentially host long runs and harder sessions are typically followed by recovery days, producing realistic within-week autocorrelation in exposure.

Training load is computed using a TRIMP-style formulation combining duration with an intensity factor derived from heart-rate reserve proxies and intensity-zone composition [35]. This provides a single load proxy that supports acute–chronic dynamics and monotony measures [13–15] while retaining separate distance-based intensity decompositions used in external benchmarks [11].

Feature families, mechanisms, and how the code generates them

SynthRun's daily representation is designed around families of predictors that recur in injury-risk and fatigue monitoring: stable runner traits, external load and its changes, recovery/physiology, and contextual modifiers. The implementation prioritises plausible ranges, realistic temporal autocorrelation, and cross-signal coupling, rather than generating each channel independently.

Stable traits (users.csv) include sex, age and anthropometrics; baseline weekly volume and long-run fraction; baseline aerobic fitness (mapped to VO_2max and easy-pace ranges); rest-day frequency; and two explicit risk traits: injury_proneness and injury_resilience. These traits mediate both training behaviour and the injury hazard in the causation layer.

External load is generated first at the session level (activities.csv) and then aggregated to day level. Each session yields distance_km, duration_min, pace, average HR proxy, elevation gain, and a distance-based intensity decomposition (kms_z3_4, kms_z5_t1_t2, kms_sprinting). These decompositions are designed to support both rich modelling and lossless translation into the Lövdal "day-approach" feature space [11].

Workload-change descriptors are then derived from the daily series. The implementation includes acute and chronic load windows (e.g., 7-day and 28-day), an ACWR-style ratio, training monotony, hard-session density over recent windows, and spike constructs that capture unusually large long runs relative to an athlete's recent maximum, consistent with RUNSAFE findings on high-risk sessions [10].

Recovery and physiology channels are generated as lagged responses to load and recovery rather than as independent noise. For example, HRV is simulated around an athlete-specific baseline with next-day suppression after high load and rebound during recovery weeks; resting heart rate rises with acute stress and can remain elevated under cumulative fatigue; sleep is influenced by both stress and training intensity; and subjective perceptions (perceived recovery, training success and exertion) reflect both the planned session difficulty and the underlying fatigue state [18,19].

Terrain exposure is represented via elevation gain at the session and day level, providing a proxy for mechanical and metabolic stress. Energetic cost varies substantially with grade, and uphill/downhill running alters biomechanics and physiological load; incorporating elevation thus improves realism for hill-dominant training contexts [25,26].

Finally, missingness is modelled explicitly. The generator simulates device wear propensity per runner (wear_rate) and emits a daily device_worn flag plus per-signal missingness indicators (e.g., missing_hrv, missing_sleep). Missingness can be correlated with behavioural disruption (e.g., illness or injury periods) and is intended as a first-class modelling signal for consumer settings where data are intermittently absent.

Injury and illness event simulation and forward-looking labels

SynthRun generates injury and illness as episodes rather than isolated labels. Each runner carries a latent state (healthy, onset, ongoing, recovery/return), allowing multi-day episodes and realistic periods of reduced training. Episode-aware generation is important because naïvely labelling every day of an injury period as a positive case can inflate apparent model performance and does not match product deployment, which primarily aims to predict onset.

Daily injury probability begins with a profile-dependent baseline hazard (higher in novices and recreational runners than in advanced/elite profiles) and is then modified by interpretable drivers capturing the best-supported mechanisms in the running injury literature: rapid workload increases, cumulative fatigue, insufficient recovery, and high mechanical-stress exposure [5–9]. In code, these drivers combine multiplicatively as risk multipliers and additively as "shock" risks for particularly hazardous exposures (e.g., sprinting and extreme long-run spikes). A simplified representation of the implemented logic is:

$$\mathrm{p\_injury}(t) = \mathrm{clip}\big(\mathrm{base\_hazard}(\mathrm{profile}) \times \mathrm{M\_load}(t) \times \mathrm{M\_fatigue}(t) \times \mathrm{M\_recovery}(t) \times (1 - \mathrm{injury\_resilience}) \times (1 + \mathrm{injury\_proneness}) + \mathrm{R\_spike}(t) + \mathrm{R\_sprint}(t)\ ,\ 0,\ \mathrm{p\_max} \big)$$

where M_load is a function of acute/chronic load and ramp behaviour; M_fatigue responds to recent hard-session density and monotony; M_recovery is driven by deviations in sleep and physiology from individual baselines; and R_spike/R_sprint represent additive risks derived from long-run spike categorisation and sprinting distance exposure. Long-run spike categories are computed relative to an athlete's rolling maximum long run and aligned with cohort evidence that exceptionally long sessions elevate injury risk [10].

When an injury onset occurs, the simulator samples a severity and assigns a recovery duration; training behaviour during the episode is modified (reduced volume, fewer hard sessions, and increased rest days), and physiology reflects elevated stress and reduced readiness. Illness episodes are generated analogously, with probability driven more heavily by physiology and sleep disruption and with shorter typical episode lengths.

Labels are forward-looking to match operational decision-making. For each day t, injury_next_7d(t)=1 if an injury onset occurs on any day in (t+1 … t+7); otherwise 0. The same convention is used for illness_next_7d. All predictor features at day t are computed using only history up to t, and model evaluation is intended to use forward-time splits per runner to prevent temporal leakage.

For indirect validation against the competitive runner benchmark, SynthRun also supports extraction of onset-anchored events with a 21-day exclusion window and deterministic construction of 7-day history vectors using the Lövdal suffix convention [11,36]. This creates a testable bridge between rich smartwatch-style synthetic data and the published event-window feature space.

Evaluating fidelity, utility, and privacy risk

Synthetic datasets intended for model development should be evaluated on three axes: fidelity (do distributions and dynamics resemble plausible real data?), utility (do models trained or tested on the synthetic data behave meaningfully?), and privacy/disclosure risk (does the synthetic data leak sensitive information about individuals, or provide false assurance of anonymity?) [27–33].

Fidelity. We recommend reporting (i) marginal plausibility checks (ranges for distance, pace, HR, sleep, HRV); (ii) multivariate similarity (correlation structure between load and physiology); (iii) temporal dynamics (autocorrelation, seasonality, recovery-week rebounds); and (iv) event realism (injury prevalence, episode length distributions, and onset clustering guards).

Utility. Downstream utility can be evaluated via (i) model discrimination and calibration on forward-time splits; (ii) sensitivity analyses showing risk increases under intended driver combinations (e.g., acute spikes with degraded sleep); and (iii) robustness to missingness and device non-wear. Importantly, utility should be assessed in the feature space that will be used in deployment, and in any reduced representations used for benchmarking.

Indirect benchmark validation (worked example). Because direct smartwatch-level benchmarks are often inaccessible, we evaluate SynthRun via a representation bridge to the Lövdal day-approach event format. First, we replicate the published standalone modelling approach on the real DataverseNL dataset, achieving ROC AUC 0.7121 compared to the paper's 0.724 ($\Delta = -0.0119$), indicating a close implementation match (see system documentation). We then convert SynthRun into the same event-window representation and score it using the model trained on real data. On the synthetic event dataset, the benchmark model achieves ROC AUC 0.6167 (86.6% of real-data AUC) with injury prevalence ~1.5% and PR AUC 0.0216. This result is conservative because the conversion discards most physiology and recovery channels (not present in the benchmark format); the primary purpose is therefore to validate that workload-driven signal patterns are plausible and generalise under a published feature representation.

Privacy/disclosure risk. Although SynthRun is not derived from a private cohort, synthetic health data can still carry disclosure and misuse risks, and "synthetic" does not imply "anonymous" in general. We therefore recommend documenting (i) whether synthetic records are learned from real individuals or purely simulated; (ii) whether any identifiers or quasi-identifiers enable singling-out; and (iii) attacks and audits relevant to the generation method, such as identity disclosure risk estimates and membership inference tests [29–31]. Broader guidance emphasises explicit trade-offs between fidelity and privacy/utility and warns against over-claiming safety from synthetic generation alone [32,33].

Discussion and future extensions

SynthRun is designed to be useful under the practical constraints that block many injury-risk projects: limited access to large, labelled wearable cohorts and difficulty benchmarking rich data representations against public datasets. The generator's transparency is a deliberate choice. Mechanistic and semi-mechanistic simulation makes assumptions inspectable, supports targeted stress-testing (e.g., spikes, down-weeks, illness), and allows developers to test whether models respond to intended drivers rather than artefacts.

There are important limitations. First, SynthRun is not a substitute for prospective cohort data, and it should not be used to make clinical or causal claims about real injury risk. Parameter choices (e.g., baseline hazards and sensitivity multipliers) are heuristics intended to produce plausible prevalence and relationships rather than estimated causal effects. Second, the generator necessarily omits many determinants of injury and illness (surface, footwear, biomechanics beyond simple proxies, prior injury history, psychosocial stressors, and context). Third, indirect validation via reduced representations is useful but incomplete: matching workload-driven signals in the Lövdal feature space does not guarantee realism in the richer physiology space.

Future work will prioritise (i) calibration of simulator parameters to prospective cohorts under informed consent; (ii) richer context channels (surface, weather, travel) and biomechanics proxies; (iii) alternative injury definitions and delayed/ noisy reporting to better mirror consumer self-report; and (iv) privacy-enhancing extensions (e.g., formal differential privacy for any future components trained on real data) alongside stronger disclosure audits.

Despite these limitations, SynthRun provides a reproducible substrate for developing end-to-end injury-risk pipelines, evaluating robustness to missingness, and producing testable hypotheses for subsequent real-world validation.

Conclusion

We introduced SynthRun, a code-driven synthetic runner cohort generator producing smartwatch-like longitudinal data with forward-looking injury and illness labels. The system is transparent by design: training is periodised, physiology responds to load and recovery with realistic lags, missingness reflects device wear, and episodes arise from an explicit causation layer aligned with load–capacity concepts. To enable indirect validation, SynthRun includes a deterministic conversion to the published competitive runner event-window representation, and preliminary benchmarking suggests plausible workload-driven signal patterns. SynthRun is intended to accelerate reproducible model development and benchmarking while enabling careful, staged claims that can later be disciplined by prospective real-world studies.

Appendix A. Dataset schema (representative fields)

Users table (users.csv)
The generator outputs this table with the following implemented columns (sample shown for readability). The full column list is provided below.

Sample fields: user_id, sex, age, height_cm, weight_kg

Full columns:

user_id, sex, age, height_cm, weight_kg, base_km_week, long_run_frac, fitness, vo2max, vo2max_adjusted, profile, rhr_base, hrv_base, hrmax, hrr, wear_rate, pace_easy_minpkm, injury_proneness, injury_resilience, rest_day_frequency, never_sprints

Daily table (daily.csv)
The generator outputs this table with the following implemented columns (sample shown for readability). The full column list is provided below.

Sample fields: user_id, date, duration_min, elev_gain_m, sleep_hours, sleep_quality, readiness_score, rhr_bpm, resp_rate_rpm, device_worn, injury_onset, injury_next_7d, illness_onset, illness_next_7d

Full columns:

user_id, date, session_type, km_total, has_double, double_session_type, double_km, is_break_week, sessions, duration_min, avg_hr_bpm, pace_min_per_km, elev_gain_m, kms_z3_4, kms_z5_t1_t2, kms_sprinting, cadence_spm, gct_ms, stride_length_cm, vertical_oscillation_cm, gct_balance, training_load, device_worn, missing_hrv, missing_rhr, missing_sleep, missing_stress, missing_resp, missing_temp, wear_7d_rate, rhr_bpm, hrv_ms, sleep_hours, stress_score, resp_rate_rpm, skin_temp_c, strength_training, hours_alternative, perceived_exertion, perceived_trainingSuccess, perceived_recovery, hrv_rmssd, hrv_sdnn, hrv_pnn50, sleep_efficiency, sleep_quality, deep_sleep_hours, load_acute7, load_chronic28, acwr, training_monotony, readiness_score, fitness, injury_proneness, injury_resilience, rest_day_frequency, illness_onset, illness_ongoing, injury_onset, injury_ongoing, injury_severity, recovery_duration, return_period_duration, days_since_recovery_end, injury_next_7d, illness_next_7d, long_run_spike_risk, long_run_spike_category, spike_absolute_risk, sprinting_absolute_risk, has_long_run_spike, had_spike_last_7d, consecutive_hard_days, consecutive_hard_risk, has_2plus_consecutive_hard, has_3plus_consecutive_hard, hard_sessions_count_7d, high_intensity_count_risk, has_3plus_hard_sessions_7d, has_4plus_hard_sessions_7d, has_5plus_hard_sessions_7d, volume_multiplier, intensity_multiplier

Activities table (activities.csv)
The generator outputs this table with the following implemented columns (sample shown for readability). The full column list is provided below.

Sample fields: user_id, date, distance_km, duration_min, avg_hr_bpm, elev_gain_m

Full columns:

activity_id, user_id, date, session_type, distance_km, duration_min, pace_min_per_km, avg_hr_bpm, elev_gain_m, kms_z3_4, kms_z5_t1_t2, kms_sprinting, cadence_spm, gct_ms, stride_length_cm, vertical_oscillation_cm, gct_balance

Abbreviations

ACWR: acute:chronic workload ratio

AUC: area under the receiver operating characteristic curve

HRV: heart rate variability

RHR: resting heart rate

TRIMP: training impulse

Conflicts of Interest

The authors are affiliated with StrideWise, which is developing products related to the SynthRun generator.

References

[1] Saragiotto BT, Yamato TP, Hespanhol Junior LC, Rainbow MJ, Davis IS, Lopes AD. What are the main risk factors for running-related injuries? Sports Medicine. 2014.

[2] van der Worp MP, ten Haaf DS, van Cingel R, de Wijer A, Nijhuis-van der Sanden MWG, Staal JB. Injuries in runners; a systematic review on risk factors and sex differences. PLOS ONE. 2015.

[3] van Poppel D, Scholten-Peeters GGM, van Middelkoop M, Verhagen AP. Risk factors for overuse injuries in short- and long-distance running: a systematic review. Journal of Sport and Health Science. 2021.

[4] Kakouris N, Yener N, Fong DTP. A systematic review of running-related musculoskeletal injuries. Journal of Sport and Health Science. 2021.

[5] Bertelsen ML, Hulme A, Petersen J, et al. A framework for the etiology of running-related injuries. Scandinavian Journal of Medicine & Science in Sports. 2017.

[6] Soligard T, Schwellnus M, Alonso JM, et al. How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury. British Journal of Sports Medicine. 2016;50(17):1030–1041.

[7] Schwellnus M, Soligard T, Alonso JM, et al. How much is too much? (Part 2) International Olympic Committee consensus statement on load in sport and risk of illness. British Journal of Sports Medicine. 2016;50(17):1043–1052.

[8] Gabbett TJ. The training–injury prevention paradox: should athletes be training smarter and harder? British Journal of Sports Medicine. 2016.

[9] Impellizzeri FM, Tenan MS, Kempton T, Novak A, Coutts AJ. Acute:Chronic Workload Ratio: Conceptual Issues and Fundamental Pitfalls. International Journal of Sports Physiology and Performance. 2020.

[10] Schuster Brandt Frandsen J, Hulme A, Parner ET, et al. How much running is too much? Identifying high-risk running sessions in a 5200-person cohort study. British Journal of Sports Medicine. 2025;59(17):1203–1210. doi:10.1136/bjsports-2024-109380.

[11] Lövdal SS, et al. Injury Prediction in Competitive Runners With Machine Learning. International Journal of Sports Physiology and Performance. 2021.

[12] Fokkema T, et al. Reasons and predictors of discontinuation of running after a running program in novice runners. Journal of Science and Medicine in Sport. 2019.

[13] Halson SL. Monitoring training load to understand fatigue in athletes. Sports Medicine. 2014;44(Suppl 2):139–147.

[14] Foster C, Florhaug JA, Franklin J, et al. A new approach to monitoring exercise training. Journal of Strength and Conditioning Research. 2001.

[15] Foster C. Monitoring training in athletes with reference to overtraining syndrome. Medicine & Science in Sports & Exercise. 1998.

[16] Tanaka H, Monahan KD, Seals DR. Age-predicted maximal heart rate revisited. Journal of the American College of Cardiology. 2001;37(1):153–156.

[17] Joyner MJ, Coyle EF. Endurance exercise performance: the physiology of champions. The Journal of Physiology. 2008;586(1):35–44.

[18] Buchheit M. Monitoring training status with HR measures: do all roads lead to Rome? Frontiers in Physiology. 2014.

[19] Plews DJ, Laursen PB, Stanley J, Kilding AE, Buchheit M. Training adaptation and heart rate variability in elite endurance athletes: opening the door to effective monitoring. Sports Medicine. 2013;43(9):773–781.

[20] Van Hooren B, Goudsmit J, Restrepo J, Vos S. Real-time feedback by wearables in running: current approaches, challenges and suggestions for improvements. Journal of Sports Sciences. 2020;38(2):214–230.

[21] Quer G, Radin JM, Gadaleta M, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. Nature Medicine. 2021.

[22] Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level influenza-like illness surveillance. The Lancet Digital Health. 2020.

[23] Mitratza M, et al. Wearables and the detection of infectious diseases: a systematic review. The Lancet Digital Health. 2022.

[24] Mason AE, et al. Detection of COVID-19 using multimodal data from a wearable device (TemPredict/Oura). Scientific Reports. 2022.

[25] Minetti AE, Ardigo LP, Saibene F. Energy cost of walking and running at extreme uphill and downhill slopes. Journal of Applied Physiology. 2002.

[26] Vernillo G, Savoldelli A, Zignoli A, et al. Influence of uphill and downhill running on biomechanics and physiology: a review. Sports Medicine. 2017.

[27] Goncalves A, et al. Generation and evaluation of synthetic patient data. BMC Medical Research Methodology. 2020.

[28] Chen RJ, et al. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering. 2021.

[29] El Emam K, Mosquera L, et al. Evaluating identity disclosure risk in fully synthetic health data. Journal of Medical Internet Research. 2020.

[30] Zhang Z, et al. Membership inference attacks against synthetic health data. Journal of Biomedical Informatics. 2022.

[31] Stadler T, et al. Synthetic data – anonymisation groundhog day. USENIX Security Symposium. 2022.

[32] Royal Society. Synthetic data – what, why and how? Survey / policy report. 2022.

[33] Adams T, et al. On the fidelity versus privacy and utility trade-off of synthetic data. 2025.

[34] Nielsen RØ, Bertelsen ML, Ramskov D, Damsted C, Brund RK, Parner ET, Sørensen H, Rasmussen S, Kjærgaard S. The Garmin-RUNSAFE Running Health Study on the aetiology of running-related injuries: rationale and design of an 18-month prospective cohort study including runners worldwide. BMJ Open. 2019;9(9):e032627. doi:10.1136/bmjopen-2019-032627.

[35] Banister EW. Modeling elite athletic performance. In: MacDougall JD, Wenger HA, Green HJ, editors. Physiological Testing of the High-Performance Athlete (2nd ed.). Champaign, IL: Human Kinetics; 1991. p. 403–424.

[36] Lövdal SS, den Hartigh RJR, Azzopardi G. Replication Data for: Injury Prediction in Competitive Runners With Machine Learning. DataverseNL. 2021. doi:10.34894/UWU9PV.