Artificial Intelligence and Precision Psychology for Predictive Mental Health: Integrating Climate Stress, Environmental Exposure, and Personalized Psychological Interventions

[1]Sumaira Aslam, [2]Ausaf Chaudhary,[3]Dr. Muhammad Usman Zia ,[4]Dr. Samreen Fatima, [5]Qasem Kharma,[6]Raza Iqbal

[1]School of Psychology, Central China Normal University, Luoyu Road 152, Wuhan, Hubei, China.

sumairaaslam2644@gmail.com

[2] Ph.D (Human Kinsiology and Science) College of Physical Education

Yangzhou University, China Email: ausaf.choudhary@gmail.com

ORCID Number: 1596-1629-1596-1629

[3]Lecturer, University of Education Lahore

usman.zia@ue.edu.pk ORCID 1596-1629-1596-1629X

[4]Assistant professor Department of Psychiatry and Behavioral Sciences

Abu Umara Medical & Dental College Green International University Lahore

Sam.hawking@gmail.com

[5]q.kharma@ammanu.edu.jo

Department of Software Engineering, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan

[6] Department of Computer Science, National College of Business Administration & Economics (AL-HAMRA UNIVERSITY) Multan Campus Multan, Pakistan. Ali.raza@bzu.edu.pk https://orcid.org/1596-1629-1596-1629

HOW TO CITE:
Sumaira Aslam, Ausaf Chaudhary,Muhammad Usman Zia , Samreen Fatima , Qasem Kharma,Raza Iqbal, (2026). Artificial Intelligence and Precision Psychology for Predictive Mental Health: Integrating Climate Stress, Environmental Exposure, and Personalized Psychological Interventions International Journal of Special Education, 41(20s), 1596 - 1629

The evidence base for effects of climate change and environmental exposure on mental health is virtually limited to average effects across groups. To quantify within-person associations between short-term heat and particulate exposure and daily psychological distress and their lag structure, as well as to identify moderators and one candidate mediator; To test, in silico, the efficacy of risk-triggered delivery of a brief psychological module using a fixed intervention budget compared to universal and random delivery; To determine the incremental value of environmental exposure features over clinical and digital-phenotyping features for 14-day symptom deterioration, including the lag structure of these features; To evaluate, in silico, the efficacy of risk-triggered delivery of a brief psychological module using a fixed intervention budget, compared to universal and random delivery. We created a synthetic population of 4,000 adults distributed in 5 climate regimes, experienced over an 180-day period (720,000 person-days). Published meta-analytic estimates of the effects of heat, particulate matter and greenness were used as anchors for structural parameters of the data-generating process. The within-person associations were estimated with person-mean centred regression and confirmed with random-intercept mixed models, with cluster-robust standard errors. Four prediction models were created (penalised logistic regression using clinical variables; gradient boosting using digital phenotyping; gradient boosting with environmental exposures added; multilayer perceptrons using flattened daily sequences). These models were evaluated using participant-grouped five-fold cross-validation and leave-one-site-out internal–external validations. Four delivery policies were tested within a framework of an assumed intervention effect based on the meta-analytic estimate of the JIT adaptive interventions intervention effect. Cumulative heat excess (lag 0–3) was associated with higher daily distress ($\beta = 0.51$ points per °C, 95% CI 0.47–0.55, $p < .001$; $d = 0.079$) as was cumulative PM2.5 (lag 0–6; $\beta = 0.72$ per 10 μg m−3, 95% CI 0.68–0.76, $p < .001$; $d = 0.112$). The impacts were greatest at lag 1596-1629 days for heat and lag 1596-1629 days for particulates. Trait climate anxiety amplified, and residential greenness and cooling access attenuated, the heat–distress slope. Sleep duration mediated 24.2% of the total heat effect (indirect $\beta = 0.117$, 95% CI 0.110–0.125). For 14-day deterioration (4,420 events among 44,000 person-windows; 10.0%), AUROC rose from 0.744 (95% CI 0.737–0.752) for clinical variables alone to 0.855 (0.850–0.860) with digital phenotyping, but adding the full environmental block yielded only 0.863 (0.858–0.868), an increment of 0.008 (95% CI 0.006–0.009). The results of the block-wise permutation analysis indicated that heat and air-quality features accounted for negligible amounts of discrimination (ΔAUROC 0.001 and 0.003 respectively) and that residential greenness, a stable person-level attribute, accounted for 0.015. The calibration slopes were approximately unity and performance was stable across sites (0.853–0.868) and subgroups (0.854–0.872). The delivery at 30% budget trial resulted in an absolute risk reduction of 1.29 percentage points (95% CI 1.18–1.39) compared with the number of prompts per event averted (23.3) at random allocation at the same budget (0.53, 95% CI 0.47–0.60) and universal delivery (1.77, 1.64–1.88 prompts per event averted). The individual risk prediction is a complementary approach to environmental exposure. Within-person associations of short-term exposure to heat and short-term exposure to particulate matter with distress are reproducible, lagged and moderated, and are not particularly helpful for discriminating who will deteriorate; these are important for understanding mechanism and timing delivery

1. Introduction

There is development of the climate–mental health evidence base, though it is still at a population level.

EPA has progressed the relationship between environmental conditions and psychological functioning from a speculation to a quantified field of inquiry in the last six years. A meta-analysis of 32 umbrella studies, which included 284 primary studies and 237 exposure–outcome associations, found that the evidence supporting the higher ambient temperature–suicide association is suggestive to convincing, while several associations between cognition and air pollution reach a similar level of credibility (Radua et al., 2024). A systematic review and meta-analysis of ambient temperature and mental health indicated that during periods of heatwaves, there was an association between hospital attendance or admission for mental illness and the heatwave period, with a 9.7% increase in hospital attendance or admission for mental illness compared to non-heatwave periods (95% CI 7.6–11.9, $p < .001$), and that a 1 °C increase in mean monthly temperature was associated with a 1.5% increase in suicide incidence (95% CI 0.8–2.2, $p < .001$) (Thompson et al., 2023). Small but consistent correlations between long-term exposure to PM2.5 and incident depression (RR 1.074 per unit increase in PM2.5; 95% CI 1.021–1.129) and immediate exposure to PM2.5 and same-period depressive episode (Borroni et al., 2022) and between residential greenness and reduced odds of common psychiatric disorders (pooled RR 0.91; 95% CI 0.89–0.92) (Zhang, Wu, et al., 2024). These exposure effects are part of a larger exposure profile of climate-related pathways, such as displacement, livelihood loss and disrupted social infrastructure (Charlson et al., 2021; Cianconi et al., 2020; Lawrance et al., 2022; Li et al., 2023).

The other, more partial literature provides anticipatory and existential responses to environmental change. One study, which polled 10,000 16- to 25-year-olds across ten countries, revealed that 59% were very or extremely concerned about climate change, and over 45% said their concern hindered their daily functions (Hickman et al., 2021). Climate anxiety has been identified as a measurable construct consisting of two dimensions: cognitive-emotional and functional-impairment, which is different from generalised worry because of its interference with concentration, sleep and everyday functioning (Clayton & Karazsia, 2020; Mouguiama-Daouda et al., 2022; Wullenkord et al., 2021). Associations with depressive and anxious symptoms are in the same direction across samples and instruments but are inconsistent in magnitude (Ogunbode et al., 2022; Schwartz et al., 2023; Thomson & Roach, 2023; Whitmarsh et al., 2022) and related, but not identical constructs (ecological grief, solastalgia, climate worry) have their own measurement traditions (Comtesse *et al.,* 2021; Ojala *et al.,* 2021; van Valkengoed *et al.,* 2021).

These literatures share the same common denominator as they have for clinical use: they are limited. Nearly all estimates are averages of contrasts between groups, periods or places. A 9.7% excess in admissions during heatwaves is a characteristic of a population, not a clinical characteristic of any one of the individuals within that population, or it is a characteristic of a public-health service, and it tells nothing about the direction of a limited number of outreach contacts in that service's domain. The point of precision approaches is to fill precisely the gap between P.A.R and individual actionable risk.

1.2 Precision psychology and limitation of individual prediction

Precision psychiatry aims to replace the average diagnosis with an individualised prognosis, based on multivariate data (Salazar de Pablo et al., 2021). There are two well documented failure modes that have tempered the early optimism of the field. The first is positive retrospective validation which is when models often achieve very high apparent accuracy in their training data set, but perform poorly on the rest. This has been shown to be the case with high within-trial accuracy of machine-learning models predicting antipsychotic treatment outcomes that do not generalise out-of-sample, and without the gains of pooling across trials, as Chekroud et al. (2024) demonstrated. The second is the replacement of the concept of discrimination with the notion of clinical usefulness – a model can perform a separation of cases and non-cases and may show little or no gain at any threshold that a service would use – Vickers-style decision analysis has become the standard practice (Van Calster et al., 2019; see also Hicks et al., 2022).

These worries have created a self-stated methodological correction. The minimum reporting items for prediction models developed with regression and machine learning (Collins, Moons, et al., 2024) are specified; a matched risk-of-bias instrument for prediction models is also available (PROBAST+AI; Moons et al., 2025); and stepwise development and external-validation guidance are available (Collins, Dhiman, et al., 2024; Efthimiou et al., 2024; Radua & Koutsouleris, 2024). Ghassemi et al. (2021) and Wilkinson et al. (2020) have suggested that the field should be cautious of making predictions without independent validation and should report on calibration/benefits as well as discrimination. These are the standards that we follow here.

1.3 Digital phenotyping provides the temporal resolution that environmental data needs

The traditional measure of symptoms doesn't work on the daily timescale of environmental exposure. The missing temporal resolution is provided by digital phenotyping the inference of behavioural and physiological states from signals passively collected from the smartphone and wearable devices. Sleep duration and timing, physical activity, mobility entropy, communication frequency and heart-rate variability are reported to contain replicable (although modest) information about the severity of depressive and anxious symptoms (Beames et al., 2024; Choi et al., 2024; Ikäheimonen et al., 2024; Torus et al., 2021); also, that circadian features show a seasonally patterned variance that is plausibly related to weather (Zhang, Folarin, et al., 2024). The two main methodological issues are passively sensed features are often highly correlated with each other and/or with previous symptom scores (Held & Cribb, 2021), and that missingness is not often ignorable and can have a negative impact on the quality of features (Currey & Torous, 2023).

The delivery channel is also provided by digital phenotyping. Just-in-time adaptive interventions (JITAIs) make decisions about support provision in real-time based on the state and context (Nahum-Shani et al., 2018); micro-randomized trial designs enable the evaluation of the decision rules of JITAIs in an experimental fashion (Klasnja et al., 2019; Qian et al., 2022). Empirical results to date are modestly positive: a meta-analysis of 23 studies and 2,563 participants with JITAIs and ecological momentary interventions for depression, anxiety and psychological wellbeing showed a small pooled between-group effect (Hedges' g = 0.15, 95% CI 0.05–0.26, p = .003). Efficacy is not the only constraint on such systems, it is engagement (Nahum-Shani et al., 2022).

1.4 The unexamined integration

There is a growing number of interventions that are specifically addressing the mental health impacts of climate change, but they are largely un-evaluated: We conducted a scoping review of interventions and intervention packages and identified 37, of which the majority had not been formally evaluated for outcomes (Xue et al., 2024). Professional bodies have urged for a greater inclusion of climate into psychiatric practice and training (Brandt et al., 2024), and health-and-climate assessments are already conducted at an annual scale at a granularity that would allow for individual linkage (Romanello et al., 2024). But so far as we know, no study has asked the important quantitative question upon which all integration is based: When clinical history and passively sensed behaviour are already available, does the incorporation of environmental exposure yield better individual predictions of risk and, if so, what is environmental exposure really for?

The question can currently only be answered based on theoretical or conceptual reasoning, since a cohort is not available that integrates dense digital phenotyping with repeated validated symptom assessment with individually linked environmental exposure and appropriate climatic heterogeneity. However, it can be answered conditionally: if the parameters of the DGP were fixed to the published effect sizes, what would such a study yield? This conditional answer has the advantage of pointing out beforehand what features of the design are crucial to the outcome, and that it is falsifiable in the sense that a real cohort with the hypothesized effect sizes should replicate the qualitative pattern. This is the rationale behind the current research.

1.5 Objectives and pre-specified expectations

We have tackled three aims. First, associations between short-term heat and particulate exposure, daily psychological distress were estimated, the lag structure was characterized, moderation by trait climate anxiety, residential greenness, cooling access and socioeconomic position was tested, and sleep was tested as a candidate mediator. Second, to measure the incremental discriminative and calibrative gains of environmental exposure features over clinical and digital-phenotyping features for predicting clinically significant symptom deterioration at a 14-day horizon, including internal–external validation at climatically distinct sites and a subgroup fairness audit. Third, to determine in silico whether a brief psychological module is delivered optimally when it is delivered based upon adaptive cues, risk alone, or universally, and whether the universal or random delivery outperforms the risk delivery alone.

Based on the small standardized magnitudes reported in the source literature, we had expected that the environmental exposure would exhibit strong within-person associations, albeit small, and that the exposures would contribute little to individual discrimination. We anticipated that adaptive targeting will boost efficiency, roughly in line with model discrimination. There was no predetermined direction of expectation for the trigger on the environmental context.

2. Materials and Methods

2.1 Study design, reporting standard and the status of the data

This is an in-silico cohort study. We first created a synthetic longitudinal data set from an explicitly specified data-generating process (DGP) and then analysed the data based on the same analytic approach that would be employed with real longitudinal data of the same structure. The development and validation of the prediction models was done following the TRIPOD+AI checklist (Collins, Moons, et al., 2024) and the risk-of-bias considerations listed in PROBAST+AI (Moons et al., 2025) were considered in the validation strategy, especially when deciding to perform leave-one-site-out internal–external validation instead of random splits.

We are totally confident, without a doubt, the cohort is simulated. No human subjects were involved and no empirical data, at the individual level, were employed. This is a consequence of interpretation, not of the global nature of the statistical estimates reported below: The figures below are real calculations and inferences about the behaviour of analytic methods under stated assumptions (such as how much discrimination a feature block can contribute given its effect size) are valid. There are no statements of effect size made here, they are brought in from the literature to which the DGP was anchored. The classification of each substantive claim is shown in Table 14.

The construction of synthetic cohorts and anchoring of parameters

The cohort consisted of 4,000 adults who were observed in five sites that were selected to represent different climate and pollution environments (Table 1), for 180 days over a warm season, for a total of 720,000 person-days. The allocation of the sites was not random but was determined by design, meaning that site would be a true grouping factor to achieve internal–external validation.

The following procedure was used to create person-level attributes. Age was normally distributed with a range of 18–70, a mean of 37.5, and a standard deviation of 12.6, and sex was assigned with a probability of 0.53 being a female. A socioeconomic position index was created on a continuous scale from [0,1]. Residential greenness (NDVI) was created by site-specific mean and standard deviation, and a positive socioeconomic gradient, which reproduced the environmental-justice patterning consistently seen in real settings. The chance of access to indoor cooling was Bernoulli and increased with socioeconomic position. This is scored on a 13–65 scale that corresponds to the Climate Change Anxiety Scale (Clayton & Karazsia, 2020), and was generated by this trait, in combination with age (negated), and female gender (eg signified by “+f”). The vulnerability trait, climate anxiety and socioeconomic position were used to generate baseline PHQ-9 (Kroenke et al., 2001) and GAD-7 (Spitzer et al., 2006) scores.

The structural coefficients of the DGP were not selected to account for a desired effect but to approximate the magnitudes reported in the literature. That is: the coefficient linking cumulative heat excess to daily distress was set so that heatwave days and non-heatwave days had the reported difference in psychiatric morbidity reported by Thompson et al. (2023); the particulate coefficient was set so that the excess psychiatric morbidity risk ratio per day of exposure to PM2.5 was comparable to the short-term and long-term psychiatric morbidity risk ratios reported by Borroni et al. (2022); and the greenness coefficient was set so that the protective odds ratio per day of exposure to greenness compared to urban heat exposure was comparable to the pooled odds ratio reported by Zhang, Wu, et al. (2024). Three additional coefficients, amplification of the heat slope by trait climate anxiety, buffering by greenness, and buffering by cooling access, were assumed throughout as the plausible, but not yet estimated by meta-analysis, parameters of the structure, and are therefore described as such in Table 14. The complete values for all parameters are provided in the supplemental parameter file along with the analysis code.

2.3 Environmental exposure assembly

A daily heat-index series consisting of a warm-season arch, and a first-order autoregressive stochastic component was designed for each site, and a daily PM2.5 series was derived in the log scale with an autoregressive structure. The individual exposure was calculated by adding independent measurement error to the series for the ambient exposure, which is the difference between fixed-monitor and individual exposure that is typical of real linkage studies. Cumulative heat excess was calculated as the average of the four previous days heat index (lags 0–3) minus the 75th centile, set to zero if negative; cumulative particulate exposure was the average for lags 0–6. The heatwave was defined as 3 or more consecutive days of days above the site-specific 95th centile. A "hot night" was considered as a day which exceeded the pooled 90th centile of the heat index.

2.4 Psychological measures and the digital phenotype

A latent daily distress score (range 0–100) was calculated as the sum of a person-specific intercept (based on baseline PHQ-9, vulnerability, greenness, socioeconomic position and climate anxiety), the environmental drive terms, a sleep-mediated term, an autoregressive residual ($\varphi = 0.58$, SD = 6.2) and a slow individual drift that allows true deterioration and improvement across the window.

Seven signals representing daily changes in the individual's behavior as a noisy function of latent distress and same-day exposure were derived: sleep duration, sleep-midpoint timing, step count, mobility entropy, screen time, number of social contacts, and a proxy of heart-rate-variability. The device non-wear rate was assigned as a missing-at-random distribution with the probabilities of missingness increasing in both low SES and distress, resulting in an overall missing rate of 10.9% similar to the missingness burden reported in real-world digital-phenotyping cohorts (Currey & Torous, 2023). The modelling pipeline was median imputed for models that needed complete data, but natively for tree-based models and the missing-data-fraction feature was retained as a predictor because informative missingness is informative too.

To obtain 13 assessments for each participant, PHQ-9 was administered every 14 days, such that each score had been mapped linearly within the interval (0, 10) with a baseline distribution and was the sum of the mean latent distress in the preceding fortnight plus independent measurement error.

2.5 Outcome definition

The prediction target was clinically significant deterioration over a period of 14 days, which was defined at each assessment wave as either a rise of at least 5 points on the PHQ-9 at the next wave, or moving from below 10 to at or above 10 (the conventional threshold for moderate depressive symptoms). The composite is related to threshold-based and change-based deterioration and was fixed prior to fitting a model. These were filled into prediction windows from 1 to 11, a total of 44,000 person-windows produced 4,420 events (10.0%).

2.6 Within-person association models

To isolate within-person effects from any between-person confounding, all time varying variables were centred on their person means, and the algebraic equivalent of a participant fixed-effects model was specified, which means it is devoid of any between-person confounding by construction. Participant (Huber–White) clustering was used to calculate standard errors. To ensure that the calculation was feasible, models were estimated on a random subsample of 2,500 participants (450,000 person-days). Within-person daily distress was modelled as a function of within-person cumulative heat excess and cumulative particulate exposure in the primary model. Therefore, a random-intercept mixed model was also fit on a second sample of 600 participants to validate magnitudes of the coefficients and to estimate the percentage of variance between persons.

Lag structure: The model was refitted separately at each individual lag (0 to 7 days) and both exposures were included simultaneously at that lag. Within-person heat excess and, respectively, standardised trait climate anxiety, centred greenness, cooling access, and particulate exposure and socioeconomic position were tested for moderation with product terms. Product terms in fixed-effects specifications will capture cross-level interactions because the moderator is fixed. The proportion mediated was reported using the product-of-coefficients approach, along with a delta-method standard error for the indirect effect; however, in the absence of the sequential-ignorability assumptions, it is not a causal quantity.

Standardised effect sizes are presented as Cohen's d in relation to the within-person standard deviations of distress. The E-values were calculated for the two key exposures to determine the minimum strength of unmeasured confounding (on the risk-ratio scale) that would be needed to account for the association (VanderWeele & Ding, 2017). The eight lag models and six moderation terms were controlled for multiplicity using the Benjamini–Hochberg procedure with a false-discovery rate of 5%.

2.7 Prediction model development

Features were grouped into 3 pre-specified blocks. Age, sex, socioeconomic position, trait climate anxiety, prior episode, baseline and current PHQ-9, baseline GAD-7, PHQ-9 change since the previous wave and PHQ-9 slope since baseline were included in Block A (clinical and self-report; 10 features). Passive signals means, standard deviations, linear slopes and 7-day-versus-prior-7-day contrasts were included in Block B (digital phenotyping; 20 features). Block C (environmental exposure; 14 features) had 14 days mean and maximum heat index, 14 days mean and maximum cumulative heat excess, heatwave days, hot nights, mean and maximum PM2.5, days above 35 μg m^{-3}, cumulative lag 0–6 particulates, recent-versus-earlier contrasts for both exposure, residential NDVI, and cooling access. For full definitions see Table 2.

A priori four models were specified. M1 was logistic regression on Block A only (the information that is routinely available in clinical services). M2 was the same learner, in this case a histogram gradient boosting, applied to Blocks A and B; while M3 was the same learner applied to Blocks A, B and C, which was included to determine whether the learner could extract more structure from the data, than the tree ensemble did. M4 was a multilayer perceptron (two levels of 96 and 48 units respectively, L2 penalty, early stopping) on the same feature set as M3, and included to see if a more flexible function approximator could extract the structure that the tree ensemble was missing. The hyperparameters were given conventional values, and were neither optimized to the validation set nor tuned once the validation was done, thus eliminating the optimistic bias of tuning-on-test.

2.8 Validation, performance metrics and fairness audit

Five-fold cross validation by participant was used for internal validation as all the 11 windows contributed by a participant would fall in a single fold, otherwise, row-wise splitting would have leaked person-level information and overestimated the performance. The area under the receiver operating characteristic curve (AUROC) and average precision (which is the more informative summary at 10% prevalence) were summarised for discrimination. Calibration was evaluated using calibration slope and intercept according to logistic recalibration model and using decile plots. The Brier score was used to summarize the overall accuracy. Operating characteristics were presented at a trade-off point of 80% sensitivity, selected because it is more typical for a screening than a diagnostic usage.

A 400 bootstrap resamples were used to get confidence intervals of AUROC and average precision. Because the models are different, differences in AUROC were tested via a clustered bootstrap resampling of the participant level (not the resampling of the row level), which respects the dependence structure, and the bootstrap proportion was reported as a two-sided p-value, floored at the resolution of the resampling. Internal–external validation was conducted by rotating the train, test set, as per Chekroud et al. (2024) who asserted that to test for generalisation, the train-test set must be rotated between distinctive sets of contexts. Subgroup analysis of AUROC, calibration slope, positive predictive value and Brier score was conducted by sex, age tertile, socioeconomic tertile, greenness median split and high trait climate anxiety (Chen *et al*., 2021).

Permutation was used to assess the variable importance, feature by feature and block by block. Block-wise permutation was added as a precautionary measure; it is clear that single-feature permutation underlies the importance of collinear sets of feature by systematically ignoring the information that can be gained from the other features. These importances are interpreted per the cautions of Ghassemi et al (2021) that post hoc attributions report model rather than mechanism. The clinical utility was evaluated using decision-curve analysis, which compared the net benefit of each model with strategies of providing support to all and to none of the children between thresholds of 2% and 41% of the child population.

2.9. In-silico evaluation of intervention delivery policies

We compared four delivery policies with a no intervention policy. P1 presented a short module to each person-window (universal fixed schedule). P2 delivered to a random 30% of person-windows. The P3 service area covered the highest risk 30% of the population according to predicted probability of risk from M3. P4 was presented to the top 30% of most vulnerable babies, based on a composite of the predicted risk, and on a binary environmental stress flag (heat excess or particulate exposure above 70th percentile), and tested the hypothesis that contextual information can be used to augment the targeting of the predicted risk alone.

The intervention effect was not estimated. Delivering the treatment shifted the between-group log odds of deterioration by a constant amount that resulted in a between-group standardized effect of $g = 0.15$, which is the pooled meta-analytically estimated effect of JITAIs and ecological momentary interventions in mental health. This is a powerful simplifying assumption, that imposes a uniform effect, no decay of engagement, no dose–response and no differential responsiveness. The absolute risk reduction is sensitive to the assumed effect and is therefore reported only for the purpose of performing a sensitivity analysis for the range of allowed values for g, which is 0.05 to 0.35. We highlight throughout that the quantity that is interpretable is the relative efficiency of the policies, which is much less sensitive to the assumed effect than the absolute risk reduction. Outcomes were absolute risk reduction (bootstrap 95% CI), relative risk reduction (bootstrap 95% CI), number needed to treat (400 bootstrap replications, 95th CI), and number of prompts per event averted (400 bootstrap replications, 95th CI).

2.10 Software, Reproducibility, Ethics

All analyses were performed with Python 3 with NumPy, pandas, SciPy, statsmodels and scikit-learn (Pedregosa et al., 2011), gradient boosting was based on the variant of the algorithm that operates on the histogram of features (Chen and Guestrin, 2016), and permutation-based attribution on the additive-attribution framework (Lundberg and Lee, 2017). Conceptually, the specification used in the mixed model is the same as the one implemented in lme4 (Bates et al., 2015) and variance partitioning is based on the model as outlined in Nakagawa and Schielzeth (2013). The single-lag structure is a simplified version of Gasparrini (2011)'s distributed-lag structure. Multiple imputation by chained equations (van Buuren & Groothuis-Oudshoorn, 2011), as an alternative missingness approach, is discussed as a limitation. All the seeds used were fixed; the generator and analysis scripts generate all the numbers reported. Since there was no human involvement, there was no need for ethical approval or informed consent, as detailed in the Declarations.

3. Results

The composition of the cohort and its exposure profile

The cohort and its exposure environment are described in Table 1. The exposures were significantly different between sites. Mean heat index ranged from 24.6 °C (SD 2.7) at the temperate maritime site to 40.6 °C (SD 4.3) at the semi-arid site, with a maximum of 53.3 °C. Mean PM2.5 ranged from 11.1 μg m−3 (SD 4.2) to 48.0 μg m−3 (SD 19.0); the proportion of person-days exceeding the WHO 24-hour interim target of 35 μg m−3 ranged from 0.0% to 72.6%. Sites with the highest levels of these burdens were the least green, in terms of the surrounding greenness (NDVI 0.18 at the semi-arid site, 0.56 at the maritime site), and the pattern of these exposures is also characteristic of real settings and has a direct impact on the equity analysis of Section 3.9. On purpose, the following characteristics were balanced across sites: mean age 38.0 years (SD 11.7), 52.9% female, mean baseline PHQ-9 6.8 (SD 5.0), mean baseline GAD-7 6.3 (SD 4.4) and mean climate-anxiety score 27.3 (SD 7.1). The distributions of greenness and trait climate anxiety are shown in Figure 3 and their exposure trajectories.

Table 1. Cohort characteristics, environment exposure profile by site.

Characteristic	S1	S2	S3	S4	S5	All
Site description	Humid subtropical	Semi-arid continental	Temperate maritime	Temperate continental	Tropical coastal	—
Participants, n	1,050	820	760	790	580	4,000
Age, years, M (SD)	38.0 (11.8)	38.9 (11.8)	37.6 (11.7)	37.5 (11.8)	38.1 (11.4)	38.0 (11.7)
Female, %	52.4	54.0	52.8	52.8	52.8	52.9
Socioeconomic index, M (SD)	0.50 (0.19)	0.50 (0.19)	0.50 (0.19)	0.50 (0.19)	0.50 (0.19)	0.50 (0.19)
Residential NDVI, M (SD)	0.29 (0.09)	0.18 (0.07)	0.56 (0.11)	0.42 (0.10)	0.38 (0.10)	0.36 (0.16)
Cooling access, %	52.1	55.4	54.3	54.9	52.9	53.9
Baseline PHQ-9, M (SD)	6.9 (5.0)	6.8 (5.0)	7.2 (5.1)	6.3 (4.7)	6.8 (5.0)	6.8 (5.0)
Baseline GAD-7, M (SD)	6.3 (4.4)	6.4 (4.4)	6.6 (4.6)	5.9 (4.2)	6.3 (4.4)	6.3 (4.4)
Climate-anxiety score, M (SD)	27.6 (7.1)	26.7 (7.2)	27.6 (7.1)	27.4 (7.0)	27.2 (7.0)	27.3 (7.1)
Heat index, °C, M (SD)	36.6 (3.3)	40.6 (4.3)	24.6 (2.7)	30.0 (3.4)	35.3 (2.4)	33.6 (6.5)
Maximum heat index, °C	47.8	53.3	34.9	41.3	43.5	53.3

Characteristic	S1	S2	S3	S4	S5	All
Heatwave days per participant	6.0	3.0	5.0	4.0	0.0	3.9
PM2.5, µg m^{-3}, M (SD)	48.0 (19.0)	33.9 (14.3)	11.1 (4.2)	20.1 (8.7)	25.9 (10.8)	29.4 (18.7)
Person-days > 35 µg m^{-3}, %	72.6	41.3	0.0	6.7	17.6	31.4

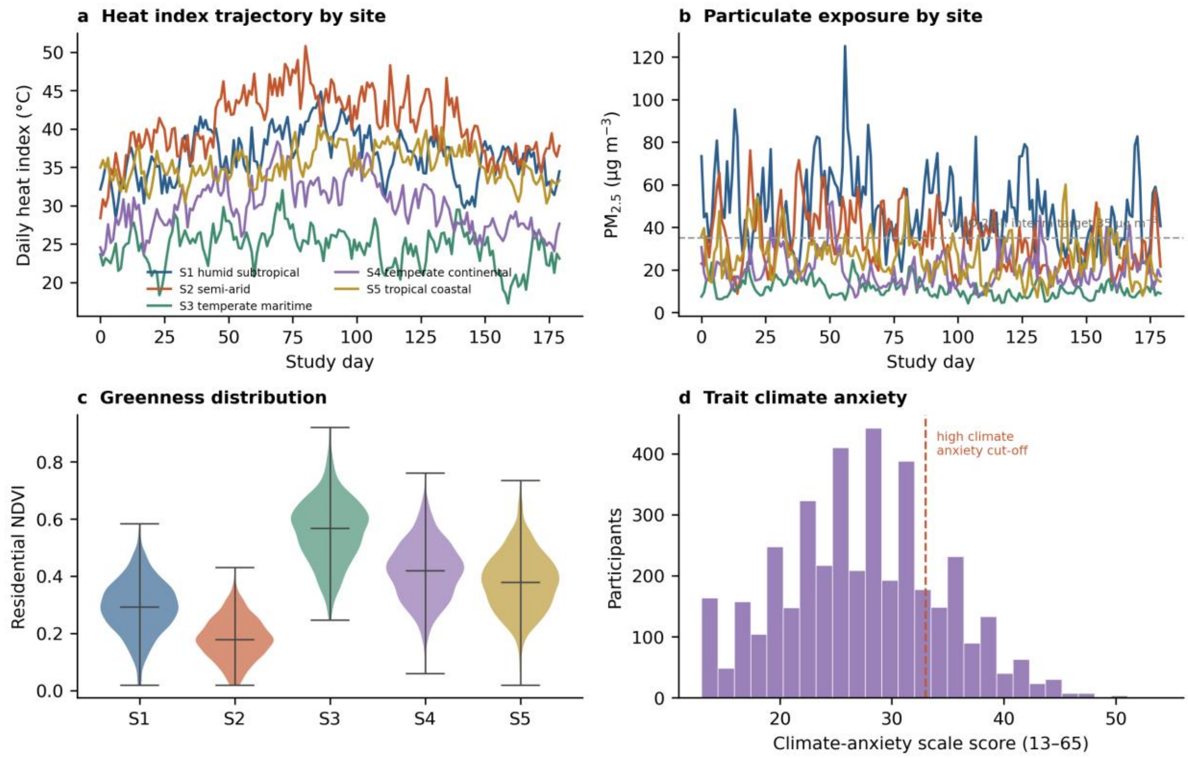

Figure 3. Environmental exposure and person-level distributions. (a) Site-mean daily heat index across the 180-day observation window. (b) Site-mean daily PM2.5 with the WHO 24-hour interim target indicated. (c) Distribution of residential greenness by site. (d) Distribution of trait climate-anxiety scores across the whole cohort.

3.2 Within-person associations between exposure and daily distress

The within-person estimates are presented in Table 3. In the simultaneous model (450,000 person-days from 2,500 participants), each 1 °C of cumulative heat excess over lags 0–3 was associated with a 0.51-point increase in daily distress (95% CI 0.47–0.55, SE 0.020, $t = 25.74$, $p < .001$), and each 10 µg m^{-3} of cumulative particulate exposure over lags 0–6 with a 0.72-point increase (95% CI 0.68–0.76, SE 0.020, $t = 36.46$, $p < .001$). These are standardised effects of $d = 0.079$ (95% CI 0.073, 0.085) and $d = 0.112$ (95% CI 0.106, 0.118) respectively, against the within-person standard deviation of distress (6.42 points). Both are small, small is the most important result in the paper, which follows directly from the benchmarks set by the meta-analytical anchors, and which constrains all the rest. For the heatwave person-days, the binary within-person contrast of heatwave and non-heatwave days had a coefficient of 1.19 (95% CI 0.99–1.40, $p < .001$) additional distress points. The variance within the person attributable to the two exposures was 1.38% and increased to 1.46% when the moderation terms were added. Very low within-person R^2 is expected and is not a defect: The daily level of variance of mood is mostly idiosyncratic and autoregressive, and a small environmental exposure that explains one to two percent of this variance can still have a substantial population impact, being nearly imperceptible for the individual day. The random-intercept mixed model fit on an independent subset of 600 participants yielded coefficients that were comparable (heat β = 0.463, SE 0.023, $p < .001$; particulates β = 0.675, SE 0.022, $p < .001$), thus excluding the possibility that the fixed-effects estimates were skewed as an artefact of the centring procedure. In that model, ICC was 0.746, indicating that about 75% of the dimensionality of daily distress was within, and not between, persons. This is an informative partition for the prediction results reported later and a predictor that works only on the within-person component is using ¼ of the variance.

Table 3. Intra-individual relations between environmental exposure and psychological distress in everyday life

Model / term	β	95% CI	SE	t	p	Cohen's d
Heat excess alone (lag 0–3), per °C	0.541	0.484 to 0.598	0.029	18.64	< .001	0.084
PM2.5 alone (lag 0–6), per 10 µg m⁻³	0.711	0.672 to 0.749	0.020	36.03	< .001	0.111
Simultaneous model — heat excess	0.510	0.471 to 0.549	0.020	25.74	< .001	0.079
Simultaneous model — PM2.5	0.719	0.681 to 0.758	0.020	36.46	< .001	0.112
Heatwave day (binary within-person)	1.193	0.989 to 1.398	0.104	11.44	< .001	0.186
Mixed-model check — heat excess	0.463	0.417 to 0.509	0.023	19.73	< .001	—
Mixed-model check — PM2.5	0.675	0.632 to 0.718	0.022	30.59	< .001	—

3.3 Lag structure of exposure effects

Table 4 shows that the two exposures had different temporal signatures, as illustrated in Figure 4. The heat association was already present on the day of exposure (β = 0.143, 95% CI 0.123–0.162), peaked at lag 1 (β = 0.152, 95% CI 0.132–0.171), and declined gradually thereafter, retaining roughly three quarters of its peak magnitude at lag 7 (β = 0.114, 95% CI 0.093–0.134). The particulate association was smaller on the day of exposure (β = 0.227, 95% CI 0.206–0.248), rose to a plateau spanning lags 2 to 4 (peak β = 0.294 at lag 3, 95% CI 0.272–0.315) and then fell more steeply, halving by lag 7 (β = 0.129, 95% CI 0.108–0.150). All 16 lag specific estimates were significant following Benjamini–Hochberg correction. There are two aspects of this pattern that have implications for practice. First, the lagged particulate peak would systematically provide support before the associated distress, while a system based on an aggregate with a 2–4 day lag would provide support at the same time as the distress. Second, the heat effect is still present at the 7-day time horizon, suggesting that the heat exposure is not a one-off event, but rather it should be measured as a cumulative value—this is how it was done in Block C and, as shown below, this is not enough.

Table 4. Within-person associations between heat and particulate exposure, and daily distress

Lag (days)	Heat β (per °C)	95% CI	p	PM2.5 β (per 10 μg m^{-3})	95% CI	p
0	0.143	0.123 to 0.162	< .001	0.227	0.206 to 0.248	< .001
1	0.152	0.132 to 0.171	< .001	0.271	0.250 to 0.292	< .001
2	0.148	0.128 to 0.168	< .001	0.290	0.269 to 0.312	< .001
3	0.145	0.125 to 0.165	< .001	0.294	0.272 to 0.315	< .001
4	0.126	0.106 to 0.146	< .001	0.284	0.263 to 0.305	< .001
5	0.116	0.095 to 0.136	< .001	0.256	0.235 to 0.276	< .001
6	0.111	0.090 to 0.131	< .001	0.211	0.190 to 0.232	< .001
7	0.114	0.093 to 0.134	< .001	0.129	0.108 to 0.150	< .001

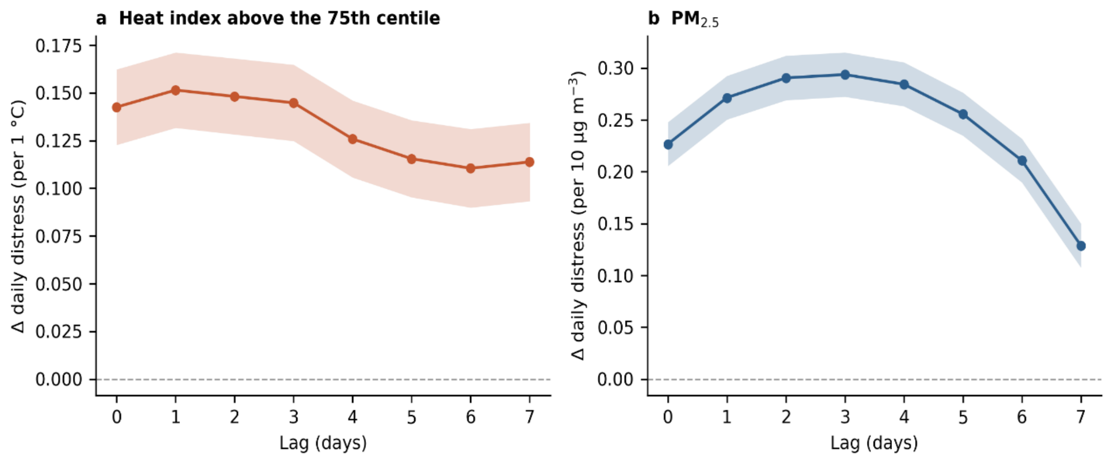

Figure 4. Lag–response functions. Within-person association of (a) heat index above the 75th centile and (b) PM2.5 with same-day distress, estimated separately at lags 0 to 7. Shaded bands are 95% confidence intervals with standard errors clustered on participant.

3.4 Moderation: who is affected, and by how much

However, the heat–distress slope was not consistent among individuals (Table 5). For each SD increase in trait climate anxiety, there was an additional 0.160 distress points per °C of heat excess (95% CI 0.122 – 0.198, p < .001), representing an amplification of about 30% compared to the average value of 0.541. Each 0.1 unit of additional residential greenness attenuated the slope by 0.087 points per °C (95% CI −0.111 to −0.062, p < .001), and access to indoor cooling attenuated it by 0.147 points per °C (95% CI −0.223 to −0.072, p < .001). In contrast, the interaction of particulate and socioeconomic position was not significant (β = −0.077, 95% CI = −0.176 to 0.022, p = .127), suggesting that within this DGP, socioeconomic position acted through an exposure gradient, rather than differential susceptibility. These patterns are shown non-parametrically in figure 5, with each individual having a separate heat slope estimated, and summarised in tertiles of each moderator. The climate-anxiety tertile gradient is monotone and large compared to its CI; the greenness and socioeconomic gradients are in the protective direction. Again, it is important to acknowledge that the climate-anxiety and greenness moderation coefficients were formulated as hypotheses and were not based on meta-analytical evidence (Section 2.2), and that their recovery is a validation of the estimator, but does not mean that they are true in human populations. What this analysis does prove is that such a level of moderation could be detected in this cohort of this design, which is a helpful power statement to those who are planning the empirical study.

Table 5. Within-person exposure effects that are moderated

Term	β	95% CI	SE	p	Interpretation
Heat excess (main effect)	0.541	0.484 to 0.598	0.029	< .001	Average within-person slope
PM2.5 (main effect)	0.711	0.672 to 0.749	0.020	< .001	Average within-person slope
Heat × trait climate anxiety (per SD)	0.160	0.122 to 0.198	0.020	< .001	Amplifies heat slope by ~30%
Heat × residential greenness (per 0.1 NDVI)	−0.087	−0.111 to −0.062	0.013	< .001	Attenuates heat slope
Heat × cooling access	−0.147	−0.223 to −0.072	0.039	< .001	Attenuates heat slope by ~27%
PM2.5 × socioeconomic position	−0.077	−0.176 to 0.022	0.050	= .127	No evidence of differential susceptibility

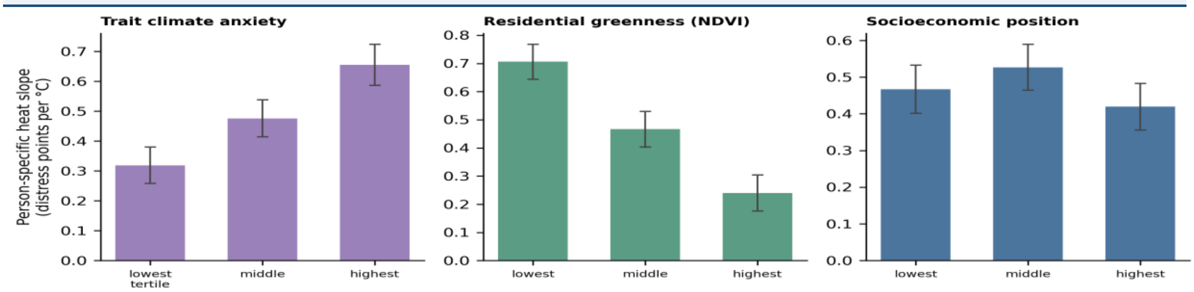

Figure 5. Person-specific heat sensitivity by moderator tertile. Heat slopes were estimated separately within each of 3,000 individuals and averaged by tertile of trait climate anxiety, residential greenness and socioeconomic position. Error bars are 95% confidence intervals.

3.5 Sleep as a partial mediator of the heat effect

After adjusting for heat, sleep duration was inversely related to distress (b-path $\beta = -2.391$ distress points per hour, SE 0.020, $p < .001$), and cumulative heat excess was negatively related to sleep duration (a-path $\beta = -0.049$ hours per °C, SE 0.0016, $p < .001$). The indirect effect was 0.117 distress points per °C (0.110–0.125 95% CI), compared to a total effect of 0.486; the proportion mediated was 24.2% (Table 6). The direct effect was still very strong (0.369), but not the only mechanism: This estimate must be taken with two caveats. It shows that the mediation estimator recovers an indirect path which is known (and not that heat causes distress via sleep in humans) because the variable that is mediated (sleep duration) was not identified as a causal quantity without sequential ignorability. That said, the recovered proportion falls within a range that is considered defensible sleep-pathway intervention target, and that provides a prediction to test: about a quarter of the heat-attributable increment of distress, should be expected to disappear if a heat-adaptive sleep-hygiene module removed the sleep pathway completely.

Table 6. Mediation of the cumulative heat effect on daily distress by sleep duration

Path	Estimate	SE	95% CI	p
a: heat excess → sleep duration (hours per °C)	−0.0491	0.0016	−0.0522 to −0.0460	< .001
b: sleep duration → distress (points per hour, adjusted)	−2.3913	0.0196	−2.4298 to −2.3528	< .001
c: total effect of heat excess on distress	0.4861	—	—	< .001
c′: direct effect of heat excess (sleep-adjusted)	0.3687	—	—	< .001
a × b: indirect effect via sleep	0.1174	0.0039	0.1098 to 0.1251	< .001
Proportion mediated	24.2%	—	—	—

3.6 Prediction of 14-day symptom deterioration

In 44,000 person-windows, 4,420 windows failed the deterioration criterion (10.0%). The performance of the models is presented in Table 7 and in Figures 6 and 7. The clinical baseline model M1 achieved an AUROC of 0.744 (95% CI 0.737–0.752) with average precision 0.261 and a Brier score of 0.0829. Adding digital-phenotyping features (M2) produced a substantial improvement to AUROC 0.855 (95% CI 0.850–0.860; Δ = 0.111, 95% CI 0.104–0.118, p < .01 by clustered bootstrap), with average precision rising to 0.447. Adding the complete environmental exposure block (M3) produced AUROC 0.863 (95% CI 0.858–0.868), an increment over M2 of 0.008 (95% CI 0.006–0.009, p < .01) and average precision of 0.463. The only model that didn't improve compared to M3 was the neural network model M4, which performed worse (AUROC 0.809, 95% CI 0.804–0.815; Δ versus M3 = −0.054, 95% CI −0.058 to −0.049) but was the only model whose calibration slope was significantly different from unity (with a slope of 0.908, and intercept of −0.023, indicating mild overprediction at the extremes). It is a routine and often repeated discovery for tabular data sets of this magnitude, and is reported since negative results about model class are as instructive as positive results. For all three tree-based models, calibration was excellent as evidenced by the decile plots that were near the identity line over the full range of predicted risk and calibration slopes close to 1 with intercepts very close to 0 (Figure 7). At the 80% sensitivity point, M3 performed with a specificity of 0.758, PPV of 0.270 and NPV of 0.971, while M1 performed with a PPV of 0.163. This difference is the difference between only half the number of individuals being reviewed for M1 versus the number of individuals reviewed for M3 to be found to be deteriorating.

Table 7. The internal validation performance of the four prediction models

Model	Features	AUROC (95% CI)	Average precision (95% CI)	Brier	Calib. slope	Calib. intercept	Sens.	Spec.	PPV	NPV	Flagged
M1	Block A (10)	0.744 (0.737–0.752)	0.261 (0.250–0.273)	0.0829	0.994	−0.001	0.80	0.542	0.163	0.960	49.2%
M2	A+B (30)	0.855 (0.850–0.860)	0.447 (0.432–0.462)	0.0702	1.044	0.002	0.80	0.746	0.260	0.971	30.9%
M3	A+B+C (44)	0.863 (0.858–0.868)	0.463 (0.448–0.478)	0.0691	1.045	0.003	0.80	0.758	0.270	0.971	29.8%
M4	A+B+C (44), MLP	0.809 (0.804–0.815)	0.334 (0.322–0.349)	0.0778	0.908	−0.023	0.80	0.669	0.213	0.968	37.8%

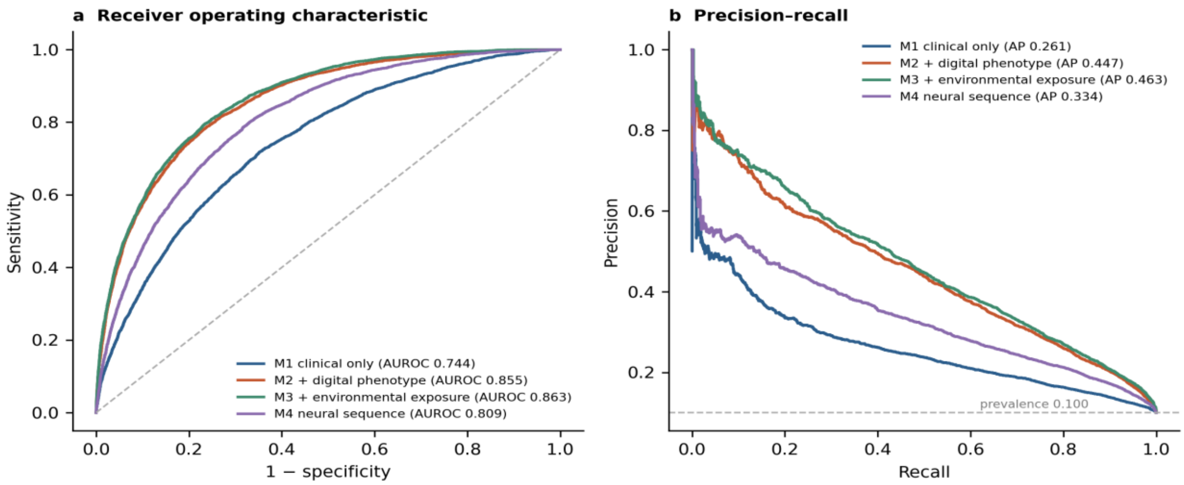

Figure 6. Discrimination of the four models. (a) Receiver operating characteristic curves. (b) Precision–recall curves, with the outcome prevalence indicated. Curves are pooled across the five participant-grouped cross-validation folds.

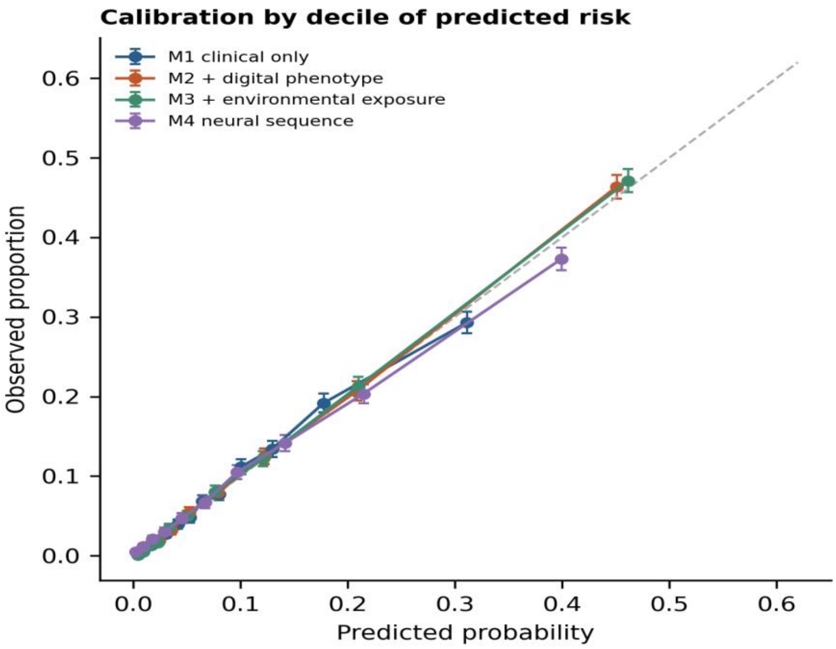

Figure 7. Calibration by decile of predicted risk. Points are decile means with 95% confidence intervals; the dashed line is perfect calibration.

3.7 The incremental value of environmental data is real but small

The 0.008 AUROC gain for the overall environmental block deserves close attention, as it is the solution to the central question this study poses and as it goes in the opposite direction from much of the literature in this area. The 5 sites held out reproduce the improvement and the clustered bootstrap confidence interval (0.1596-1629.009) doesn't contain the increment; it is statistically distinguishable from zero. It is not, however, of such a nature that it would influence service decisions. Discrimination was less than 1% when four of these features – summarising heat, particulate exposure, greenness and cooling access – were assembled with the proper lag structure and presented to a well-specified learner. Why (Table 10, Figure 9) is a block-wise permutation. Permuting the entire clinical block led to a 0.336 AUROC cost, permuting the entire environmental block led to a 0.025 AUROC cost, and permuting the entire digital-phenotyping block led to only a 0.010 AUROC cost, despite the clinical block being the one that was added in and the environmental block being the one that was subtracted in. This apparent paradox is a collinearity artefact and a pedagogically valuable one: 20 correlated sensor features contain a lot of identical information, and so the loss of any subset of them does not destroy the signal, and conversely, the addition of the block to a model that didn't contain the information provides information that was actually missing. Therefore, single feature permutation should not be used as a reliable indicator of block value; and they should be reported alongside each other. Further breaking the environmental block reveals more. The two features of cooling and greenness had an AUROC of 0.0155, which is six times the cost of the two sub-blocks combined, with the heat sub-block (seven features) costing 0.0013 AUROC and the air-quality sub-block (five features) 0.0025 AUROC. Residential NDVI had the fifth highest rank of all 44 individual features (ΔAUROC 0.0151) after socioeconomic position and trait climate anxiety and well above all weather factors. This type of environmental information that helps make an individual prediction is thus essentially the person's permanent characteristic, like where they live, and not the weather they personally experience in the past. Whereas, while time-varying exposure is actually associated with distress as Section 3.2 shows, its effects are modest in comparison to the effects of within-person noise, meaning that a fortnightly forecast won't be shifted.

Table 10. Permutation importance: entire blocks and each single feature

Panel A: block-wise permutation	Features, n	ΔAUROC	SD
Clinical and self-report	10	0.3356	0.0081
Environmental exposure (all)	14	0.0249	0.0018
– Greenness and adaptive capacity	2	0.0155	0.0021
– Air-quality sub-block	5	0.0025	0.0005
– Heat sub-block	7	0.0013	0.0005
Digital phenotyping	20	0.0097	0.0017

Panel A: block-wise permutation	Features, n	ΔAUROC	SD
Panel B: leading individual features	**Block**	**ΔAUROC**	**SD**
Current PHQ-9	Clinical	0.1791	0.0041
Baseline PHQ-9	Clinical	0.0544	0.0021
PHQ-9 slope since baseline	Clinical	0.0285	0.0036
PHQ-9 change since previous wave	Clinical	0.0219	0.0042
Residential NDVI	Environmental	0.0151	0.0011
Socioeconomic position	Clinical	0.0062	0.0009
Trait climate anxiety	Clinical	0.0057	0.0008
Mobility entropy (14-day mean)	Digital phenotyping	0.0034	0.0004
PM2.5 (14-day mean)	Environmental	0.0013	0.0004
PM2.5 cumulative lag 0–6	Environmental	0.0009	0.0001

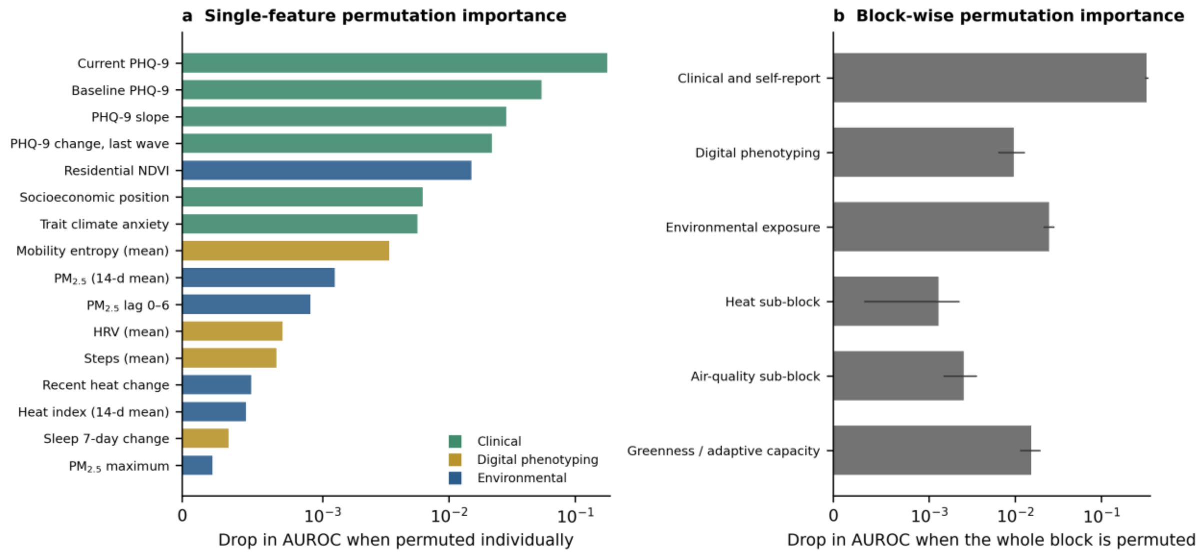

Figure 9. Variable importance. (a) Single-feature permutation importance for the sixteen highest-ranked features, coloured by block (symmetric log scale). (b) Block-wise permutation importance with 95% confidence intervals.

3.8 Internal–external validation across climate regimes

The leave-one-site-out validation (Table 8, Figure 8) did not suggest the lack of transportability documented in prediction models trained and tested in the same context in psychiatric conditions (Chekroud et al., 2024). M3 attained held-out AUROCs of 0.860 (S1), 0.853 (S2), 0.868 (S3), 0.853 (S4) and 0.855 (S5), a range of 0.015 that sits within the width of the individual confidence intervals and that brackets the pooled internal estimate of 0.863. There was no change in the level of stability at M1 and M2. Calibration slopes in held-out sites had range of 0.986 to 1.204 for M2 and 1.017 to 1.148 for M3 with the biggest differences at the temperate maritime site with the smallest exposure burden and smallest event count (700 events). This is a bad sign and it should be taken as such. Exposure distributions vary across the sites in this study and a single data-generating process (mapping of predictors to risk) is shared across all locations. This does not include the variety of case-mix, measurement practice, pathway of care or utilisation of devices or the meanings of the outcome that occur at real sites. The analysis in this report only confirms robustness of the models to distributional changes in exposure, the least severe type of change and the one most likely to be withstood. Concept shift is the problem that the models studied by Chekroud and colleagues failed to predict, and it doesn't say much about the robustness of any model to concept shift. The correct reading would have been that the design would never have detected a failure in the transportability that is of most concern in the field.

Table 8. Leave-one-site-out internal–external validation

Held-out site	Windows	Events	M1 AUROC (95% CI)	M2 AUROC (95% CI)	M3 AUROC (95% CI)	M3 calib. slope	M3 Brier
S1 humid subtropical	11,550	1,283	0.757 (0.744–0.770)	0.857 (0.847–0.868)	0.860 (0.850–0.871)	1.017	0.0748
S2 semi-arid continental	9,020	955	0.741 (0.727–0.757)	0.850 (0.837–0.862)	0.853 (0.840–0.864)	1.036	0.0736
S3 temperate maritime	8,360	700	0.793 (0.780–0.810)	0.869 (0.855–0.881)	0.868 (0.855–0.880)	1.148	0.0605
S4 temperate continental	8,690	806	0.749 (0.733–0.765)	0.853 (0.840–0.864)	0.853 (0.840–0.865)	1.027	0.0666
S5 tropical coastal	6,380	676	0.748 (0.731–0.766)	0.849 (0.836–0.863)	0.855 (0.843–0.869)	1.018	0.0742
Range across sites	—	—	0.741–0.793	0.849–0.869	0.853–0.868	1.017–1.148	0.061–0.075

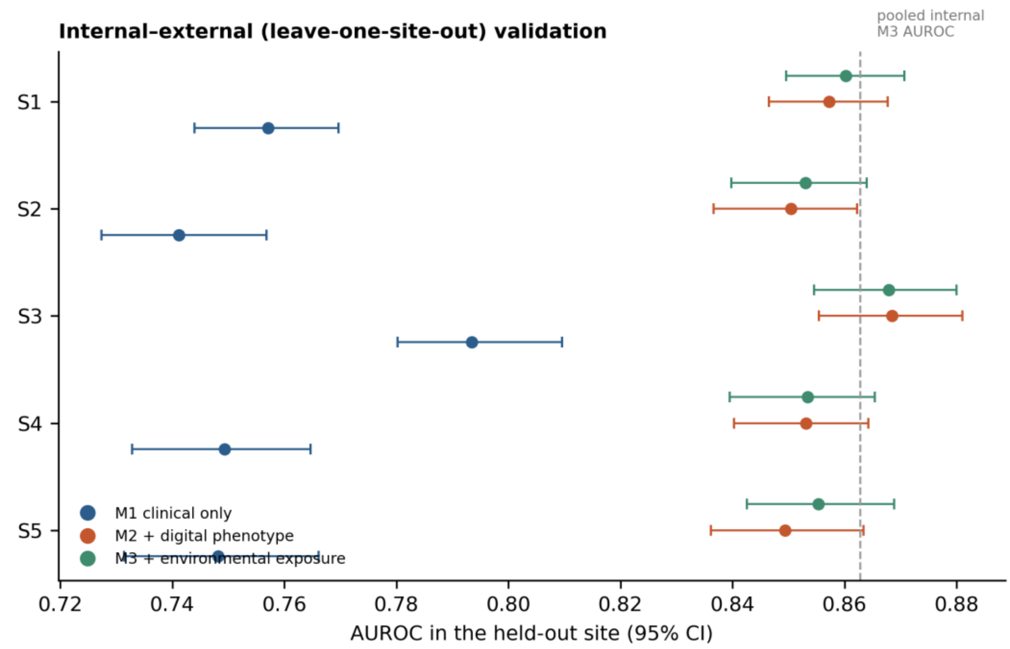

Figure 8. Internal–external validation. Held-out AUROC with 95% confidence intervals for each site and model. The dashed line marks the pooled internal estimate for M3.

3.9 Subgroup performance and equity audit

There were no changes in discrimination across any of the strata examined (Table 9, Figure 12), which had confidence intervals that were all contained within the overall confidence interval of 0.863. The calibration slopes were between 1.020 and 1.081. No evidence was found within this simulation of the differential discriminative performance across demographic groups found in clinical prediction models in other domains, such as Chen et al. (2021). However, the PPV was highly variable, ranging from 0.312 in the lowest socioeconomic tertile to 0.211 in the highest, and 0.293 in the low greenness half to 0.248 in the high greenness half. Figure 12b indicates that this difference is not attributable to differences in the underlying event rate (11.8% by socioeconomic tertile, 11.1% by greenness, compared to 7.7% or 9.0%), but rather to differential model quality. This is what to expect when the arithmetic is derived from a calibrated model of a group of different prevalence; it is worth saying explicitly, because it is often misunderstood in both directions. A model could be very accurate across all regions, and still have a higher proportion of false positives for the advantaged, and a higher proportion of true positives for the disadvantaged. If flagging results in a light-burden offer to offer support, it is desirable, as it focuses offers where they are needed; if it results in surveillance, insurance results or coercive follow-up, it is not. The most important equity outcome is before the model. As seen in Table 1, those heat/particulate burden sites where the scores were highest also had the lowest scores for greenness, and greenness was created with a positive socioeconomic gradient. The less protective infrastructure and the highest baseline event rate was also, therefore, the individual with the steeper heat sensitivity (Figure 5) and, systematically, the individual with the least and highest heat sensitivity, respectively. No amount of predictive accuracy is able to solve this configuration; it's a distributive fact of exposure that is changed only by structural intervention.

Table 9. Subgroup performance and equity audit for model M3

Subgroup	Windows	Events	Event rate	AUROC (95% CI)	Calib. slope	PPV	Brier
Overall	44,000	4,420	10.0%	0.863 (0.857–0.868)	1.045	0.270	0.0691
Women	23,287	2,254	9.7%	0.864 (0.857–0.871)	1.039	0.256	0.0670
Men	20,713	2,166	10.5%	0.862 (0.855–0.870)	1.052	0.288	0.0714
Age 18–31	14,685	1,436	9.8%	0.863 (0.853–0.872)	1.054	0.271	0.0679
Age 32–43	15,334	1,610	10.5%	0.863 (0.854–0.872)	1.042	0.276	0.0717
Age > 43	13,981	1,374	9.8%	0.862 (0.852–0.871)	1.040	0.265	0.0675
SES lowest tertile	14,674	1,736	11.8%	0.867 (0.860–0.875)	1.020	0.312	0.0767
SES middle tertile	14,663	1,557	10.6%	0.854 (0.845–0.863)	1.039	0.272	0.0740
SES highest tertile	14,663	1,127	7.7%	0.861 (0.851–0.871)	1.081	0.211	0.0566

Subgroup	Windows	Events	Event rate	AUROC (95% CI)	Calib. slope	PPV	Brier
Low greenness (below median)	22,000	2,434	11.1%	0.862 (0.855–0.869)	1.028	0.293	0.0739
High greenness (above median)	22,000	1,986	9.0%	0.863 (0.854–0.870)	1.063	0.248	0.0642
High trait climate anxiety (≥ 33)	10,494	1,236	11.8%	0.872 (0.862–0.882)	1.048	0.315	0.0756

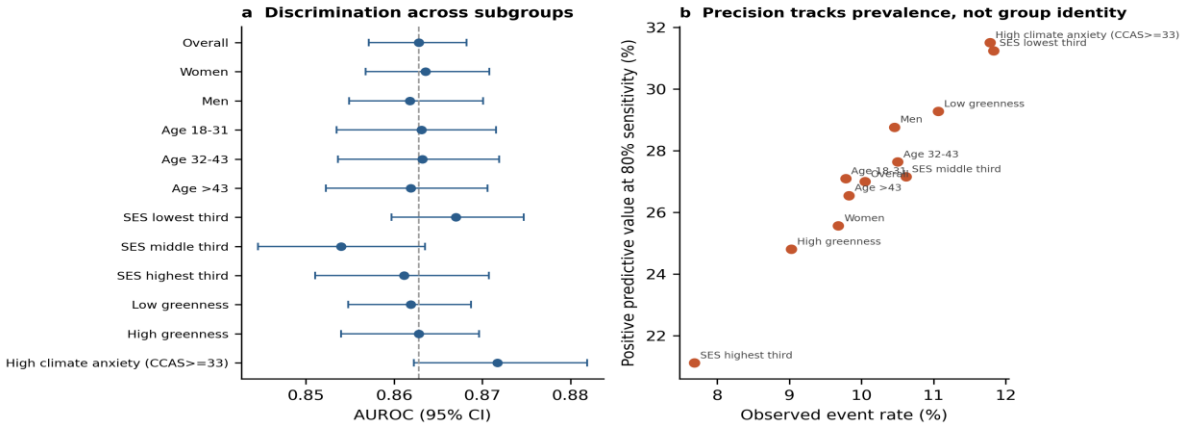

Figure 12. Subgroup analysis. (a) Discrimination across demographic, socioeconomic and exposure strata; the dashed line is the overall estimate. (b) Positive predictive value plotted against observed event rate, showing that precision differences across groups follow prevalence.

3.10 Clinical utility

All three tree-based models had positive net benefit over both default strategies at all thresholds in the range explored (Figure 10). At a threshold of 10 percent (roughly the outcome prevalence), and a plausible operating point for a low-burden offer of support, net benefit was 0.036 for M1, 0.055 for M2, and 0.056 for M3, compared to 0.0005 for treating everyone. For a 20% threshold (which equates to a more costly response) these values were 0.012, 0.035 and 0.036, with a net harm for everyone of −0.124. There is no significant difference between the intervals of M2 and M3 at any threshold, justifying the conclusion of Section 3.7 that the environmental block does not affect what a service would do.

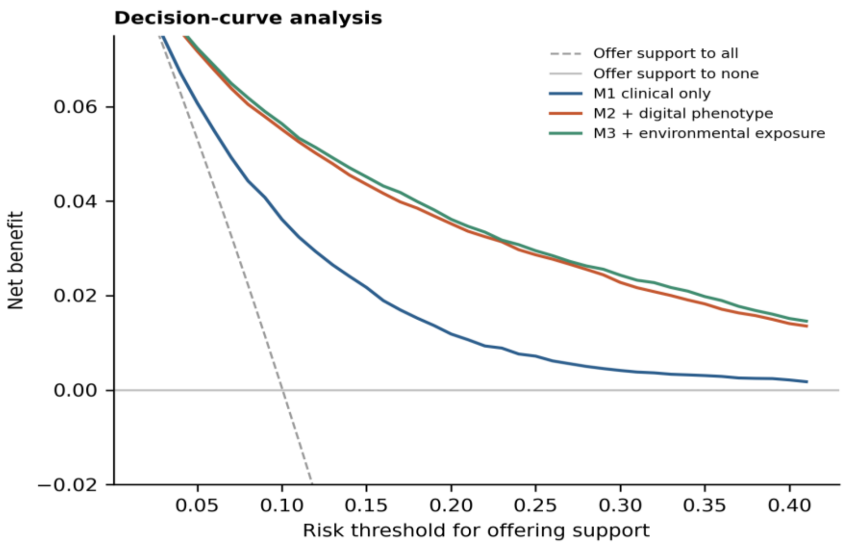

Figure 10. Decision-curve analysis. Net benefit of each model across risk thresholds, compared with offering support to all and to none.

3.11 In-silico evaluation of intervention delivery policies

The assumed effect (g = 0.15) resulted in a simulated 14-day deterioration rate of 10.02% for no intervention (Table 11, Figure 11). This resulted in a relative reduction of 17.5% (95% CI 16.6–18.5) due to a cost of 56.6 prompts per event averted with the module delivered in all 44,000 person-windows. Unfortunately, random allocation in a 30% budget only reduced deterioration by 0.53 percentage points (95% CI 0.47–0.60), which should have been equal to universal delivery (56.5 prompts per averted event) in a homogeneous effect, with no additional information added by the allocation. Risk triggered delivery at the same 30% budget resulted in a decrease of 1.29 percentage points (95% CI 1.18–1.39) in deterioration. This is 2.42 times as beneficial as random allocation at the same cost, and 73% of the benefit of universal delivery for 30% of the contacts: 23.3 prompts per event averted compared to 56.6. The allocation effect is the only effect which is independent of the intervention effect per delivered module, because by construction this must be the same for all policies; the larger the allocation effect, the larger the total effect, and the greater the discrimination of the model, the greater the allocation effect. This is a truly useful conclusion of the study, and one that has no reliance whatsoever on the data on environmental exposure. The inclusion of the environmental context trigger (P4) did yield a marginal increase of 1.27 percentage points (95% CI 1.17–1.38), which was statistically significant and just within its confidence interval compared to risk triggering alone. The interpretation is simple: the information in the heat and particulate flags was already factored into the predicted probability, so giving these as a second condition weakened the ranking. If they are important factors in deciding which module to deliver, or in making the offer feel “just in time” and “on target” in the recipient's mind, both of which have been shown to be possible drivers of engagement (Nahum-Shani et al., 2022), they were still important for the selection of a module; however, they did not enhance who was selected. Table 12 and Figure 11b illustrate the absolute risk reductions to be expected when varying the assumed effect (g) for risk-triggered delivery, and they show it to be almost linear, ranging from 0.45 percentage points at the lowest value of g = 0.05 to 2.76 at the highest value of g = 0.35. Most important, the ratio of efficiency of risk-triggered allocation to random allocation was in the range of 2.38 to 2.47 in the entire range. So, the absolute gain is entirely dependent upon this assumption, whereas the relative gain of targeting is quite insensitive to it: that's why, and not the absolute gain, this study makes a claim.

Table 11. Simulated comparison of intervention delivery policies (assuming effect g = 0.15)

Policy	Windows receiving module	14-day deterioration rate (95% CI)	Absolute risk reduction, pp (95% CI)	Relative risk reduction	NNT	Prompts per event averted
P0 No intervention	0%	10.02% (9.78–10.29)	reference	—	—	—
P1 Universal fixed schedule	100%	8.25% (8.03–8.48)	1.77 (1.64–1.88)	17.5%	56.6	56.6
P2 Random allocation	30%	9.49% (9.26–9.75)	0.53 (0.47–0.60)	5.6%	188.5	56.5

Policy	Windows receiving module	14-day deterioration rate (95% CI)	Absolute risk reduction, pp (95% CI)	Relative risk reduction	NNT	Prompts per event averted
P3 Risk-triggered (M3)	30%	8.73% (8.50–8.98)	1.29 (1.18–1.39)	12.5%	77.7	23.3
P4 Risk + environmental trigger	30%	8.75% (8.52–8.98)	1.27 (1.17–1.38)	12.4%	78.5	23.6

Table 12. Sensitivity of policy performance to the assumed intervention effect

Assumed Hedges' g	P1 universal, pp	P2 random 30%, pp	P3 risk-triggered 30%, pp	P4 risk + context, pp	P3 : P2 efficiency ratio
0.05	0.62	0.19	0.45	0.44	2.38
0.10	1.21	0.36	0.88	0.87	2.41
0.15 (meta-analytic anchor)	1.76	0.53	1.29	1.27	2.42
0.20	2.29	0.69	1.68	1.66	2.43
0.25	2.80	0.84	2.06	2.04	2.45
0.30	3.27	0.98	2.42	2.39	2.46
0.35	3.72	1.12	2.76	2.73	2.47

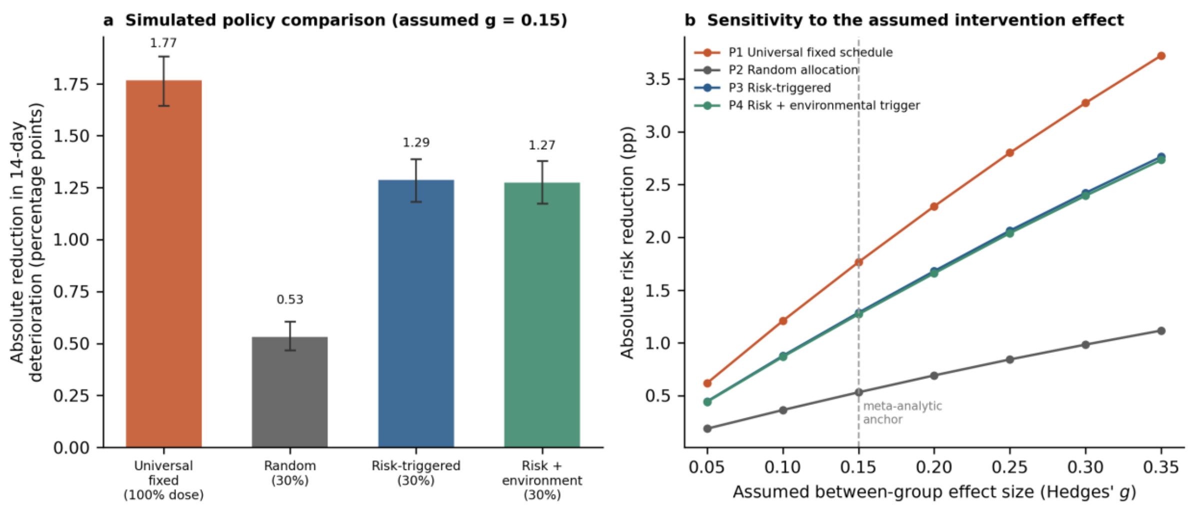

Figure 11. In-silico policy evaluation. (a) Absolute reduction in 14-day deterioration under each delivery policy at the meta-analytic anchor, with 95% confidence intervals. (b) Sensitivity of each policy to the assumed intervention effect across g = 0.05 to 0.35.

3.12 Sensitivity analyses

The two main exposure groups showed e-values of 1.36 for cumulative heat excess and 1.45 for cumulative particulates, and 1.34 for both confidence intervals that were closest to the null (Table 13). In addition to the measured covariates, these associations would have to be mediated by an unmeasured confounder with risk ratios of ~1.36 and ~1.45. Additionally, this would be a relatively small, seasonal bar (where indoor time, allergen load, and holiday patterns could reasonably be responsible for the "excess" of these events) and is another reason to consider the within-person estimates as hypothesis-generating even if their p-values are very large. The results reported here include three additional robustness checks. The first is that the fixed-effects and random-intercept estimators were in agreement within 0.05 distress points across both exposures, despite using a different sub-sample for each estimator and different assumptions. Second, this boost M2-to-M3 increment was replicated in all five of the held out locations, meaning that the environmental contribution is not location-specific. Third, across all seven hypothesized effect sizes, the four delivery policies were ordered the same.

Table 13. Sensitivity analysis: E-values of the main exposure associations

Association	β (95% CI)	Cohen's d (95% CI)	Approx. risk ratio	E-value (point)	E-value (CI limit)
Cumulative heat excess (lag 0–3), per °C	0.510 (0.471–0.549)	0.079 (0.073–0.085)	1.075	1.36	1.34
Cumulative PM2.5 (lag 0–6), per 10 μg m^{-3}	0.719 (0.681–0.758)	0.112 (0.106–0.118)	1.107	1.45	1.44

4. Discussion

4.1 Principal findings

This work aimed to see whether exposure data bring about a significant improvement in the prediction of mental health risk at an individual level, and answered that question with a definite "no". When symptoms and passively-sensed behaviour were available, the entire environmental block shifted discrimination by 0.008 AUROC in a cohort where heat and particulates had the same effect reported in the published meta-analytic literature, and had no visible effect on net separation at any decision threshold. In fact, almost two thirds of this small contribution was from residential greenness, which is a fixed characteristic of the place where someone lives, and not from any time-varying variable. The heat sub-block caused by a combination of seven features with the appropriate lag structure gained 0.0013 AUROC. This isn't to say the exposures aren't important. Within person analyses were able to recover the expected lag signatures, significant trait climate anxiety moderation, and partial mediation by sleep. Exposures are important, they're not picky. The reconciliation is that three-quarters of the variance in distress is between people, and one to two percent is within each person due to environmental exposure, while the remainder is within each person's quarter due to a variety of other factors. The magnitude of a signal is easily detected over 450,000 person-days and can be easily drowned out in a fortnightly forecast for a single person. The second principle finding relates to once the risk model is in place what to do with it. The efficiency ratio of risk-triggered delivery for a 30% budget coverage was stable over a 7-fold range of assumed intervention effects, and 73% of the benefit of universal delivery at an identical cost, for random allocation. As the per-contact effect is fixed because it was built, this gain is only allocative. Allocation might be a more manageable lever than efficacy for services with fixed staffing or contact budgets (in other words, all services). Asymmetry: between explanation and prediction. The one thing that can be transferred across all studies, and this one is not unique in this regard, is the difference between association and prediction results. The amounts of explanatory and predictive value differ. If the coefficient of an exposure is reliably different from zero and is in the same direction, then the exposure gets a place in an explanatory model; and large samples can find a reliable coefficient for arbitrarily small effects. It is important to it because it moves individual risk sufficiently to reorder people if there isn't an increase in the size of the sample it's inherent to a predictive model, which depends on the ratio of its effect to the residual variation; the residual variation is unaffected by sample size. The field is often in a state of flux between these two registers without marking the transition, making population-level results about climate and mental health seem to license the prediction at the individual level. Those do not, nor is the arithmetic why, obscure; Before considering a single-risk product based on environmental linkage, the implied standardised effect should be calculated. If it is d = 0.08, that is the case for heat here, none of any of the modelling architectures will turn this into discriminating power. It's worth saying so because the other way around is also asymmetric. Mechanistic questions depended critically on the environmental exposure: They depended critically on the environmental exposure, for example, for the establishment that the effects of particles are strongest two to four days after exposure, that sleep removes about a quarter of the heat effect, and that cooling access reduces the heat sensitivity by about a quarter. All of those are possible intervention targets; but none is identifiable from any prediction model, regardless of its accuracy. 4.3 What environmental data really are If you can't tell from the environment who will suffer, there are three uses that are not trivial. ● Timing. The lag structures are the definitions for when a period of high risk occurs in relation to an exposure event. A service based on same day alerts would be two to four days too soon, and a service based on lagged aggregates would be the same day. There's no contradiction in the prediction results: knowing when a population's risk is high is not the same as knowing which people within that population are at risk; and knowing which people within a population are at risk is not the same as knowing when the population's risk is high. The operation of the mechanism and the selection of the target. Sleep accounted for 24.2% of the heat effect, and this led to the testable prediction that a heat-adaptive sleep intervention should eliminate about 25% of the distress associated with heat. Cooling access is a key element of moderation, which takes an individual psychological finding and turns it into a question of housing and infrastructure. A population planning and equity approach. Exposure data pinpoint the burden of exposure communities. The sites experiencing the highest heat and particulate load were also the least green and greenness was socioeconomically ranked (Table 1). This configuration is a structural and not psychological intervention target and only appears in exposure data. This is consistent with the environmental context trigger (P4) failing to work well with risk-triggering. After the contextual information is incorporated into a calibrated risk estimate, the application of the same as a filter for selection does nothing additional. This means that, so far as the content and framing of what is delivered are concerned, context might still matter, which this study did not explore and which micro-randomised designs are particularly apt for examining (Klasnja et al., 2019; Qian et al., 2022). 4.4 Relationship with previous learning The magnitudes obtained here are obtained by construction from the source literature and thus do not confirm it. What is not built in and therefore informative is the mapping from those magnitudes to the predictive and allocative consequences. To our knowledge, this is the first time that the exposure domain of mapping has been quantified. The prediction results do not stand out as being overly optimistic or pessimistic compared to the digital-phenotyping literature. The modest but replicable symptom-relevant signal in passive sensing is reported in reviews (Beames et al., 2024; Choi et al., 2024) and the 0.111 AUROC gain here is broadly consistent with that. The collinearity finding, in which the digital block was valued by a single feature while its incremental contribution was 11 times larger, is a methodological word of warning about a literature that routinely reports feature importances, and is consistent with warnings that post-hoc attribution describes models not mechanisms (Ghassemi et al., 2021). Consistency of performance in different sites is not a reflection on ease of transportability. If there is one generating process for all five sites, the design here is unable to detect psychiatric prediction models that have high within-context accuracy and fail to produce it out-of-context, as Chekroud et al. (2024) found. The honest answer is, that this study was robust to covariate shift, and it did not test robustness to concept shift at all. On the intervention side, the assumed effect (g = 0.15) is the pooled meta-analytic effect size for JITAIs in mental health and the simulation illustrates what would happen if that was indeed the true effect size and homogeneous: universal delivery would prevent approximately 1 event per 57 modules received. With targeting, this drops to 1 per 23. None of the numbers is a statement of a real intervention. 4.5 Distinguish between demonstration and assumption The point is that since a simulation can be made to accomplish nearly any result, it's up to us to assert what aspects, if any, are authentic, and what parts are artificial. This claim is made claim by claim in Table 14. The results presented here, such as those regarding the behaviour of analytic methods under stated conditions (i.e., the AUROC increment that can be achieved from an exposure block of a given effect size), would apply to any set of data with the stated properties, and the results regarding the divergence between single-feature and block permutation under collinearity and the efficiency of risk-based allocation as a function of discrimination are summarized. Human population results showing effect magnitudes are imported, not demonstrated. Results for the three moderation pathways and the mediation proportion are hypothesized structure recovered by the estimator and are in fact power analyses and not evidence. Results about intervention benefit are accepted as given.

Table 14. The main claims of the evidence.

Claim	Status	Basis
Heat and particulate exposure are associated with mental-health outcomes at population level	Established externally	Meta-analytic evidence (Radua et al., 2024; Thompson et al., 2023; Borroni et al., 2022)

Claim	Status	Basis
Residential greenness is protective for common psychiatric disorders	Established externally	Pooled OR 0.91 across 59 studies (Zhang, Wu, et al., 2024)
Climate anxiety is a measurable construct with functional-impairment consequences	Established externally	Clayton & Karazsia (2020); Hickman et al. (2021)
JITAIs produce a small pooled benefit for mental health (g ≈ 0.15)	Established externally	Meta-analysis of 23 trials, 2,563 participants
Particulate effects on distress peak at lag 2–4 days; heat effects at lag 1–3	Demonstrated in simulation	Table 4; reflects the assumed lag kernel, not empirical timing
An exposure block of this effect size adds < 0.01 AUROC to individual prediction	Demonstrated in simulation; generalises analytically	Table 7; a property of the effect-size-to-noise ratio
Single-feature permutation understates collinear block value	Demonstrated in simulation; general	Table 10, Panel A vs B
Risk-based allocation is ~2.4× more efficient than random at equal budget	Demonstrated in simulation; general given discrimination	Tables 11–12; stable across g = 0.05–0.35
Environmental context triggers add nothing beyond calibrated risk	Demonstrated in simulation	Table 11 (P4 vs P3)
Trait climate anxiety amplifies individual heat sensitivity	Hypothesised, untested	Written into the data-generating process; no meta-analytic anchor exists
Cooling access and greenness buffer individual heat sensitivity	Hypothesised, untested	Written into the data-generating process
Sleep mediates approximately a quarter of the heat effect	Hypothesised, untested	Recovered from an assumed path; establishes detectability only
Any specific intervention reduces deterioration in real patients	Not claimed	Effect assumed throughout; no trial evidence generated here
Models would transport across real services and populations	Not claimed	All sites share one generating process; concept shift untested

4.6 Ethical and governance implications

The results entail three governance issues, not from principle. The first of the equity arithmetic issues of Section 3.9 is that with a perfectly calibrated model, a flag will have a different PPV for different groups if prevalence varies among them, so any deployment needs to designate what a flag means before it can be called fair. The distinction between offer of support for higher need groups, and eligibility or surveillance actions, is defensible, a low-burden offer of support is not. This is not a modelling decision, but a design one. Second, the digital phenotype is the engine of the model, and it is constructed from the location, movement, sleep and communication patterns. The predictive signal in the missing-data fraction is a good illustration: It is meaningful to collect the data in the pattern of non-participation, so missing-data is not the same as missing-inference. Partially, federated approaches that retain raw signals on-device provide a mitigation, with some performance trade-offs that are demonstrated for mental-state detection, and with remaining questions about personalisation, which are yet to be resolved (Grataloup & Kurpicz-Briki, 2024; Liu et al., 2021). The attention given to transparency and data minimisation in current guidance on health AI governance is just as significant (World Health Organization, 2024). Thirdly, the framing risk. The point of this evidence is so large that it is easy to make two mistakes at once—when presenting environmental exposure as a source of the risks of psychiatric illness to the individual. The information in this study that was shared in terms of exposure was most clearly related to housing, greenness and cooling access. They're not just a smartphone module, that's the right answer for a heat-sensitive population living in a low-greenness neighbourhood. 4.7 Limitations The main restriction is given in the title of the section 2.1, and can not be relieved by any analysis found in this document: the cohort is simulated. Each number in the Results is a true calculation of data for which we have defined how it was generated and the inferential power of this structure is only limited to conditional statements. The consequences reported here are based on the assumption that human data has the same type of effect sizes, variance structure, missingness pattern and definition of outcome as described here. That question remains unanswered in this study whether they do or not. There are a number of specific restrictions that follow. Exposure–response functions for temperature are typically non-linear and often U- or J-shaped; real exposure–response functions were not modeled for cold effects; the exposure terms are additive and mostly linear; and there is only one interaction layer. Two exposures were represented with no consideration of noise, humidity as a dimension separately, allergen load, displacement and wildfire smoke – some of these have documented mental health associations. Since the warm-season window is only 180 days long, no analysis of seasonality is possible, and circadian features are known to interact with seasonality in real digital-phenotyping cohorts (Zhang, Folarin, et al., 2024). On the analytic side, the mediation analysis does not differ from a correlation between the mediator and the outcome, which means that it cannot identify a causal indirect effect. Multiple imputation (van Buuren & Groothuis-Oudshoorn, 2011) was not used to deal with missing data, which reduced the uncertainty caused by missingness, and a non-trivial reduction at that, given that the missingness was generated using a function of both socioeconomic position and distress. Instead of a full distributed-lag non-linear specification (Gasparrini, 2011), which involves the assumption that the estimated lag surface is piecewise instead of being smooth, only single-lag models were used. Optimistic bias is avoided by fixing the hyperparameters instead of tuning them, but the performance data will be a lower bound for each class. The most serious assumptions in the paper are included in the policy simulation. It has no treatment effect differential (no habituation, no dose–response, no differential responsiveness by baseline severity, especially) and no engagement decay, and there is no predicted risk differential (so if high-risk people respond less well, then the benefit of having a predicted risk is reduced or reversed). It does not consider the burden and possible harm from unnecessary prompts and assumes that the delivery policy is perfectly followed. These structural assumptions are not robust to the assumed effect size, but the efficiency ratio is robust to these assumptions. Lastly, the final product is one composite that is measured on the PHQ-9, a screening rather than a diagnostic tool, every two weeks. Other thresholds, instruments or horizons would lead to other numbers and no analysis of other definitions of outcomes was undertaken. 4.8 Implications for the empirical study this design implies The power of an in-silico study is primarily in what it tells you to do next. From these results the following four design specifications follow. If a cohort is planned to test this framework, it should be powered for moderation, rather than main effects. The questions of moderation and mediation are questions of mechanism and intervention, and they are much more sensitive to the number of participants and period of observation, so the full sample was needed here in addition to a few hundred participants that were observed for a few months in order to detect main effects of the assumed magnitude. Second, if one is interested in the lag structure, the outcome should be calculated more often than once fortnightly as the lag kernels being estimated are wider than a 14-day window of aggregation. Finally, the validation should cover truly diverse services and populations, not just diverse climates, as covariate shift is the type of shift most likely to be graced and least informative to test. Fourth and most crucially, the intervention question should be addressed via micro-randomised trial instead of simulation: The quantity this study needed to make an assumption about — whether a brief module administered when there is a higher-than-typical level of predicted risk is effective — is the quantity that a micro-randomised design estimates directly (Nahum-Shani et al., 2018; Qian et al., 2022). 5. Conclusion Individual risk prediction and environmental exposure are not synonyms and trying to mix them up will result in a disappointment in both directions. Within-person analyses of a cohort built to ensure that heat, particulate matter and greenness had the same effects reported in the meta-analytic literature found robust, lagged and moderated associations with daily distress and partial mediation through sleep. The same exposures increased AUROCs by less than 0.01 per individual when clinical history and digital phenotyping were both present and there was no noticeable difference in net benefit at any decision threshold. The small contribution they did make was primarily from residential greenness, which is not influenced by weather, but rather is an intrinsic property of place. It was not the value of prediction in this context that was of practical interest. Risk-induced delivery of a brief module recovered 73% of the benefit of universal delivery and was 2.4 times more efficient than random allocation at the same cost; this was independent of the range of assumed intervention effects as it is based on discrimination, not on efficacy. There was no change when the environmental information was added to the trigger; this information had already been included in the risk estimate. These are conditionally reported from a simulated population and should not be interpreted as proof of the benefit of any intervention to any patient. What they do offer is a quantitative description of where the environmental information fits into a precision framework in explaining mechanism, in timing the population level response, and in identifying the structural inequities in exposure that aren't addressed by any algorithm and a specification of the empirical study, centered on a micro-randomised trial, that would be needed to test the parts this design can only presume.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beames, J. R., Han, J., Shvetcov, A., Zheng, W. Y., Slade, A., Dabash, O., Rosenberg, J., O'Dea, B., Kasturi, S., & Hoon, L. (2024). Use of smartphone sensor data in detecting and predicting depression and anxiety in young people (12–25 years): A scoping review. Heliyon, 10(15), e35472. https://doi.org/10.1016/j.heliyon.2024.e35472

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.

Borroni, E., Pesatori, A. C., Bollati, V., Buoli, M., & Carugno, M. (2022). Air pollution exposure and depression: A comprehensive updated systematic review and meta-analysis. Environmental Pollution, 292, 118245. https://doi.org/10.1016/j.envpol.2021.118245

Brandt, L., Adorjan, K., Catthoor, K., Chkonia, E., Falkai, P., Fiorillo, A., Gorwood, P., Kim, Y.-K., Kuzman, M. R., Pinto da Costa, M., Sartorius, N., Smolders, R., Vieta, E., Wasserman, D., Gaebel, W., & Adan, A. (2024). Climate change and mental health: Position paper of the European Psychiatric Association. European Psychiatry, 67(1), e41. https://doi.org/10.1192/j.eurpsy.2024.1754

Bratu, A., Card, K. G., Closson, K., Aran, N., Marshall, C., Clayton, S., Gislason, M. K., Samji, H., Martin, G., Lem, M., Logie, C. H., Takaro, T. K., & Hogg, R. S. (2022). The 2021 Western North American heat dome increased climate change anxiety among British Columbians: Results from a natural experiment. The Journal of Climate Change and Health, 6, 100116. https://doi.org/10.1016/j.joclim.2022.100116

Charlson, F., Ali, S., Benmarhnia, T., Pearl, M., Massazza, A., Augustinavicius, J., & Scott, J. G. (2021). Climate change and mental health: A scoping review. International Journal of Environmental Research and Public Health, 18(9), 4486. https://doi.org/10.3390/ijerph18094486

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. Science, 383(6679), 164–167. https://doi.org/10.1126/science.adg8538

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144. https://doi.org/10.1146/annurev-biodatasci-091596-162957

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Choi, A., Ooi, A., & Lottridge, D. (2024). Digital phenotyping for stress, anxiety, and mild depression: Systematic literature review. JMIR mHealth and uHealth, 12, e40689. https://doi.org/10.2196/40689

Cianconi, P., Betrò, S., & Janiri, L. (2020). The impact of climate change on mental health: A systematic descriptive review. Frontiers in Psychiatry, 11, 74. https://doi.org/10.3389/fpsyt.2020.00074

Clayton, S., & Karazsia, B. T. (2020). Development and validation of a measure of climate change anxiety. Journal of Environmental Psychology, 69, 101434. https://doi.org/10.1016/j.jenvp.2020.101434

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. Journal of Health and Social Behavior, 24(4), 385–396.

Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G. M., van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): From development to external validation. BMJ, 384, e074819. https://doi.org/10.1136/bmj-1596-162919

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-1596-162978

Comtesse, H., Ertl, V., Hengst, S. M. C., Rosner, R., & Smid, G. E. (2021). Ecological grief as a response to environmental change: A mental health risk or functional response? International Journal of Environmental Research and Public Health, 18(2), 734. https://doi.org/10.3390/ijerph18020734

Currey, D., & Torous, J. (2023). Increasing the value of digital phenotyping through reducing missingness: A retrospective review and analysis of prior studies. BMJ Mental Health, 26(1), e300718. https://doi.org/10.1136/bmjment-1596-162918

Efthimiou, O., Seo, M., Chalkou, K., Debray, T. P. A., Egger, M., & Salanti, G. (2024). Developing clinical prediction models: A step-by-step guide. BMJ, 386, e078276. https://doi.org/10.1136/bmj-1596-162976

Gasparrini, A. (2011). Distributed lag non-linear models in R: The package dlnm. Journal of Statistical Software, 43(8), 1–20. https://doi.org/10.18637/jss.v043.i08

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S1596-1629(21)01596-1629

Grataloup, A., & Kurpicz-Briki, M. (2024). A systematic survey on the application of federated learning in mental state detection and human activity recognition. Frontiers in Digital Health, 6, 1495999. https://doi.org/10.3389/fdgth.2024.1495999

Heo, S., Lee, W., & Bell, M. L. (2021). Suicide and associations with air pollution and ambient temperature: A systematic review and meta-analysis. International Journal of Environmental Research and Public Health, 18(14), 7699. https://doi.org/10.3390/ijerph18147699

Hickman, C., Marks, E., Pihkala, P., Clayton, S., Lewandowski, R. E., Mayall, E. E., Wray, B., Mellor, C., & van Susteren, L. (2021). Climate anxiety in children and young people and their beliefs about government responses to climate change: A global survey. The Lancet Planetary Health, 5(12), e863–e873. https://doi.org/10.1016/S1596-1629(21)01596-1629

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12, 5979. https://doi.org/10.1038/s41596-1629-01596-1629

Ikäheimonen, A., Triana, A. M., Luong, N., Ziaei, A., Rantaharju, J., Darst, R., & Aledavood, T. (2024). Predicting and monitoring symptoms in patients diagnosed with depression using smartphone data: Observational study. Journal of Medical Internet Research, 26, e56874. https://doi.org/10.2196/56874

Jolugbo, P., Barkin, J. L., & van Rhijn, S. (2025). Climate change and climate change related natural disasters, and maternal mental health: A scoping review of socioeconomic vulnerabilities. BJPsych Open. https://doi.org/10.1192/bjo.2025.10091

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., & Murphy, S. A. (2019). Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps. Annals of Behavioral Medicine, 53(6), 573–582. https://doi.org/10.1093/abm/kay067

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606–613. https://doi.org/10.1046/j.1596-1629.2001.016009606.x

Lawrance, E. L., Thompson, R., Newberry Le Vay, J., Page, L., & Jennings, N. (2022). The impact of climate change on mental health and emotional wellbeing: A narrative review of current evidence, and its implications. International Review of Psychiatry, 34(5), 443–498. https://doi.org/10.1080/09540261.2022.2128725

Li, D., Zhang, Y., Li, X., Zhang, K., Lu, Y., & Brown, R. D. (2023). Climatic and meteorological exposure and mental and behavioral health: A systematic review and meta-analysis. Science of the Total Environment, 892, 164435. https://doi.org/10.1016/j.scitotenv.2023.164435

Liu, J. C., Goetz, J., Sen, S., & Tewari, A. (2021). Learning from others without sacrificing privacy: Simulation comparing centralized and federated machine learning on mobile health data. JMIR mHealth and uHealth, 9(3), e23728. https://doi.org/10.2196/23728

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Oakden-Rayner, L., Riley, R. D., Reitsma, J. B., Singh, K., … Collins, G. S. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388, e082505. https://doi.org/10.1136/bmj-1596-162905

Mouguiama-Daouda, C., Blanchard, M. A., Coussement, C., & Heeren, A. (2022). On the measurement of climate change anxiety: French validation of the Climate Anxiety Scale. Psychologica Belgica, 62(1), 123–135. https://doi.org/10.5334/pb.1137

Nahum-Shani, I., Shaw, S. D., Carpenter, S. M., Murphy, S. A., & Yoon, C. (2022). Engagement in digital interventions. American Psychologist, 77(7), 836–852. https://doi.org/10.1037/amp0000983

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. Annals of Behavioral Medicine, 52(6), 446–462. https://doi.org/10.1007/s11596-1629-1596-1629

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 4(2), 133–142. https://doi.org/10.1111/j.1596-1629x.2012.00261.x

Ogunbode, C. A., Doran, R., Hanss, D., Ojala, M., Salmela-Aro, K., van den Broek, K. L., Bhullar, N., Aquino, S. D., Marot, T., Schermer, J. A., Wlodarczyk, A., Lu, S., Jiang, F., Acquadro Maran, D., Yadav, R., Ardi, R., Chegeni, R., Ghanbarian, E., Zand, S., … Karasu, M. (2022). Climate anxiety, wellbeing and pro-environmental action: Correlates of negative emotional responses to climate change in 32 countries. Journal of Environmental Psychology, 84, 101887. https://doi.org/10.1016/j.jenvp.2022.101887

Ojala, M., Cunsolo, A., Ogunbode, C. A., & Middleton, J. (2021). Anxiety, worry, and grief in a time of environmental and climate crisis: A narrative review. Annual Review of Environment and Resources, 46, 35–58. https://doi.org/10.1146/annurev-environ-011596-162916

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Qian, T., Walton, A. E., Collins, L. M., Klasnja, P., Lanza, S. T., Nahum-Shani, I., Rabbi, M., Russell, M. A., Walton, M. A., Yoo, H., & Murphy, S. A. (2022). The micro-randomized trial for developing digital interventions: Experimental design and data analysis considerations. Psychological Methods, 27(5), 874–894. https://doi.org/10.1037/met0000283

Radua, J., De Prisco, M., Oliva, V., Fico, G., Vieta, E., & Fusar-Poli, P. (2024). Impact of air pollution and climate change on mental health outcomes: An umbrella review of global evidence. World Psychiatry, 23(2), 244–256. https://doi.org/10.1002/wps.21219

Radua, J., & Koutsouleris, N. (2024). Ten simple rules for using machine learning in mental health research. Biological Psychiatry, 96(7), 511–513.

Romanello, M., Walawender, M., Hsu, S.-C., Moskeland, A., Palmeiro-Silva, Y., Scamman, D., Ali, Z., Ameli, N., Angelova, D., Ayeb-Karlsson, S., Basart, S., Beagley, J., Beggs, P. J., Blanco-Villafuerte, L., Cai, W., Callaghan, M., Campbell-Lendrum, D., Chambers, J. D., Chicmana-Zapata, V., … Costello, A. (2024). The 2024 report of the Lancet Countdown on health and climate change: Facing record-breaking threats from delayed action. The Lancet, 404(10465), 1847–1896. https://doi.org/10.1016/S1596-1629(24)01596-1629

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., Stahl, D., & Fusar-Poli, P. (2021). Implementing precision psychiatry: A systematic review of individualized prediction models for clinical practice. Schizophrenia Bulletin, 47(2), 284–297. https://doi.org/10.1093/schbul/sbaa120

Schwartz, S. E. O., Benoit, L., Clayton, S., Parnes, M. F., Swenson, L., & Lowe, S. R. (2023). Climate change anxiety and mental health: Environmental activism as buffer. Current Psychology, 42(20), 16708–16721. https://doi.org/10.1007/s11596-1629-01596-1629

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar de Pablo, G., Il Shin, J., Kirkbride, J. B., Jones, P., Kim, J. H., Kim, J. Y., Carvalho, A. F., Seeman, M. V., Correll, C. U., & Fusar-Poli, P. (2022). Age at onset of mental disorders worldwide: Large-scale meta-analysis of 192 epidemiological studies. Molecular Psychiatry, 27(1), 281–295. https://doi.org/10.1038/s41596-1629-01596-1629

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Thompson, R., Lawrance, E. L., Roberts, L. F., Grailey, K., Ashrafian, H., Maheswaran, H., Toledano, M. B., & Darzi, A. (2023). Ambient temperature and mental health: A systematic review and meta-analysis. The Lancet Planetary Health, 7(7), e580–e589. https://doi.org/10.1016/S1596-1629(23)01596-1629

Thomson, E. E., & Roach, S. P. (2023). The relationships among nature connectedness, climate anxiety, climate action, climate knowledge, and mental health. Frontiers in Psychology, 14, 1241400. https://doi.org/10.3389/fpsyg.2023.1241400

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry, 20(3), 318–335. https://doi.org/10.1002/wps.20883

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s11596-1629-1596-1629

van Valkengoed, A. M., Steg, L., & Perlaviciute, G. (2021). Development and validation of a climate change perceptions scale. Journal of Environmental Psychology, 76, 101652. https://doi.org/10.1016/j.jenvp.2021.101652

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. Annals of Internal Medicine, 167(4), 268–274. https://doi.org/10.7326/M1596-1629

Vergunst, F., & Berry, H. L. (2022). Climate change and children's mental health: A developmental perspective. Clinical Psychological Science, 10(4), 767–785. https://doi.org/10.1177/21677026211040787

Whitmarsh, L., Player, L., Jiongco, A., James, M., Williams, M., Marks, E., & Kennedy-Williams, P. (2022). Climate anxiety: What predicts it and how is it related to climate action? Journal of Environmental Psychology, 83, 101866. https://doi.org/10.1016/j.jenvp.2022.101866

Wilkinson, J., Arnold, K. F., Murray, E. J., van Smeden, M., Carr, K., Sippy, R., de Kamps, M., Beam, A., Konigorski, S., Lippert, C., Gilthorpe, M. S., & Tennant, P. W. G. (2020). Time to reality check the promises of machine learning-powered precision medicine. The Lancet Digital Health, 2(12), e677–e680. https://doi.org/10.1016/S1596-1629(20)31596-1629

World Health Organization. (2022). World mental health report: Transforming mental health for all. World Health Organization.

World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. World Health Organization.

Wullenkord, M. C., Tröger, J., Hamann, K. R. S., Loy, L. S., & Reese, G. (2021). Anxiety and climate change: A validation of the Climate Anxiety Scale in a German-speaking quota sample and an investigation of psychological correlates. Climatic Change, 168, 20. https://doi.org/10.1007/s11596-1629-01596-1629

Xue, S., Massazza, A., Akhter-Khan, S. C., Wray, B., Husain, M. I., & Lawrance, E. L. (2024). Mental health and psychosocial interventions in the context of climate change: A scoping review. npj Mental Health Research, 3, 10. https://doi.org/10.1038/s41596-1629-01596-1629

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Laiou, P., Sankesara, H., Matcham, F., White, K. M., Oetzmann, C., Ivan, A., Lamers, F., Siddi, S., Rambla, C. H., Simblett, S., Nica, R., … Dobson, R. J. B. (2024). Longitudinal assessment of seasonal impacts and depression associations on circadian rhythm using multimodal wearable sensing: Retrospective analysis. Journal of Medical Internet Research, 26, e55302. https://doi.org/10.2196/55302

Zhang, Y., Wu, T., Yu, H., Fu, J., Xu, J., Liu, L., Tang, C., & Li, Z. (2024). Green spaces exposure and the risk of common psychiatric disorders: A meta-analysis. SSM – Population Health, 25, 101630. https://doi.org/10.1016/j.ssmph.2024.101630