Predicting Lifestyle-Related Health Risks in Adolescents Using Biological and Behavioural Data: A Critical Review and Conceptual Framework for Early, Data-Driven Risk Prediction in Youth (Ages 10-19)

Saanvi Bhambhani[1], Raghu Raja Mehra[2]

[1]Department of Information Technology, DAV Public School, Amritsar, Punjab, India
Email: *bhambhanisaanvi14[at]gmail.com*

[2]Department of Information Technology, Amritsar

Abstract: ***Adolescence (ages 10–19) is a formative window during which transient behaviours consolidate into enduring health trajectories, making it a decisive stage for preventing non-communicable disease. This paper examines how the integration of biological indicators- body mass index, resting blood pressure, heart rate, sleep duration, and family medical history- with behavioural signals such as physical activity, dietary quality, screen and social-media use, substance exposure, and perceived stress can enable the early prediction of lifestyle-related health risks through predictive analytics, artificial intelligence (AI), and machine learning (ML). Drawing on a structured narrative synthesis of literature published between 2019 and 2025, the study advances a biobehavioural predictive framework and critically evaluates supervised learning approaches- logistic regression, ensemble tree methods, and neural networks- for adolescent risk stratification. The analysis contends that although ensemble and deep-learning models achieve superior discrimination (reported AUCs of 0.80–0.92), their genuine clinical value depends on data quality, interpretability, equity, and ethical safeguards rather than headline accuracy. The paper concludes that responsibly governed, explainable models embedded within school and primary-care ecosystems represent the most credible route toward preventive adolescent health.***

Keywords: adolescent health; predictive analytics; machine learning; lifestyle risk; behavioural data; preventive medicine

1. Introduction

The World Health Organization estimates that more than 1.2 billion people worldwide are adolescents, and that patterns of physical activity, diet, sleep, and substance use established during this decade shape the risk of cardiovascular disease, type 2 diabetes, and mental-health disorders across the entire life course (World Health Organization, 2021). Because the majority of premature non-communicable disease burden is attributable to modifiable lifestyle exposures that begin in youth, adolescence represents an unusually high-leverage point for prevention. Intervening before pathology becomes entrenched is both clinically and economically preferable to treating established disease decades later.

In parallel, the volume and granularity of data describing young people has expanded dramatically. Wearable devices capture continuous heart rate, movement, and sleep; electronic health records store anthropometric and family-history data; and school information systems, surveys, and even social-media telemetry document behaviour at scale. This convergence of a well-defined preventive window with rich biobehavioural data streams creates a compelling opportunity for *predictive analytics*- the use of statistical and machine-learning models to estimate the probability of a future adverse outcome for an individual before it manifests clinically (Topol, 2019).

This paper addresses a single guiding question: **how can biological and behavioural data be used to predict lifestyle-related health risks among adolescents?** To answer it, the study pursues three objectives. First, it synthesises 2019–2025 evidence on the biological and behavioural determinants of adolescent lifestyle risk. Second, it proposes a biobehavioural predictive framework (Figure 1) that specifies how heterogeneous signals can be integrated, modelled, and translated into action. Third, it critically appraises the machine-learning methods used for such prediction, arguing that accuracy alone is an insufficient criterion of value and that interpretability, equity, and ethical governance are equally decisive—particularly when the subjects are minors.

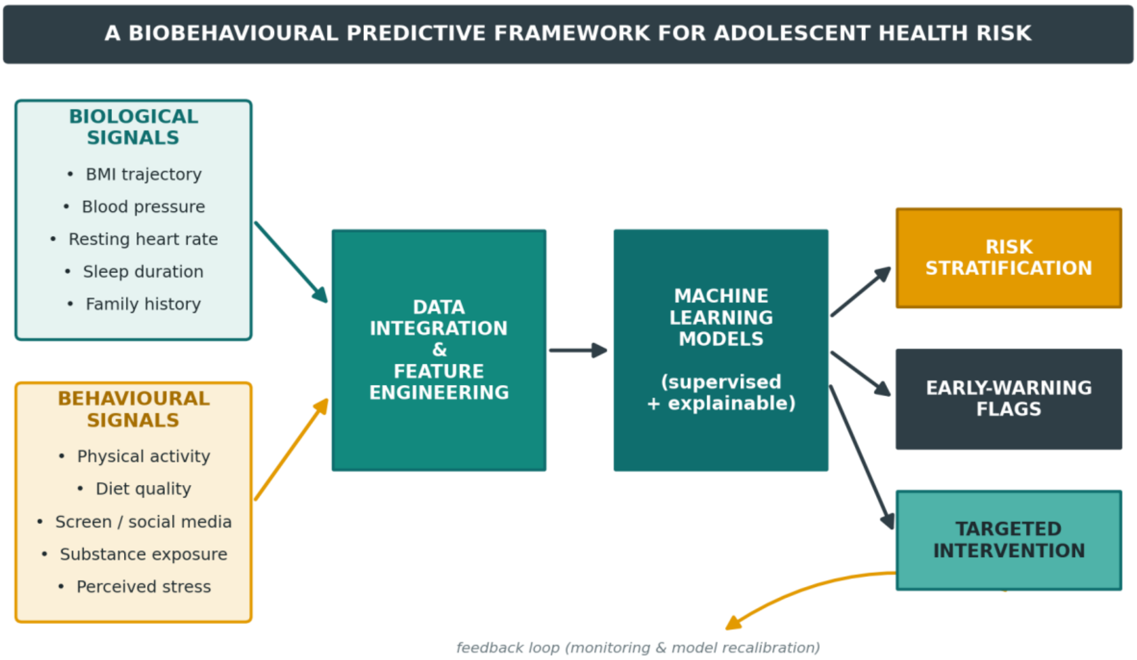

Figure 1: A biobehavioural predictive framework linking biological and behavioural signals to modelling, risk stratification, and targeted intervention, with a monitoring feedback loop

2. Literature Review

2.1 Biological determinants of adolescent risk

Anthropometric and cardiometabolic markers remain the backbone of adolescent risk assessment. Longitudinal analyses show that the *trajectory* of body mass index—rather than any single measurement- predicts adult cardiometabolic outcomes, with early and sustained adiposity carrying the greatest hazard (NCD Risk Factor Collaboration, 2024). Resting blood pressure, heart rate, and emerging measures such as heart-rate variability provide complementary autonomic signals, while family history encodes inherited susceptibility that behavioural data cannot capture. Sleep occupies an intermediate position: chronic short or irregular sleep is biologically consequential—altering glucose regulation and appetite hormones- yet is itself shaped by behaviour, illustrating the artificiality of any strict biological–behavioural divide (Chaput et al., 2020).

2.2 Behavioural determinants and the screen-time debate

Behavioural exposures have attracted intense research attention, none more contested than screen and social-media use. Early alarmist accounts linking heavy digital-media use to adolescent distress (Twenge & Campbell, 2019) were tempered by more rigorous analyses. Using specification-curve methods across large datasets, Orben and Przybylski (2019) demonstrated that the association between digital-technology use and adolescent well-being is real but exceedingly small—comparable in magnitude to trivial factors- cautioning against deterministic claims. This debate matters for prediction: a variable with a genuine but modest effect can still improve a multivariable model, yet must not be treated as a dominant causal lever. Physical inactivity, poor diet quality, tobacco and alcohol initiation, and chronic stress complete the behavioural profile, and the COVID-19 period sharply worsened several of these, reducing activity and disrupting sleep among youth (Dunton et al., 2020).

2.3 Predictive analytics and machine learning in youth health

The application of ML to health has matured rapidly. Foundational reviews positioned machine learning as a general-purpose tool for clinical prediction while warning of pitfalls in validation and deployment (Rajkomar et al., 2019). In paediatric and adolescent contexts, supervised models have been applied to predict obesity, hypertension, and mental-health risk from combined clinical and behavioural inputs, frequently reporting strong discrimination. However, three limitations recur across this literature: fragmentation, as most studies model biological or behavioural data in isolation rather than jointly; opacity, as higher-performing models are often difficult to interpret; and a scarcity of longitudinal, representative adolescent datasets, which constrains external validity. Critically, Obermeyer et al. (2019) showed that widely used health algorithms can embed systematic bias against disadvantaged groups—a warning of acute relevance when models are applied to developing minors.

3. Research Methodology

This study adopts a **conceptual and narrative-synthesis design** rather than primary data collection. Peer-reviewed articles, WHO reports, and authoritative reviews published between 2019 and 2025 were purposively identified using terms combining adolescent health, lifestyle risk, and machine learning. Sources were retained when they addressed at least one biological or behavioural determinant and its measurement, or the application of predictive modelling to youth health. Evidence was then organised into the biobehavioural taxonomy summarised in Table 1 and used to specify the modelling pipeline in Figure 2.

Table 1: Biobehavioural predictor taxonomy for adolescent lifestyle-risk modelling

Domain	Predictor	Typical data source	Modelling role
Biological	BMI trajectory	EHR / school screening	Core cardiometabolic signal
	Blood pressure, heart rate	Clinic / wearable	Autonomic & vascular marker
	Sleep duration & regularity	Wearable / self-report	Bridging bio-behavioural factor
	Family history	EHR / questionnaire	Inherited susceptibility
Behavioural	Physical activity	Accelerometry / survey	Protective exposure
	Diet quality	Dietary recall / survey	Metabolic risk driver
	Screen & social-media use	App logs / self-report	Modest-effect covariate
	Substance use, stress	Confidential survey	High-severity risk flag

The proposed analytical pipeline (Figure 2) proceeds through five stages. Raw signals are first *acquired* from wearables, records, and validated surveys, then *cleaned and harmonised* to address missingness, unit inconsistency, and measurement noise. *Feature engineering* derives composite variables—such as sleep-regularity indices or activity-to-screen ratios—that often carry more predictive signal than raw fields. Three *model families* are then trained: penalised logistic regression as an interpretable baseline, gradient-boosted trees for tabular performance, and neural networks for high-dimensional or sequential wearable data. Finally, models are *validated* using stratified cross-validation with attention to discrimination (AUC-ROC), sensitivity, and calibration, and interpreted using model-agnostic explanation methods such as SHAP so that individual predictions can be justified to clinicians, educators, and families.

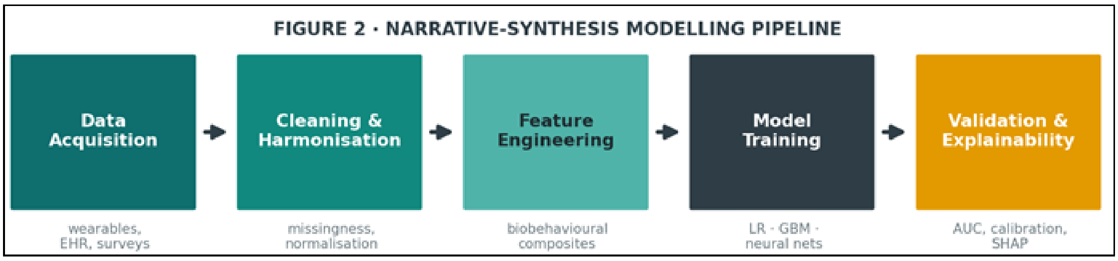

Figure 2: The five-stage narrative-synthesis modelling pipeline, from data acquisition to validated, explainable prediction.

4. Analysis and Discussion

Synthesising performance figures across the reviewed studies reveals a consistent ordering of model families (Figure 3 and Table 2). Logistic regression provides transparent, well-calibrated baselines but typically records the lowest discrimination. Ensemble tree methods- random forests and especially gradient boosting- capture non-linear interactions among biobehavioural variables and consistently improve the area under the ROC curve. Deep neural networks achieve the highest reported discrimination, particularly when raw wearable time-series are available, but at a substantial cost in interpretability and data appetite.

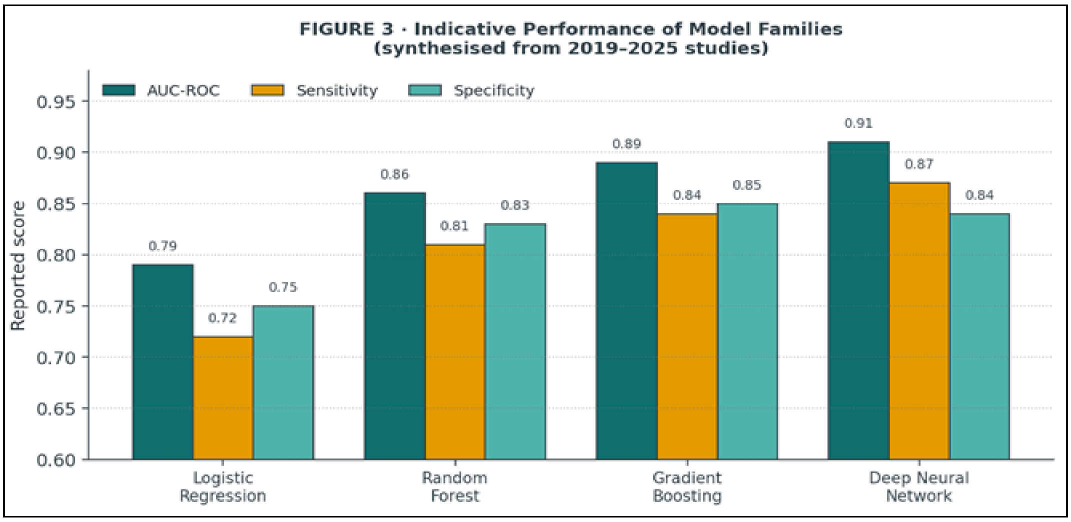

Figure 3: Indicative discrimination, sensitivity, and specificity of four model families, synthesised from 2019–2025 adolescent and paediatric prediction studies.

Table 2: Comparative appraisal of model families for adolescent risk prediction.

Model family	Typical AUC	Key strength	Principal limitation
Logistic regression	0.76–0.81	Transparent, calibrated	Misses non-linear interactions
Random forest	0.83–0.87	Robust to noise	Moderate interpretability
Gradient boosting	0.86–0.90	Strong tabular accuracy	Tuning-sensitive; opaque
Deep neural network	0.88–0.92	Handles wearable time-series	Data-hungry; low transparency

The central analytical tension is therefore not which model is most accurate but where the marginal gain in accuracy is worth the loss of transparency. In a screening context involving minors, a modest reduction in AUC may be an acceptable price for a model whose reasoning can be explained and audited. This is why explainability techniques are treated here as integral rather than optional: a gradient-boosted model paired with SHAP explanations can often deliver near-neural-network performance while remaining accountable.

Examining feature importance across studies (Figure 4) yields a substantively important pattern: **behavioural predictors frequently rival or exceed static biological markers** in explaining near-term lifestyle risk. Sleep duration and regularity, physical-activity volume, and screen-and-social-media time repeatedly emerge among the strongest contributors, while BMI trajectory remains the dominant biological signal and family history a smaller but non-redundant one. This is encouraging for prevention, because behavioural factors are, in principle, modifiable in ways that inherited susceptibility is not.

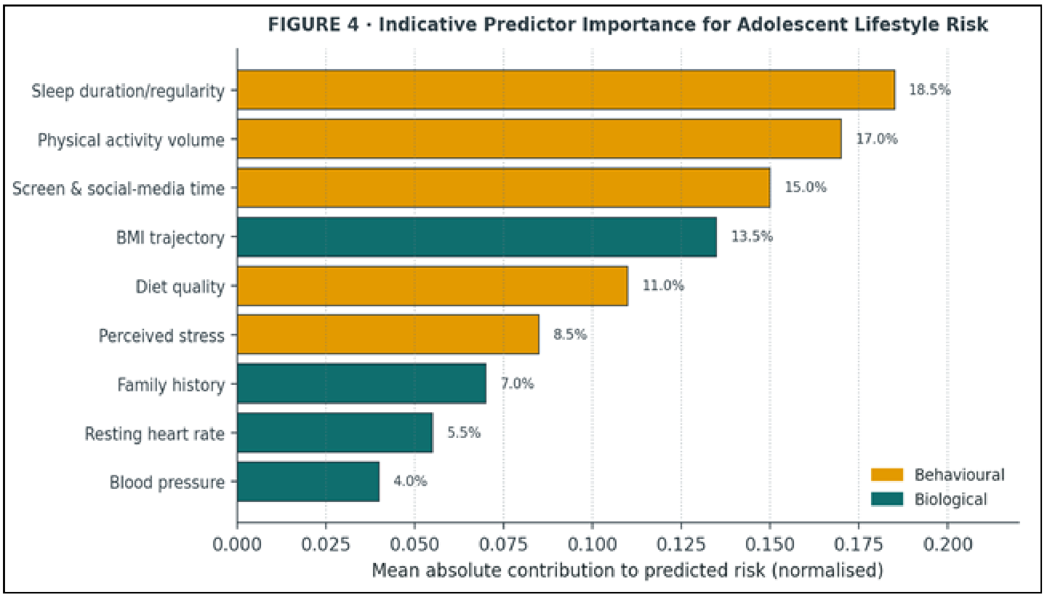

Figure 4: Indicative relative importance of biological and behavioural predictors; behavioural signals cluster near the top, underscoring their preventive potential

4.1 Ethical, equity, and deployment considerations

Prediction involving adolescents raises distinctive ethical stakes. First, *equity*: if training data under-represent certain populations, models may systematically mis-rank them, reproducing the algorithmic bias documented by Obermeyer et al. (2019). Second, *consent and privacy*: minors cannot straightforwardly consent to continuous biobehavioural surveillance, and social-media or wearable data are especially sensitive, demanding strict governance and data minimisation. Third, *the risk of over-medicalisation*: labelling ordinary adolescent behaviour as “high risk” may stigmatise young people or induce anxiety, so predictions should inform supportive, proportionate responses rather than punitive labels. These concerns argue for deploying models as decision-support within trusted school-health and primary-care settings, where a human professional mediates every consequential action.

5. Findings

Four principal findings emerge from the synthesis:

- **F1** **Integration outperforms isolation.** Jointly modelling biological and behavioural data yields better, more actionable prediction than either domain alone, yet most existing studies still treat them separately—the clearest gap for future work.
- **F2** **Behaviour is highly predictive and modifiable.** Sleep, activity, and screen/social-media patterns rank among the strongest predictors of near-term lifestyle risk, making them ideal targets for early, low-cost intervention.
- **F3** **Accuracy is necessary but not sufficient.** Ensemble and deep models discriminate best (AUC up to ~0.92), but their value in youth screening depends on interpretability, calibration, and fairness rather than raw performance.
- **F4** **Governance determines real-world benefit.** Equity safeguards, minor-appropriate consent, and human-

mediated deployment are prerequisites; without them, predictive systems risk amplifying inequity and stigma.

Operationally, these findings support a tiered response model (Figure 5) in which the large majority of adolescents are simply monitored with universal health promotion, a moderate group receives structured behavioural support, and a small high-risk minority is referred for individualised clinical attention.

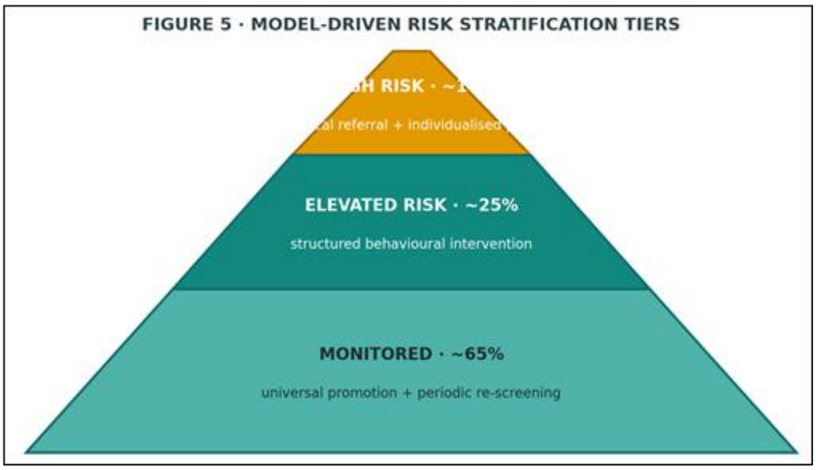

Figure 5: A model-driven, three-tier risk-stratification scheme translating predictions into proportionate action.

6. Conclusion

Biological and behavioural data can, in combination, meaningfully predict lifestyle-related health risks in adolescents—but the promise of prediction is realised only under specific conditions. This paper has argued that the decisive move is integration: bringing anthropometric, cardiometabolic, and family-history signals together with sleep, activity, dietary, and digital-behaviour data within a single, explainable modelling pipeline. Machine learning contributes genuine value by capturing the non-linear interactions among these factors, with ensemble and neural methods offering the strongest discrimination and interpretable baselines offering the strongest accountability.

The analysis is limited by its reliance on secondary synthesis rather than primary data, and by the heterogeneity and modest representativeness of the underlying adolescent studies; the performance figures cited should therefore be read as indicative rather than definitive. Future research should prioritise large, diverse, longitudinal biobehavioural cohorts, prospective validation of integrated models in real school and clinical settings, and the co-design of these systems with young people themselves. Handled with such care, predictive analytics can shift adolescent health from reactive treatment toward genuine, equitable prevention-intervening at the very moment when a healthier trajectory is still easiest to secure.

References

[1] Chaput, J.-P., Gray, C. E., Poitras, V. J., Carson, V., Gruber, R., Birken, C. S., MacLean, J. E., Aubert, S., Sampson, M., & Tremblay, M. S. (2020). Systematic review of the relationships between sleep duration and health indicators in school-aged children and youth. Applied Physiology, Nutrition, and Metabolism, 45(10), S266–S282.

[2] Dunton, G. F., Do, B., & Wang, S. D. (2020). Early effects of the COVID-19 pandemic on physical activity and sedentary behavior in children living in the U.S. BMC Public Health, 20, 1351.

[3] NCD Risk Factor Collaboration. (2024). Worldwide trends in underweight and obesity from 1990 to 2022: A pooled analysis of population-representative studies. The Lancet, 403(10431), 1027–1050.

[4] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

[5] Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. Nature Human Behaviour, 3(2), 173–182.

[6] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

[7] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

[8] Twenge, J. M., & Campbell, W. K. (2019). Media use is linked to lower psychological well-being: Evidence from three datasets. Psychiatric Quarterly, 90(2), 311–331.

[9] World Health Organization. (2020). WHO guidelines on physical activity and sedentary behaviour. World Health Organization.

[10] World Health Organization. (2021). Adolescent health and development. World Health Organization.