DNA methylation-based predictors of health: applications and statistical considerations

Paul D. Yousefi, Matthew Suderman, Ryan Langdon, Oliver Whitehurst, George Davey Smith and Caroline L. Relton✉

Abstract | DNA methylation data have become a valuable source of information for biomarker development, because, unlike static genetic risk estimates, DNA methylation varies dynamically in relation to diverse exogenous and endogenous factors, including environmental risk factors and complex disease pathology. Reliable methods for genome-wide measurement at scale have led to the proliferation of epigenome-wide association studies and subsequently to the development of DNA methylation-based predictors across a wide range of health-related applications, from the identification of risk factors or exposures, such as age and smoking, to early detection of disease or progression in cancer, cardiovascular and neurological disease. This Review evaluates the progress of existing DNA methylation-based predictors, including the contribution of machine learning techniques, and assesses the uptake of key statistical best practices needed to ensure their reliable performance, such as data-driven feature selection, elimination of data leakage in performance estimates and use of generalizable, adequately powered training samples.

Genome-wide association studies
(GWAS). Studies that examine the statistical correlation or 'association' between a set of genetic polymorphisms large enough to capture most of the variation in the human genome and a given phenotype of interest.

Medical Research Council Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Bristol, UK.

✉*e-mail: caroline.relton@bristol.ac.uk*

https://doi.org/10.1038/s41576-022-00465-w

Common complex diseases are major contributors to the global burden of disease. Anticipating and stratifying patients according to their disease risk or likelihood of response to treatment is therefore a major translational goal motivating research across clinical medicine, public health and epidemiology, an approach sometimes referred to as precision or personalized medicine[1]. Information on both heritable and environmental factors related to disease can be harnessed to discriminate such risks, including family history, medical history and lifestyle. For example, genome-wide association studies (GWAS) have sought to predict heritable components of disease risks and phenotypes by estimating weighted combinations of risk alleles, called polygenic risk scores (PRSs). Although their explanatory power varies, large consortium-driven efforts have generated PRSs that are strongly predictive of complex phenotypes, including for height, body composition and cardiovascular disease[2–4]. However, PRSs have inherent limits to their ability to predict disease incidence[5,6]: because genotype is established at fertilization, they are only able to estimate lifetime disease liability, and their explanatory capacity for any phenotype reaches a theoretical limit at the traits' broad-sense heritability[7,8].

By contrast, DNA methylation (DNAm) varies dynamically throughout the life course owing to a complex set of endogenous biological processes, such as disease pathology[9–11], density of CpG sites[12], time- and lineage-varying cell specificity[13–16], as well as in response to exogenous environmental exposures[17,18]. The presence or absence of methyl groups at the more than 30 million CpG sites across the human genome establishes cell-lineage-specific patterns of gene expression involved in cellular differentiation throughout development, which are often maintained into adulthood[13–15]. Measurement of the relative proportion of methyl groups present at a particular CpG site across the many DNA strands isolated from a population of cells is possible using genotyping technology following bisulfite treatment of DNA, which converts the presence or absence of methylation into a genetic polymorphism. This approach has allowed DNAm levels to be measured at many CpG sites simultaneously using microarrays and sequencing approaches in a manner that is sufficiently cost-effective for use at scale in human population studies[19,20]. These advances have contributed to an era of studies that perform genome-level, locus-agnostic examinations of levels of DNAm, commonly called epigenome-wide association studies (EWAS) or methylome-wide association studies (MWAS). The population-based design, large multiple-testing burden and statistical framework for EWAS have parallels with GWAS, albeit with distinct biological interpretation.

Results of EWAS are likely to capture tissue- and time-specific information distinct from that of static genetic estimates and have the potential to refine or improve

Box 1 | **Explanatory versus predictive modelling**

A fundamental and underappreciated distinction exists between modelling aimed at explanations and modelling aimed at predictions. Explanatory models are used to test hypothesized causal relationships or theoretical descriptions of observed phenomena, whereas predictive models aim to estimate future or yet-to-be-sampled observations[204]. Explanatory models are more familiar to scientific practitioners as they are used widely across disciplines. However, the disparity between the two approaches can easily be appreciated by considering the shift in emphasis that predictive models take compared with explanatory models in four key aspects of their methodology.

Association over causation
As predictive models attempt to produce only the best possible guess of an outcome in new observations instead of evaluating the soundness of causal–theoretical assertions, they are concerned only with identifying associations between observed inputs and outputs.

Data over theory
The point of explanatory models is to evaluate possible competing testable theories of the causal relationships between relevant inputs and outcomes. Predictive modelling takes an opposite approach and only builds the most convenient theoretical model from the observed data, as causal interpretability is not necessarily required.

Variance over bias
The expected prediction error models generated can be deconvolved into three categories: first, the variance inherent to the outcome; second, bias or errors in accuracy of the model specification; and third, estimation variance or differences in precision from estimating the model in a finite sample[176]. The first is unavoidable, but the second and third are functions of the modelling method implemented. Whereas explanatory models prefer reductions in model bias, suggesting that their theoretical assertions more closely adhere to the causal reality, predictive models, which value 'getting it right' above all else, find it more acceptable for error reduction to stem from improved precision.

Prospective over retrospective
Predictive models aim to estimate future values or unobserved instances of an outcome whereas explanatory models are focused on deciphering causal explanations of past observations. For predictive models, more importantly than whether inputs causally precede outcomes is whether the observations of input variables will be available at the time the prediction is performed, called ex ante availability. A subset of prediction problems, which we refer to as forecasting, target unobserved outcomes that explicitly occur in the chronological future and require time-lagged or time-series data (for example, given tumour stage at time *t*, what is predicted survival at $t+5$ years?). However, prediction models can be applied in cross-sectional settings when there are data collection or information gaps such that some outcomes are yet to be observed (for example, if we have not asked someone if they smoke, can we predict whether they do so from the DNAm levels we observe ex ante at their *AHRR* gene?).

These differences in priorities reflect the disparate aims of explanatory and predictive modelling but also have ramifications for a host of decisions made throughout the modelling process. These range from the types of variables and models deemed appropriate for use owing to expectations of interpretability to crucial decisions on allocation of resources that arise from varying demands on study design, data collection and replication or validation of results.

PRS phenotype predictions beyond the limit of phenotype heritability. Sample sizes from even the largest EWAS to date have remained well below that of contemporary GWAS and have largely been performed in DNA isolated from blood. However, spurred by technical advances, such as multiplexed detection of hundreds of thousands of CpG sites by microarray-based technology using bisulfite-converted DNA[21,22], DNAm-based predictors of health-related phenotypes have begun to proliferate. Already, these DNAm predictors show additive capacity to explain some phenotypes compared with genetic predictors[23–26].

Much hype has emerged regarding how machine learning approaches can improve prediction by automating the process of building or 'training' predictive models using one data set and then evaluating or 'testing' their performance in another data set. Such automation has increasingly become necessary to ensure that models make full use of the predictive information available in data sets that are too large for manual inspection, such as population-scale DNAm data. This Review identifies the key work to date, areas of opportunity, limitations, trade-offs and fundamental design constraints of DNAm predictors. We survey the applications and phenotypes that have so far received the greatest attention as targets for the development of DNAm predictors. In the first two sections, we summarize the progress that has been made in producing DNAm predictors of health risk factors and outcomes, respectively. In the next section, we examine the spectrum of statistical approaches that have been deployed to optimize prediction performance. Lastly, we examine where application gaps are likely to exist and provide a description of best practices in DNAm predictor development.

Predictors of risk factors and exposures

To date, analysis of DNAm variation has most commonly aimed to determine whether differences in DNAm patterns between comparison groups have some causal or mechanistic relevance, as assessed by explanatory modelling methods (BOX 1). However, variation in DNAm levels with unknown mechanistic function may have relevance for health if it indexes other endogenous risk factors that contribute to disease risk or progression or a response to environmental exposures, particularly if these are rare or expensive to measure via biomonitoring. A few risk factors and exposures emerged early on as candidates for development of informative predictors because they have large numbers of associated CpG sites and substantial effect sizes, notably chronological age and tobacco smoking behaviour. Since then, the development of DNAm predictors has expanded to target a growing number of health risk factors and exposures (FIG. 1).

Age. As far back as the late 1960s, there was extensive reporting of strong associations between age and DNAm in various organisms and tissues[27]. The early 2010s saw the development of DNAm or epigenetic clocks, mathematical algorithms that use machine learning to derive an estimate of the epigenetic age of a DNA source[27–29]. For example, Horvath used penalized regression to develop a clock that uses the output of a linear model to integrate DNAm levels measured at 353 CpG sites across the genome; this widely used clock produces estimates of age that correlate strongly with chronological age across a wide variety of tissues and cell types[27]. Well over a dozen further clocks have since been developed[30]. The sudden explosion of interest was motivated in part by the increasing availability of large-scale DNAm data sets; whereas an early data set that was based on 93 DNAm profiles was used to uncover hundreds of age associations across the genome[31], Horvath was able to assemble 8,000 DNAm profiles, all of them publicly available, to derive his clock[20]. A second innovation was Horvath's hypothesis that the epigenetic clock may provide an

Polygenic risk scores (PRSs). Weighted sums of risks for a phenotype conferred by genetic polymorphisms within an individual where the weights used are coefficients from the relevant genome-wide association studies (GWAS). GWAS loci are typically selected for inclusion in the score by applying a *P* value threshold, commonly that of genome-wide significance ($P<5\times10^{-8}$).

Broad-sense heritability
The proportion of phenotype or trait variance attributable to genetic factors.

DNA methylation
(DNAm). An epigenetic modification whereby a methyl group (CH_3) is covalently attached to a DNA base in a mitotically stable bond. In mammals, DNAm occurs mainly at cytosine residues in CpG sites.

CpG sites
Specific sequences of DNA bases where cytosines are followed by guanines. The 'p' indicates the phosphate bond separating the two residues in sequence in the 5′ to 3′ direction.

Epigenome-wide association studies
(EWAS). Studies that examine the association between a large number of epigenetic variables and a phenotype or exposure of interest. As most have been performed using DNA methylation levels, we treat EWAS and methylome-wide association studies as synonyms.

DNAm-based predictors
Any statistical models (for example, linear model) of observed data employed to predict values of an outcome (for example, exposure, phenotype or disease) in which many or all of the of the input variables are levels of DNA methylation (DNAm) measured at CpG sites.

Machine learning
Algorithms and statistical models that improve their performance from experience or by optimization through training on earlier data collection.

Epigenetic clocks
Estimators of biological age or other ageing phenotypes that use levels of DNA methylation or other epigenetic measurements as inputs.

Penalized regression
Linear regression modelling methods that apply some numerical penalty on the total size of all input variable coefficient values. Examples include lasso, ridge and elastic net regression.

estimate of biological age, and that discrepancies between clock age and chronological age would thus identify instances of accelerated or decelerated ageing. This hypothesis has since been supported by reports of hundreds of associations between DNAm age acceleration and various exposures (for example, traumatic stress[32], lifetime stress[33]), phenotypes (for example, obesity[34], fitness[35]) or health outcomes (for example, lung cancer[36], mortality[37,38]).

Many associations between DNAm and age could subsequently be explained by age-related changes in the proportions of white blood cell types[39]. Variation in the relative amount of various cell types occurs both within and between individuals, and DNAm patterns differ strongly between cell types, as determining cell fate is a major regulatory function of DNAm. Nevertheless, other associations and clock performances have been derived and replicated in multiple tissues and cell types[27]. The mostly tissue-independent behaviour of some clocks makes them attractive for clinical and epidemiological applications where they can be conveniently applied to peripheral tissues such as blood or saliva. Horvath further refined existing clocks to be either 'intrinsic', aimed at capturing ageing independently of cell type, or 'extrinsic', combining ageing information from both DNAm and blood cell type proportions[40]. Extrinsic clocks have increased health-relevant explanatory power over intrinsic clocks owing to their additional cell type information. However, there is growing evidence against the hypothesis that DNAm age acceleration is driven by ageing-related changes in DNAm. Most notably, one study showed that differences between intrinsic clock estimates and chronological age decreased as training data set size increased, suggesting that DNAm within single cell types tracks age, and cell composition reflects ageing[41]. Thus, it is possible that sufficiently large training data sets could produce clocks without any biologically relevant departure from chronological age. Indeed, the authors show that associations with mortality diminish as training set size increases[41] (BOX 2).

Inspired by the success of extrinsic clocks, a second generation of clocks has been developed that are trained to estimate ageing-associated phenotypes and diseases instead of, or in addition to, age itself. Prominent examples include PhenoAge, a DNAm predictor of nine clinical biomarkers of health[42], GrimAge, a model of seven plasma proteins, smoking and age[43], and DunedinPoAm, a DNAm predictor of 18 biomarkers of 'organ-system integrity'[44]. These more recent clocks strongly outperform previous clocks and even their underlying biomarkers for most ageing outcomes. For example, the statistical inference for the association between mortality and PhenoAge is greater than 20 orders of magnitude stronger than with the best first-generation clocks[42]. This strong improvement in performance cannot be explained by larger training set sizes, as the training set sizes for PhenoAge, DunedinPoAm and GrimAge were much smaller (n = 900, 1,000, 1,700, respectively) than for Horvath's original clock (n = 8,000).

Although superfluous to their function as purely predictive tools (BOX 3), the widespread application of age-related clocks has prompted questions about the mechanisms that underlie their mode of action. GWAS of both first- and second-generation clocks have

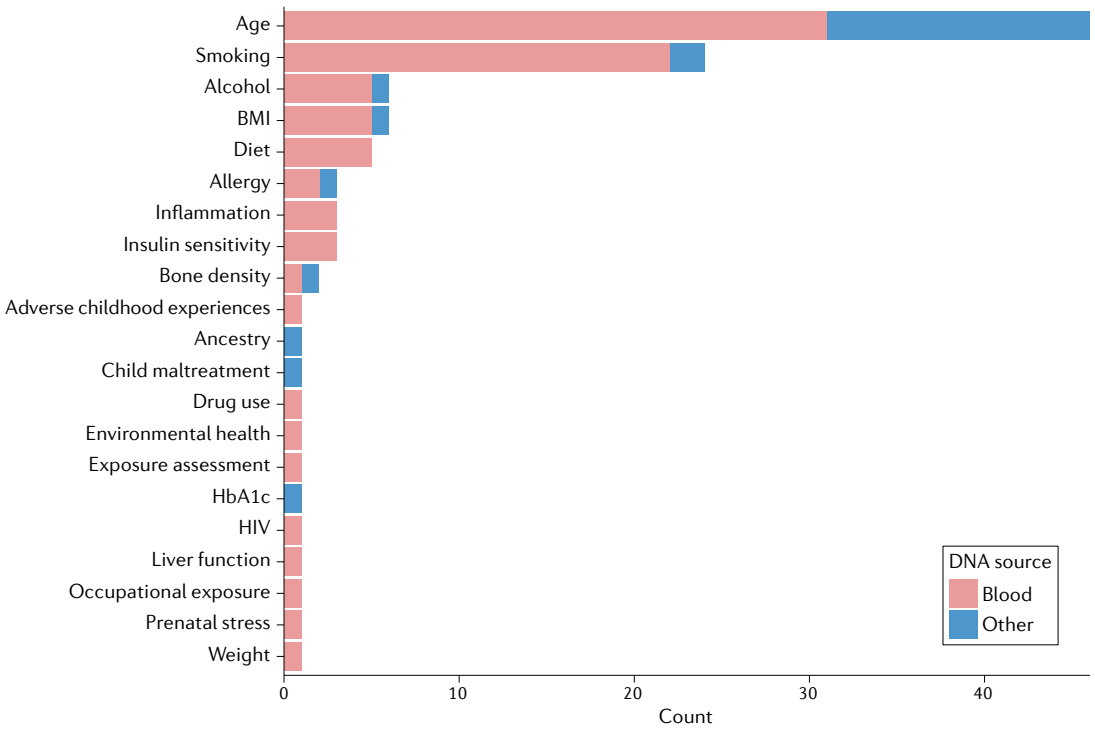


Fig. 1 | **Illustrative distribution of application areas for studies of DNA methylation (DNAm) prediction of health risk factors and exposures by DNA source tissue for the majority of relevant studies published before April 2020.** BMI, body mass index; HbA1c, haemoglobin A1c.

Box 2 | **Ground truth in predictor development**

Ground truth is the set of true values derived by either observation or measurement. Predictors are trained to produce outputs that match the ground truth for the variable being predicted. For example, for a predictor of weight, the measured weights are the ground truth. In many cases, including DNA methylation (DNAm) applications, ground truth is not available; nearly every measurement is subject to some level of measurement error. Hence, the measured values can only be considered approximations of ground truth. Exposures such as cigarette smoke, alcohol or nutrient intake can be subject to high levels of measurement error[205–207]. In more extreme cases, it may be impossible to actually measure ground truth: examples include biological age, health, intelligence, obesity and well-being. In these cases, alternatives are used in place of ground truth. For example, body mass index is measured as a proxy for obesity, even though there is substantial controversy over limitations in its ability to reflect obesity[208]. Instead of biological age, first-generation epigenetic clocks were trained to model chronological age. Prediction errors were assumed to indicate differences between chronological and biological age. Second-generation clocks were trained to model combinations of ageing and health biomarkers. Although these biomarkers were closer to biological age, they are still not equivalent to biological age itself. In fact, there is no consensus definition of biological age. However, this has not prevented epigenetic clocks from being indicators of past health-related exposures and predictors of future health outcomes.

identified genetic associations but have so far failed to identify clear underlying mechanisms[45–47]. All genetic studies have indicated a role for the gene *TERT*, which encodes a regulator of telomere length; yet, evidence suggests that DNAm clocks and telomere length capture distinct aspects of ageing[48]. At least some of the genetic variation associated with most clocks is linked to cell count variation, supporting the hypothesized existence of some confounding, even for 'intrinsic' clocks, by cell composition[47]. Unsurprisingly, owing to the complex combination of biomarkers that underlie the new generation of clocks, the mechanisms influencing them have not yet been pinpointed and seem to be equally complex.

Smoking. Smoking is the exogenous exposure that has the greatest known impact on DNAm levels, with EWAS consistently reporting large numbers of robust associations for various smoking-related phenotypes, including smoking status[49], number of cigarettes per day[50] and time since smoking cessation[51], leading to the successful derivation of DNAm predictors of each of these phenotypes[26,52–54]. As with age, there is evidence that at least some DNAm–smoking associations are cell-type or lineage specific[55,56], but the specific implications for DNAm predictor performance have yet to be fully characterized. However, given the large variance in the life spans of different white blood cell types, cell-specific effects may contribute to the wide range of half-lives observed for smoking-altered DNAm levels at different CpG sites. For some CpG sites, smoking-related DNAm signals can persist for more than 30 years[49], capturing information on historical exposure well beyond the half-lives of traditional smoking biomarkers, such as serum or urinary cotinine[57]. For some CpG sites, smoking-associated DNAm reverts to levels observed in never-smokers soon after smoking cessation, for some individuals in as little as 1 year[51,53]. For others, associations with in utero smoking exposure are known to persist into middle age[58,59]. The precise mechanisms explaining the perpetuation of the DNAm signal in blood cells in the absence of continued exposure remain to be defined.

An added value of DNAm predictors of smoking is their ability to explain more variance in health-related outcomes than self-reported smoking, in part owing to their ability to capture lifelong exposure with less reporter bias[60]. For example, in the Lothian Birth Cohort 1936, a DNAm predictor of smoking explained greater proportions of variance in cognitive function, structural brain integrity, inflammatory markers and other smoking-related health measures than phenotypic smoking[54]. These findings suggest that improved precision and dose-dependent DNAm characterization of smoking may further delineate the relationship between smoking and disease.

Alcohol consumption. Alcohol consumption information is notoriously biased, and objective assays are informative only in the short term (for example, breathalysers) or for detecting extreme exposure (for example, liver enzymes[61]). DNAm in peripheral tissues such as blood has been explored as a potential biomarker given the success of DNAm predictors of exposure to cigarette smoke[49,62]. However, although EWAS have identified hundreds of associations with alcohol intake, the observed magnitude of effects has been small and resulted in DNAm predictors with moderate ability to explain levels of alcohol intake[63], albeit with substantially improved performance over the <1% of variance explained that is commonly seen for PRS[64,65]. Whereas DNAm at a single CpG site in *AHRR* (cg05575921) was found to explain 66% of the variance in cigarette consumption[66], the best alcohol intake model based on 144 CpG sites explained only 0.8–14.3% of the variance, with higher values being observed in populations with long-term and high average alcohol consumption[67]. Interestingly, this alcohol intake model actually tended to explain slightly more variation in a measure of alcohol use disorder than intake itself in models that adjusted for white blood cell type proportions. Although undesirable for accurately estimating alcohol intake, it could prove useful for detecting and studying alcohol abuse.

Adiposity and lipid profiles. A growing body of literature has identified associations between peripheral blood DNAm and several aspects of metabolism associated with a wide range of adverse health outcomes. These include body mass index (BMI)[68–72], central adiposity[73,74], childhood adiposity[75], diet[76–79], lipid profiles[80] and a range of prenatal exposures such as maternal dysglycaemia, diet, BMI and lipid profiles[81–85]. Adiposity as measured by BMI and its association with DNAm is the most commonly investigated of these to date, likely because it is the easiest to measure in large populations. Similarly, because blood is more frequently available, few studies have been conducted in adipose tissue, even though DNAm levels in adipocytes have much stronger associations with BMI[86–88]. A further limitation of blood DNAm is that it seems to be of little use in predicting future BMI beyond current BMI[68,70,71,89,90]. Evidence from these studies based on Mendelian randomization and longitudinal modelling overwhelmingly support reverse

Linear model
A statistical description of the relationship between one or many input variables *X* and an observed level of an output *Y*, where each *X*–*Y* association is summarized by the slope or coefficient of the line plotted between them.

Biological age
The hypothesis that the phenotypical age of a DNA source (for example, cell, tissue or organ) may be greater (that is, accelerated) or less (that is, decelerated) than chronological age at any given point in time.

Mendelian randomization
An analytical method that uses genetic variants as instrumental variables to evaluate putative causal relationships between modifiable risk factors and disease outcomes.

causation, whereby changes in BMI drive changes in DNAm rather than vice versa.

Psychosocial factors and stressful environments. There is an extensive literature investigating the plasticity and stability of epigenetic processes in response to stress and the social environment[91,92]. EWAS have improved understanding of epigenetic changes associated with many pre- and postnatal psychosocial stressors, such as socioeconomic position, depression, violence and childhood abuse, among others[93–99]. A common feature of mechanistic studies investigating stress and adverse social environments is a prolonged activation of the hypothalamic–pituitary–adrenocortical (HPA) axis[100]. However, there are few examples in which DNAm robustly indexes complex, sometimes hard-to-measure psychosocial phenotypes. Some studies have shown that DNAm can detect dysregulation of the HPA axis, which subsequently can predict impaired mental health[101,102]. However, performance of the limited number of published DNAm predictors of specific psychosocial phenotypes is generally poor, with performances in independent data only slightly better than random (for example, area under the curve (AUC) = 0.58 for major depressive disorder[103], and AUC = 0.5 for suicide attempt[104]; BOX 4).

Box 3 | **Does causality improve prediction?**

Although predictive modelling is ultimately only concerned with reliably guessing an unobserved outcome, *Y*, given a set of inputs, X, regardless of cause, that does not mean that causal factors are not involved. That is, a network of causal relationships still determines the value of *Y* whether or not predictive models choose to incorporate that information when making their estimates. The causal inference literature often uses causal graphs, called directed acyclic graphs (DAGs)[209], to formalize these networks, and although it may be practically impossible, any set X can in theory be graphically expressed in relation to the causal process that produces *Y*. If no causal link exists between X and *Y*, then the graph would reflect their independence.

In fact, for any input *X* in X to provide signal regarding *Y* beyond pure noise, it must by definition have some causal connection to *Y*. This may seem counterintuitive at first but becomes clearer in the context of particular examples. For example, consider how carrying a lighter in one's pocket is a reliable predictor of later getting lung cancer. The causal link that underpins this prediction is not apparent until one recognizes that it is driven by an unobserved common cause of both *X* and *Y*: cigarette smoking. Similarly, consider the role of a genetic polymorphism, SNP A, that contributes to the genetic risk prediction for schizophrenia through its inclusion in a polygenic risk score. Because of the correlation structure of the genome, SNP A may be predictive for schizophrenia risk without functionally contributing to the development of the disease. Instead, it need only be correlated with another SNP that does have a causal role in schizophrenia. Here again, the non-causal association is due to another relationship, which is causal, that is, the assortment of SNP A into close proximity with a true genetic cause of the outcome explains the predictive capacity of a particular input variable.

Several different types of non-causal relationship can connect *X* to *Y*, owing to confounding, colliders and reverse causation, by interacting directly with *Y* or its causal antecedents (known as 'parents' in causal graphs) and consequences ('descendants'). However, recent work that combined principles emerging from the field of 'causal learning'[210] and analysis of Bayesian networks[211] suggests that, assuming no measurement error, the optimal set of variables for predicting an outcome belong to the 'Markov Blanket' or the variables that are parents of the outcome variable, all of its descendants and all parents of those descendants[212]. Although the direction of causal relationship between *X* and *Y* may have implications for predictive generalizability — reverse causes will tend to be less generalizable than forward causes — direction alone does not determine predictive performance. In fact, reverse causation has the potential to produce very strong predictors. In many instances in which causal information may be prohibitively expensive and/or inexpedient to discover, causally agnostic predictors may yield valuable information.

Predictors of health outcomes

In addition to providing an index of exposure history, DNAm predictors can be developed to aid anticipation and surveillance of disease. In that regard, DNAm predictors can convey information on a range of environmental risk factors as well as detect early and progression-related disease impact. To date, much of this work has been performed in cancer settings, with substantial focus on identifying early disease indicators; because they rely on disease initiation for their signal, these applications have an inherent ceiling to their prediction or forecasting performance. Predictors that pool information across disease and environmental risk are only beginning to be considered. Beyond applications that attempt to provide early detection of existing disease, much space still remains for addressing more nuanced clinical questions regarding disease progression trajectory following identification, subtyping and response to treatment.

Oncology. Altered patterns of DNAm are a hallmark of both oncogenesis and the pathophysiology of cancer progression[105–107]. Owing to the well-known functional role of DNAm in tumorigenesis, there has been considerable interest in translating functional discoveries into clinical biomarker applications. Perhaps one of the best-characterized successes of clinical translation is the relationship between O^6-methylguanine-DNA methyltransferase (*MGMT*) methylation and response to chemotherapy in patients with glioblastoma[108]. After a body of evidence emerged showing that tumour *MGMT* methylation has a clear functional role in the mechanism of action of temozolomide, tumour *MGMT* promoter methylation status was validated in phase III trials as a biomarker for favourable outcome in patients with glioblastoma treated with temozolomide chemotherapy[109,110].

Most oncology applications for DNAm predictors have been developed for neurological and breast cancers, followed in turn by prostate and head and neck cancer (FIG. 2). The development of DNAm predictors for cancer to date has been driven by several factors, including overall amount of research interest, poor diagnostics and over-screening issues (for example, prostate[111], breast[112]), clinical need for early detection (for example, lung[113,114], ovarian[115] and pancreatic[116]) and the existence of a functional, biologically relevant signal (for example, *MGMT* in glioblastoma detection[108]).

Most DNAm predictors have been derived from DNAm profiles of tumour samples to predict progression and survival (FIG. 2). As such, applications are limited to where biopsy tissue is available, preventing use in other critical scenarios such as early detection of cancer. For early detection, typically only non-invasive, peripheral tissues are likely to be available, in which DNAm has less direct mechanistic relevance to tumour biology but has still proved informative in some applications[117–120] (BOX 5). Predictor development in such settings has tended to incorporate information from a greater number of CpG sites and be trained in large general populations rather than targeted patient samples[121–123].

Cell-free DNA
(cfDNA). Non-nucleated DNA found circulating in blood plasma. Sources can include lysed cells from any number of tissues, including tumour cells, which are commonly of greatest interest.

Studies in peripheral tissue that have seen the most progress so far have involved diagnosis of lung and pancreatic cancer. DNAm predictors of lung cancer derived from peripheral blood have consistently performed well (AUC >0.75 across multiple studies) using three or fewer CpGs and have shown improvements over clinical prediction models[113,124–126]. Some attempts have been made to develop blood-based DNAm predictors for pancreatic cancer diagnosis, with reasonable classification performance (AUC >0.75)[127,128]. However, all these applications still await external validation.

Cell-free DNA collected from peripheral sources such as blood plasma and urine has increasingly become a source for cancer diagnostic applications mostly focused on hepatocellular carcinoma[129]. These studies identify circulating DNA fragments originating from tumour

Box 4 | Measuring performance

To characterize the extent to which the predictions of a model adhere to observed outcomes, it is often helpful to generate a summary statistic, called a performance metric, that quantifies the overall predictor error. The types of statistics used to evaluate such predictive performance differ depending on whether the outcome being targeted is a continuous number (that is, a regression problem) or a discrete set of values (that is, a classification problem).

In regression settings, common metrics include the coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute deviation (MAD). The popular R^2 summarizes the linear concordance between predicted and observed outcomes and benefits from a standardized interpretation ranging from 0 (perfect disagreement) to 1 (perfect agreement). Conversely, RMSE and MAD both provide estimates of the average amount of predictor error in the same scale or units as the outcome being predicted.

For classification, nearly all metrics attempt to summarize the simple cross-tabulation of agreement and disagreement between observed and predicted outcomes, called a confusion matrix (see the figure, panel **a**). Accuracy is perhaps the best known of the metrics derived from this matrix, taking the counts for which observed and predicted outcomes agree (shown along the diagonal) as a percentage of all observations. However, sensitivity, specificity, precision, recall, and positive or negative predictive value are all reasonable alternative metrics that can be trivially computed from confusion matrices.

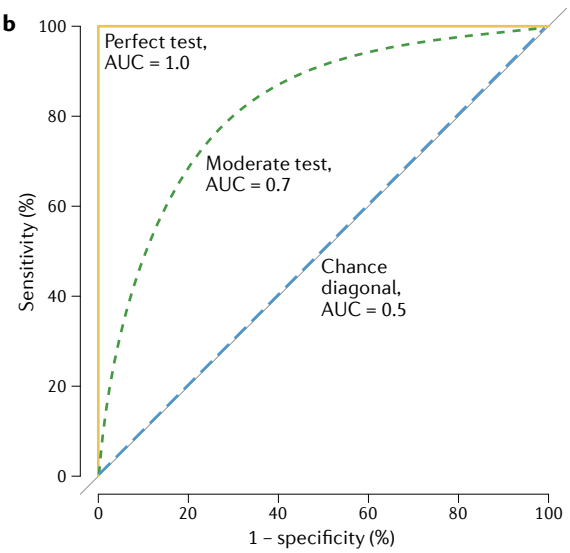
A common complication involves summarizing performance of models with continuous or risk probability outputs for discrete outcomes. In this case, a threshold value needs to be established to convert continuous predictions into discrete classes. Unfortunately, optimal thresholds are rarely replicable between studies, especially in studies that use microarray-generated DNA methylation (DNAm) profiles, where batch effects can be substantial. A common solution is to summarize model performance across all possible threshold values. For each possible threshold cut-off, sensitivity (true positive rate) and 1 – specificity (false-positive rate) are calculated from the corresponding confusion matrix. The resulting calculations can be visualized as a curve in a plot of sensitivity versus 1 – specificity, called a receiver operating characteristic (ROC) curve, where the area under the curve (AUC) or *c*-statistic provides a summary of overall performance (see the figure, panel **b**). As the threshold changes from one extreme of model outputs to the opposite extreme, sensitivity and 1 – specificity, which are initially 0%, gradually increase until both reach 100%. A predictor that generates purely random predictions will tend to produce an equal number of false positives and false negatives. The resulting ROC curve will lie along the diagonal and have an AUC roughly equal to 0.5. By contrast, a predictor that perfectly separates the outcomes will produce a curve with AUC = 1.0 that goes straight up to sensitivity 100% and 1 – specificity 0% and then horizontally to sensitivity and 1 – specificity 100%. Informative but imperfect predictors will produce curves somewhere between these two extremes, with AUC values ranging between 0.5 and 1.0. Better predictors will produce curves that diverge more from the diagonal and have greater AUC values. Similarly, a curve can also be generated using precision and recall as the *y* and *x* axes, respectively, again using the AUC across the range of threshold values as a summary for predictor performance across these two metrics. As precision and recall both assess performance focused on the target event or positive class frequency (which conventionally is the less abundant class), the AUC from this graph tends to be more robust to drastic class imbalances. Here, a non-informative curve will be a horizontal line with a precision in proportion to the target event frequency, while an informative curve will bow towards the 1,1 point and have an AUC approaching 1.0.

For calibration, or the agreement between observed and predicted probabilities, a scatter plot of these two quantities is typically the first step in assessing performance. Visual inspection of whether the linear trend or calibration curve from this plot adheres closely to a 45-degree line is a valuable assessment, and indeed the calibration curve slope has perhaps been the most common measure for quantifying calibration, where a slope of 1 is ideal. Several alternative metrics have been proposed that use slightly different heuristics for quantifying linear agreement[213].

a Confusion matrix

		Reference/true classes	
		Event	No event
Predicted classes	Event	A True positives	B False positives
	No event	C False negatives	D True negatives

Accuracy	(A+D)/(A+B+C+D)	% of predicted 'event' and 'no event' that are true
Sensitivity, recall or true positive rate	A/(A+C)	When it is a true 'event' how often it predicts 'event'
Specificity, selectivity or true negative rate	D/(B+D)	When it is a true 'no event' how often it predicts 'no event'
Precision or positive predictive value	A/(A+B)	When predicting 'event' how often it is correct
Negative predictive value	D/(D+C)	When predicting 'no event', how often is it correct



cells by detecting tumour-specific mutations or DNAm patterns that provide targeted tumour information. Early examples evaluated tumour-specific mutations only and reported low sensitivity for early detection owing to low levels of both circulating tumour DNA and recurrent tumour-specific mutations[130,131]. Since then, investigations have grown to consider tumour-specific DNAm patterns, which cover large sections of the genome and can sometimes provide information about the tissue of origin. Some early studies have reported promising results for the early diagnosis of various cancers[130,132] including intracranial tumours[133], renal cell carcinoma[134], pancreatic cancer[135], bladder cancer[136] and colorectal cancer[137]. One such assay platform[132], for the simultaneous early detection and tissue-of-origin determination of over 11 cancers, is being trialled for clinical use by the UK National Health Service[138]. However, sensitive detection performance in initial results was limited to late-stage tumours[139], and the overall clinical impact remains to be seen.

Cardiovascular disease. Evidence generated on blood-based DNAm profiles supports the existence of substantial DNAm signals linked to outcomes such as blood pressure[140,141] and cardiovascular disease[142,143]. So far, DNAm predictors have been developed for the diagnosis of large-artery atherosclerosis stroke[144], hypertension[140,141] and risk of cardiovascular disease[145,146], with some success (for example, AUC >0.70 for stroke observed for six separate CpGs[144]). Also, increasingly studies have been conducted to identify biomarkers for risk of related phenotypes, such as metabolic syndrome[147] and type 2 diabetes mellitus[148,149]. However, few of these studies have progressed beyond identifying differential methylation and attempted to formally construct and evaluate DNAm predictors. A recent exception developed a DNAm predictor of incident cardiovascular disease that was associated (hazard ratio = 1.58, $P < 6 \times 10^{-10}$) with time to myocardial infarction in an independent population[146].

Neurological and psychiatric disease. An advantage of brain biomarker studies, compared with imaging approaches, has been their capacity for functional molecular interpretation. However, molecular studies of brain, including DNAm, are often limited to being conducted in samples collected post mortem or in peripheral tissues such as blood or saliva. To date, studies using peripheral tissues have had some success discriminating neurological conditions such as Alzheimer disease and Parkinson disease, in which treatment may have more favourable results if initiated earlier in the disease process (for example, AUC >0.84 for Alzheimer disease cases versus controls)[150,151]. Other psychiatric conditions with blood-based predictors include postpartum depression[152,153], post-traumatic stress disorder[154], bipolar disorder[155], suicide[104] and autism spectrum disorder[156]. However, prediction of disease progression and response to therapy has been limited. One study identified a model based on three CpGs that predicts resistance to antidepressant treatment with a sensitivity of 0.6; although not currently sufficiently powerful for clinical use, this model may warrant further investigation[157].

Statistical methods of DNAm predictors

As the previous two sections have summarized, DNAm has begun to be used to characterize a diverse range of exposures and health phenotypes. However, the scientific and clinical value of these results ultimately depends on their appropriate use of reproducible and generalizable study designs, as well as robustly implemented statistical procedures. Unfortunately, neither has been applied consistently in the development of DNAm predictors to date.

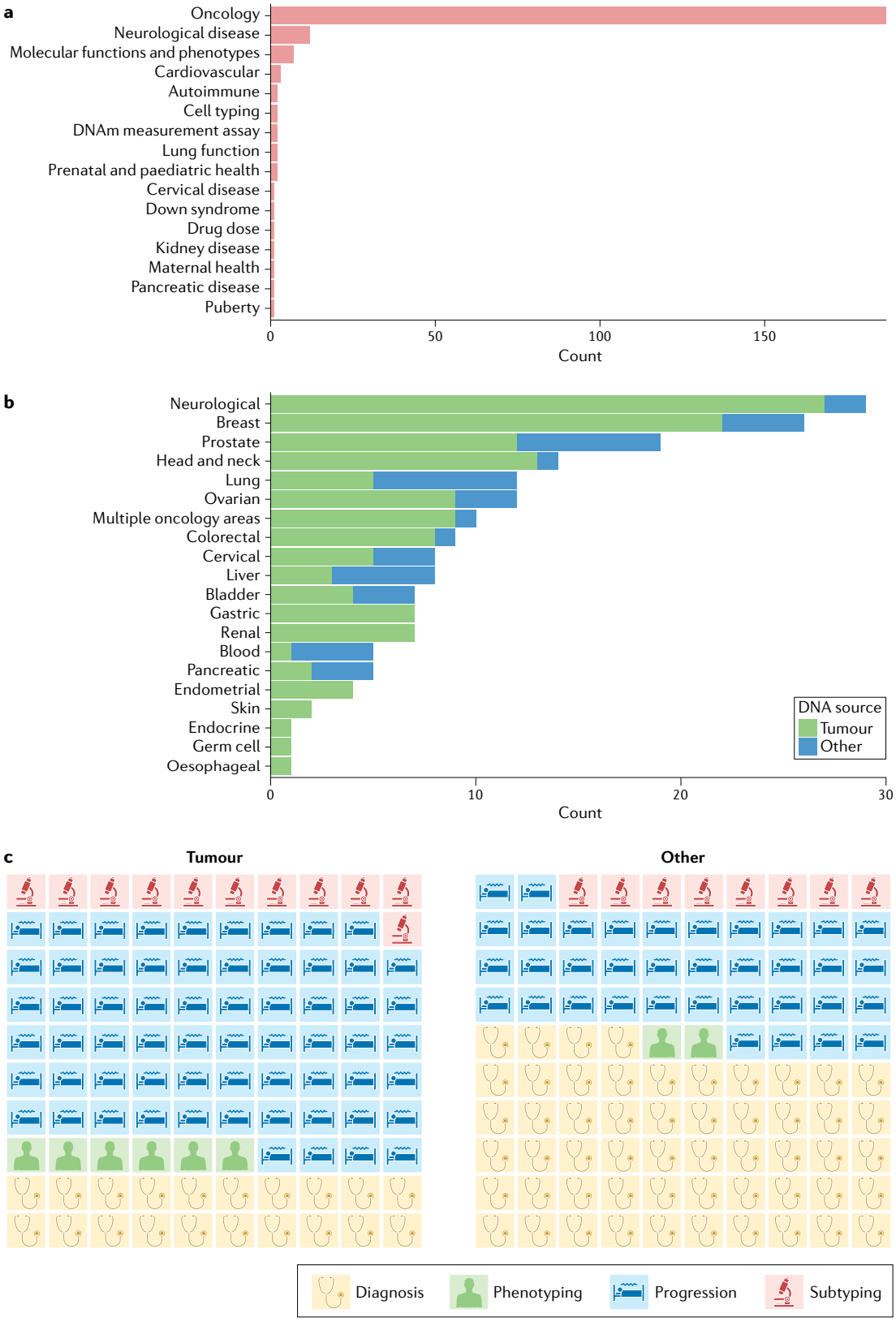
Feature selection and engineering. Feature selection, the process of identifying a set of possible input variables or features that best explains or predicts an outcome of interest, is a key step in prediction model development. Historically, where measurement cost per feature was high and candidates had to be identified before data collection, independent observations were simply sourced by using the results of earlier studies to select the relatively small number of most likely predictive features. However, as the cost of DNAm data generation has decreased, the large number of prospective features leads to encounters with the curse of dimensionality, whereby linearly increasing the number of input variables produces a corresponding exponential increase in the number of variable level combinations, meaning that the number of samples required to observe all combinations of levels will also increase exponentially. The number of possible combinations of features thus drastically exceeds the number of available observations. Here, researcher-agnostic, data-driven methods are needed to select combinations of features to identify signal relative to noise. For a detailed introduction to these concepts, we refer readers elsewhere[158].

One popular solution has been to apply a minimum P value threshold to EWAS results to identify the CpGs most strongly associated with the outcome of interest[124,159–161], an approach that has been successfully applied to genetic associations identified by GWAS for the development of polygenic scores. However, the resulting models tend to suffer from winner's curse, which is the tendency of the top associations among many tests to have inflated statistics[162]. High collinearity between features is a further complication of this approach that may produce unstable models. PRSs tend to skirt issues of collinearity by filtering SNPs on known correlation or linkage disequilibrium (LD) structure. Such structure is less well characterized in DNAm data, but techniques for removing collinear signals similar to PRS have occasionally been applied[161].

Penalized regression provides an alternative data-driven approach to feature selection and has recently become quite popular, particularly elastic nets, which are weighted combinations of lasso and ridge regressions. Penalized regression selects features and generates corresponding models simultaneously, reducing the danger of overfitting. Elastic nets have been the method of choice for several well-known predictors, including both first- and second-generation ageing clocks[27,42–44]. Perhaps the greatest perceived advantage of these models is their ease of interpretation: results are provided as familiar regression coefficients that, when inputs have been

Winner's curse
The phenomenon that strength of association is commonly overestimated in initial discovery samples and often experiences a regression to the mean in subsequent validation.

Linkage disequilibrium
(LD). Greater than chance co-occurrence or association of alleles at various loci due to nonrandom assortment.



centred and scaled, provide an expected measure of the importance of any particular variable. As researchers have often conflated predictive and explanatory aims in their modelling and sought to extract functional insights from their DNAm predictors, these approaches have clear relationships with those aims, although their results should not be used to draw causal inference.

However, it is notoriously difficult to predetermine which set of predictors or models will perform best, and a plethora of other methods are available from the machine

◀ Fig. 2 | **Illustrative distributions for studies of DNA methylation (DNAm) prediction of health outcomes for the majority of relevant studies published before April 2020.** Relevant studies (n = 278) were identified by PubMed query (search terms: "('DNA Methylation'[Mesh] AND prediction[All Fields]) AND 'humans'[MeSH Terms]") producing 822 initial results that were subsequently restricted to primary research studies, performed in humans, where DNAm measurements had been used to perform variable prediction. **a** | Overall distribution of health outcome application areas for DNAm prediction studies. **b** | Among studies with oncology applications, specific cancer application area distribution by DNA source tissue. **c** | Within oncology application studies, percentage of studies aimed at developing DNAm predictors of cancer diagnosis, phenotyping, progression or subtyping end points stratified by studies using tumour versus other DNA source tissues.

learning literature to identify a set of predictive features. Some reduce the number of choices by first applying one or more dimension reduction methods to construct new composite features that capture more variance, a process increasingly referred to as feature engineering[158]. The features that result from such approaches can use complicated combinations of inputs, and it often proves difficult to disentangle the contributions of specific measured variables after the fact. This can have knock-on ramifications for interpretability and even clinical translation; as such, 'black box' models are generally unacceptable for clinical use[163].

Performance assessment. To ensure their performance is generalizable to other data sets, prediction models should ideally minimize out-of-sample prediction error or errors observed when prediction models are applied to independent data. This is crucial because the alternative, in-sample prediction error or errors derived from observations used in estimating the predictive model, are subject to the bias–variance trade-off, whereby increasing model complexity improves in-sample performance by overfitting (that is, reducing in-sample bias) but proves unreproducible in novel data sets as small changes in the underlying data produce highly variable estimates. In-sample error estimates, such as within-sample AUC statistics, model or incremental coefficient of determination (R^2) values, are more appropriate for explanatory modelling settings (BOX 1) but have been commonly applied in this field in prediction contexts where they likely provide inflated estimates of performance[115,150,164,165].

Increasingly, however, DNAm predictors are being developed that target out-of-sample error by estimating performance in observations 'held out' during model building. The most basic approach is to randomly partition a data set into two subsets or folds. One subset is used for model training and the remaining subset, which is withheld from training to approximate new observations, is used for evaluating performance. More sophisticated methods, jointly referred to as resampling techniques, including cross-validation and bootstrapping, repeatedly split or resample the available data to generate multiple validation sets for estimating out-of-sample error. The resulting multiple rather than single observation of error provides more robust estimation of out-of-sample performance. Use of validation sets, or the observations split off for error assessment in the process of resampling, are particularly useful for model comparisons and for determining optimal values of important model hyperparameters (for example, the 'penalty' in penalized regression or number of trees in random forests). However, it is best practice to withhold another set of data, called a test set, completely aside for final prediction error assessment after modelling is finalized. We refer interested readers to an accessible introductory text on machine learning[166].

Data leakage. Even when using resampling techniques, estimates of prediction error are easily biased by data leakage, which occurs when the separation between training and test data is breached, in other words, when information regarding occurrence of an outcome is inadvertently passed from the training set to the evaluation set such that knowledge regarding the outcome in the evaluation set is no longer analogous to an event yet to occur in the chronological future. Under these circumstances, test-set-specific information influences model training, and the resulting models make assumptions that would not hold in newly collected data from the same target population, often leading to artificially inflated measures of predictor performance.

For resampling to produce valid estimates of prediction error, data partition must be maintained throughout the entire sequence of predictor development. Predictors that use data with substantial pre-processing requirements, as is common with DNAm data, may be especially vulnerable to unintentional violations, as data cleaning and normalization steps often borrow information across samples. Even simple steps such as centring and scaling variables before data partition can produce leaks by revealing mean and variance information during training. Frustratingly, many such violations are near impossible to detect without access to primary data.

The risk of data leakage is not limited to pre-processing steps. For example, leakage may occur during feature selection, as in one study[63] in which data were adequately prepared and partitioned when training a regression-based DNAm predictor of alcohol intake; however, when the final model was applied in the test set, the coefficients were re-estimated in the test data rather than predicting directly from the training set estimates, effectively 'peeking' at the test data[167].

Performance metrics. Model selection is highly dependent on choice or interpretation of performance metrics. Descriptions of fundamental performance metrics for both continuous and binary outcomes are provided in BOX 4, all of which have relative advantages and disadvantages regarding their ease of interpretation, stability, robustness to outliers and appropriateness for specific applications[168].

In discrete outcome classification settings, one of the most popular metrics is accuracy because it seems to have a straightforward interpretation. However, in settings with class imbalances — that is, when the relative amounts of outcome classes (for example, cases versus non-cases) differ substantially between training and testing sets — accuracy can give distorted impressions of performance, because it does not account for such frequency differences. For example, by always predicting a class that makes up 90% of the test set, one's predictions would be 90% accurate. Similar distortions can occur for

Feature engineering
The process of transforming or combining possible inputs (for example, by taking their principal components or rescaling their values) to make novel super-features that better explain or predict an outcome.

Out-of-sample prediction error
The discrepancy between estimates of an outcome $\hat{Y}$ generated by a predictive modelling function f and values of Y observed in a sample of data that was not available to f during model training.

In-sample prediction error
The discrepancy between estimated values of an outcome $\hat{Y}$ generated by a modelling function f and values of Y observed in a sample of data that was available to f during model training.

Resampling
Splitting, partitioning or sampling available data to generate subsamples in which model predictions can be tested and used to estimate distributions of out-of-sample errors.

Accuracy
The percentage of times all levels of a classifier agree with observed values of those levels.

Box 5 | **Peripheral versus target tissues**

Many DNA methylation (DNAm) disease studies are conducted in peripheral tissues, such as blood or saliva, instead of in more directly affected tissues, such as the brain for neurological diseases or the pancreas in diabetes mellitus. In some cases, this may be an example of the 'streetlight effect', a bias in which one searches for something where it is easiest to search for instead of where it is most likely located. However, there is some evidence that blood may contain clinically useful biomarkers for many diseases. In some cancers, dying cells in a growing tumour release DNA into the blood. Detecting the presence of this DNA could be used to diagnose the presence of the tumour or even identify tumour attributes such as its host organ, subtype or grade. This approach, using a commercial cell-free DNA (cfDNA)-based platform[132], is currently being trialled for clinical use by the UK National Health Service[138]. Alternatively, molecular patterns may arise in multiple tissues, including blood immune cells, owing to a common cause. For example, one study showed that more than half of the associations between gene expression and cognitive traits and psychiatric disorders discovered in brain could have been discovered in blood[214]. This was due to similar effects of genetic variants on gene expression in both brain and blood. Environmental exposures often also have similar effects on multiple tissues. Cigarette smoking is a prime example, with similar effects on DNAm across adipose tissue, blood, skin and lung[215]. A couple of studies have investigated DNAm in brain, saliva and blood samples collected from the same individuals[216,217]; both studies confirmed that DNAm variation at individual CpG sites was highly correlated between all pairs of tissues, but that CpG sites tended to have higher correlations between saliva and brain than between blood and brain. However, there remain large differences between saliva and brain, both of which are complex tissues composed of multiple cell types, each with their own DNAm patterns. Consequently, saliva and other peripheral tissues can never be expected to fully capture all phenotypic variation found in an organ as complex as the brain. Single-cell RNA sequencing suggests that there may be more than two dozen different cell types in the prefrontal cortex[218]. Beyond blood and saliva, there is a range of accessible DNA sources that have been included to date in only a small number of studies, including urine, faeces, fat, synovial fluid, skin, nasal epithelium, colonic epithelium, cord and placenta. These seem to capture at least some complementary biological variation that could be used to enhance predictive performance in peripheral tissues.

Confusion matrix
A frequency table of agreement and disagreement between observed and predicted values of an outcome variable. It is used to compute many classification metrics, including, among others, accuracy, sensitivity and specificity.

Cohen's kappa
A confusion matrix metric ranging from −1 (total disagreement between observed and predicted classes) to 1 (total agreement), where class imbalances are corrected by normalizing to the expected error rate.

Matthews correlation coefficient
A numerical summary of agreement in a confusion matrix, ranging from −1 (total disagreement) to 1 (total agreement), that seeks to correct for class imbalances using a method similar to that of a χ^2 statistic.

Calibration
The extent to which predicted outcome risk matches observed outcome proportions.

other statistics that naively summarize a confusion matrix without considering such imbalances, including positive and negative predictive value, among others. Alternative metrics, including Cohen's kappa and Matthews correlation coefficient, directly account for such imbalances and offer more robust measures of model performance[169,170].

Most performance metrics naively give equal weight to all types of error, which often does not correspond to the true value of errors encountered for many clinical applications. For example, when treatment is expensive or when there may be substantial personal consequences to the individual receiving treatment from an unnecessary intervention, false positives may need to be minimized. Or in a converse example, when failure to treat results in death, reducing false negatives is more important. In either case, careful examination of the resulting confusion matrix is needed to select for or against specific error categories.

In a continuous setting, such as when predicting risk probability, model calibration can be equally impactful for clinical decision-making where acceptable risk thresholds can vary between patients. Even for models that correctly rank patients by risk, true risk estimates could be inaccurate, leading to unacceptable clinical decision-making. For example, patients with higher breast cancer risk estimates could indeed be more likely to die within the next 5 years, but if risk estimates are systematically lower than true risk, then many patients may underestimate the potential benefit of treatment.

Sample generalizability. It is important that the sample population used for DNAm predictor development be generalizable to the target population to which the predictor will be applied. Systematic differences between development and application populations can have unpredictable and misleading interpretations. For example, the application of a gestational age clock trained without preterm births gives spurious results when applied to samples with large numbers of preterm individuals[171].

Given substantial historical over-representation of European participants in genomic studies[172], populations of non-European ancestries are particularly vulnerable to such mis-characterizations from lack of representation in DNAm training data. Uninformed cross-population applications of DNAm predictors can reinforce or exacerbate health inequalities[173]. Researchers should be aware of these historical injustices and approach sample selection intentionally instead of relying on convenient samples that are likely to confirm biases. Considerations on reporting standards and metric selections for model generalizability being proposed in the machine learning literature are also applicable to DNAm predictor development[174,175].

Sample size. The considerations relevant to determining appropriateness of sample size for predictive modelling differ from those relevant to analyses with explanatory aims. Most notable is the need to reserve adequate numbers of observations for model validation and testing, to estimate out-of-sample performance. The optimal proportion of observations to be allocated for each will vary by application depending on the signal to noise ratio. One commonly cited rule of thumb recommends "50% for training, and 25% each for validation and testing"[176], but other optimal splits, including reserving 10% of observations for testing, have been proposed[177].

A priori estimates of appropriate training set sample sizes are complicated when many potential input features of unknown strength are being considered, as is often the case for development of DNAm predictors, or when the number of model parameters is uncertain, which is typical of complicated, so-called black box machine learning models. At present, few if any, methods for power analysis exist that apply generally to a broad range of machine learning models. A commonly used rule of thumb for classification settings — that ten observations should be available for each prediction parameter included[178] — has been demonstrated to be inappropriate under various reasonable modelling assumptions[179]. Riley and colleagues provide some of the most thorough explorations of the considerations that influence power for penalized regression models, which can provide a valuable starting point for producing power calculations for machine learning applications[179,180].

Future directions and opportunities

The study design and methodology used to date to develop DNAm predictors have already begun to generate useful results for researcher and clinicians. However, molecular phenotyping and large-scale data availability are likely to expand in the near future, potentially substantially increasing the value of resulting predictors.

Sequencing. Most predictors available to date have relied on microarray data. However, sequencing is increasingly considered the gold standard for measuring DNAm, because it provides a binary read-out of the methylation status of individual cytosines within DNA fragments that is easier to interpret than probe signal intensities. With the costs of whole-genome DNAm sequencing (WGMS) continually decreasing[181], microarrays may soon be replaced as the standard for data generation in population-based epidemiological studies.

Analysing WGMS will be challenging, not least because managing the increase in genomic coverage from <3% to nearly the complete genome will exacerbate the curse of dimensionality. The problem could be somewhat offset by the improved characterization of the methylome and incorporation of this characterization into more sophisticated dimension reduction techniques. However, current methods are still rudimentary[182] and improvements difficult, as the methylome does not have the simple LD-style correlation structure of the genome[183]. Therefore, sample sizes will almost certainly need to increase to discover and make use of the additional biological signal.

Targeted methylation sequencing of subsets of the genome is an emerging compromise between the increased interpretability of sequencing and the prohibitive cost of genome-wide sequencing[184,185]. Genomic targets may be identified by hypothesis or from WGMS studies of small sample size or with low sequencing depth. Target sets range from small panels consisting of tens of genomic loci up to all gene promoter regions or all exons.

Nanopore long-read methylation sequencing, although currently even less feasible for population-based studies than WGMS, is undergoing rapid development and will likely follow a path towards feasibility similar to that of WGMS[186,187]. The current challenges are reduction of measurement noise and both financial and computational costs. Beyond WGMS, long-read methylation sequencing will allow interrogation of many regions with repetitive DNA sequence, enabling models of DNAm to include detailed information about DNAm dependencies between cytosines hundreds of bases apart. Even the limited dependency information that can be derived from short-read sequencing enables improved prediction of gene expression levels from DNAm[188].

Multi-omics. As omic technologies proliferate and become more widely available, the range of applications that will seek to combine multi-omic measures into single prediction models will grow. Genomics, transcriptomics, proteomics and metabolomics are all becoming more common[189,190]. Some early examples have demonstrated that combining genetic and DNAm data can provide added value for prediction by supplying complementary sources of information[145,191]. Although greatly dependent on the specific phenotype and sample type used, DNAm has shown additive capacity to explain some health phenotypes beyond genetic predictors given similarly sized or even smaller amounts of training observations[25,26,192]. Already, some specific variance components modelling approaches have sought to take advantage of the diverse range of factors in addition to DNAm, including other omics, that can be drawn upon to improve complex trait prediction[193,194]. However, combining multiple omic measures again expands the set of available features, adding to the curse of dimensionality. The total number of features added will differ for each new class of omic measurement considered and may or may not be of a similar amount to adding WGMS data. Multi-omic predictors will improve on their single omic competitors where the additional omics adds information on other dimensions of exposures and health outcomes[195–198] such as the time frame they operate on, cell type specificity and proximity to exposure or outcome source[199]. However, there is already some evidence to suggest overlap between omics, allowing one omic molecular phenotype to be effectively predicted by another, for example, when DNAm predictors were developed for proteomic measurements[200].

Best practices. Regardless of the expansion of data acquisition across any number of dimensions, prediction modelling using DNAm data has substantial potential for improvement by greater adoption of statistical best practices. Prediction modelling applications are vulnerable to several common design and implementation pitfalls that can bias and seriously restrict the impact of results. To help future researchers avoid the worst of these, we have adapted a practical checklist of best practices[201,202] (BOX 6).

Beyond improvements in methodology and reporting surrounding predictor development needed from individual investigators, advances are also needed from the broader research environment to encourage and incentivize 'reproducible science' in this area. As things stand, there are substantial incentives for researchers to report inflated DNAm predictor performance (greater publication impact, grant funding) and insufficient mechanisms to ensure accountability and transparency. One key way that the latter could be improved would be through greater adoption of outcome-specific, clinically relevant benchmarking data sets[203].

As mentioned earlier, omic predictor development must include greater representation of individuals of diverse ancestry. Diversity is a key factor influencing the value of predictors in clinical or research use[173].

Conclusions

At present, prediction is often a secondary goal in studies that examine DNAm levels after primary mechanistic hypotheses have been pursued. This leaves substantial opportunity to expand development of DNAm-based predictors across diverse application areas and with improved clinical and research value, even using existing data collections. However, achieving these improvements will require increased use of more rigorous statistical approaches and performance evaluations. This includes greater appreciation of the likelihood of data leakage and its propensity to inflate reported performance. Progress in the application of machine learning statistical approaches will require improvements to facilitate biological understanding of black box methods and enable clinical uptake. Other critical improvements include increased

Box 6 | **Checklist of statistical best practices for DNAm prediction development**

We highlight below a few aspects of prediction modelling that have proved problematic in the DNA methylation (DNAm) predictor literature to date. For a more comprehensive discussion of prediction modelling best practices see the TRIPOD statement[201,202].

Study design

- Any outcome definition used should be selected in light of providing the greatest value to the intended audience. For example, if a cumulative exposure measure is more informative for clinical decision-making than a binary exposed/unexposed categorization, the former should be targeted.
- The population used for predictor development needs to generalize to the population in which it will be applied. Systematic differences between these populations can skew interpretation, often in incomprehensible ways. Differences could include things like age or sex distribution, but perhaps most important to consider is genetic ancestry.
- Training sample size needs to be large enough to achieve desired performance levels. Early studies will tend to be small but should be large enough to indicate potential for usefully improving on current clinical practice. Follow-up studies should then aim to determine whether clinically relevant performance improvements are possible.

Resampling and performance metrics

- Out-of-sample measures of prediction error are most informative for assessing prediction performance. Estimates of out-of-sample error can be easily generated within a single data set through resampling techniques. However, it is still useful to observe results in an independent test set. Ideal performance characterization would compare results using a pre-specified, community agreed, independent benchmarking data set.
- To avoid data leakage and artificially inflated estimates of performance, data splits generated must be respected throughout the entire sequence of data processing, feature selection and/or engineering and model fitting.
- Class imbalances need to be considered carefully when measuring predictor performance. Many familiar metrics such as accuracy are easily distorted by such class imbalances. Advanced metrics (for example, Cohen's kappa, Matthews correlation coefficient), class weighting, as well as over- and under-sampling approaches can provide more stable estimates of prediction performance, but may require extra attention for interpretation.

Model development

- Feature selection and model fitting are two distinct processes, but data leakage can occur at either step if data splits are not maintained throughout.
- Stepwise regression techniques have well-characterized flaws and limitations. They should be avoided as a general rule.
- Confirm any model-specific prerequisites to ensure valid interpretation of results. For example, many methods require input features to be centred and scaled before implementation.
- Engineered features or models that operate as black boxes require post hoc interpretation to characterize the relationships between observed input variables and model prediction or risk losing clinical value.

Reporting of results

- Clear, transparent reporting of development methodology is crucial to establishing the value of any DNAm predictor for clinical and research audiences. Participant selection, the precise outcome definition used, which features have been retained or engineered, and the final model implemented are all essential for readers to adopt the results into their own practice. We suggest including the following.
  - A detailed flow diagram describing how samples have been selected and split across all training, validation and testing data partitions (including *N*s)
  - Exact description of the samples used for both feature selection and model fitting, ideally integrated into the same flow chart mentioned above
  - The precise operational definition of the outcome being targeted, including all relevant scale, thresholding and cut-point details
  - Sufficient model reporting that independent researchers could reasonably recreate the prediction model, in accordance with the basic principles of reproducible science. For example, when possible, model functional form should be reported as formulae. When infeasible owing to model complexity, a reproducible, open-source software implementation of the model should be distributed.

sample size, data accessibility, phenotyping and participant diversity in data sets for both training and testing. In the meantime, available data can often be used more efficiently by careful consideration of observation partitioning and use of resampling procedures. Lastly, further attention to choice of outcome metric and the types of error being minimized can be used to more clearly and cost-efficiently communicate predictor relevance to target audiences. These improvements will unlock the enormous potential that exists to harness the properties of DNAm in prediction studies, be they for the purpose of diagnosis, prognosis, treatment response or otherwise.

Published online 18 March 2022

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **396**, 1204–1222 (2020).
2. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
3. Khera, A. V. et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* **177**, 587–596.e9 (2019).
4. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
5. Roberts, N. J. et al. The predictive capacity of personal genome sequencing. *Sci. Transl. Med.* **4**, 133ra58 (2012).
6. Adeyemo, A. et al. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat. Med.* **27**, 1876–1884 (2021).
7. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2017).
8. Ala-Korpela, M. & Holmes, M. V. Polygenic risk scores and the prediction of common diseases. *Int. J. Epidemiol.* **49**, 1–3 (2020).
9. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571**, 489–499 (2019).
10. Teschendorff, A. E. et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440–446 (2010).
11. Petronis, A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**, 721–727 (2010).
12. Baubec, T. & Schübeler, D. Genomic patterns and context specific interpretation of DNA methylation. *Curr. Opin. Genet. Dev.* **25**, 85–92 (2014).
13. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
14. Kim, M. & Costello, J. DNA methylation: an epigenetic mark of cellular memory. *Exp. Mol. Med.* **49**, 49 (2017).
15. Russo, V. E. A., Martienssen, R. A. & Riggs, A. D. *Epigenetic Mechanisms of Gene Regulation* (Cold Spring Harbor laboratory Press, 1996).
16. Lappalainen, T. & Greally, J. M. Associating cellular epigenetic models with human phenotypes. *Nat. Rev. Genet.* **18**, 441–451 (2017).
17. Hou, L., Zhang, X., Wang, D. & Baccarelli, A. Environmental chemical exposures and human epigenetics. *Int. J. Epidemiol.* **41**, 79–105 (2012).
18. Perera, F. & Herbstman, J. Prenatal environmental exposures, epigenetics, and disease. *Reprod. Toxicol.* **31**, 363–373 (2011).
19. Laird, P. W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
20. Foley, D. L. et al. Prospects for epigenetic epidemiology. *Am. J. Epidemiol.* **169**, 389–400 (2009).
21. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **8**, 389–399 (2015).
22. Sandoval, J. et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).

23. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
24. Bell, C. G. et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* **20**, 249 (2019).
25. McRae, A. F. et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol.* **15**, R73 (2014).
26. McCartney, D. L. et al. Epigenetic prediction of complex traits and death. *Genome Biol.* **19**, 136 (2018).
This paper systematically demonstrates that DNAm could predict a whole range of risk factors and exposures, with explanatory capacity roughly equal to or better than polygenic risk predictors.
27. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
This early epigenetic clock is broadly applicable owing to its multi-tissue training set and accordingly saw widespread use as a biomarker of biological ageing in many epidemiological studies.
28. Bocklandt, S. et al. Epigenetic predictor of age. *PLoS ONE* **6**, e14821 (2011).
This is the first paper to report a DNAm predictor of age, or epigenetic clock.
29. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
30. Crimmins, E. M., Thyagarajan, B., Levine, M. E., Weir, D. R. & Faul, J. Associations of age, sex, race/ethnicity and education with 13 epigenetic clocks in a nationally representative US sample: the Health and Retirement Study. *J. Gerontol. Ser. A biol. Sci. Med. Sci.* **76**, 1117–1123 (2021).
31. Rakyan, V. K. et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* **20**, 434–439 (2010).
32. Boks, M. P. et al. Longitudinal changes of telomere length and epigenetic age related to traumatic stress and post-traumatic stress disorder. *Psychoneuroendocrinology* **51**, 506–512 (2015).
33. Zannas, A. S. et al. Lifetime stress accelerates epigenetic aging in an urban, African American cohort: relevance of glucocorticoid signaling. *Genome Biol.* **16**, 266 (2015).
34. Horvath, S. et al. Obesity accelerates epigenetic aging of human liver. *Proc. Natl Acad. Sci. USA* **111**, 15538–15543 (2014).
35. Marioni, R. E. et al. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *Int. J. Epidemiol.* **44**, 1388–1396 (2015).
36. Levine, M. E. et al. DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. *Aging* **7**, 690–700 (2015).
37. Marioni, R. E. et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 25 (2015).
38. Marioni, R. E. et al. The epigenetic clock and telomere length are independently associated with chronological age and mortality. *Int. J. Epidemiol.* **45**, 424–432 (2016).
39. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
40. Horvath, S. & Ritz, B. R. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. *Aging* **7**, 1130–1142 (2015).
41. Zhang, Q. et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* **11**, 887–897 (2019).
42. Levine, M. E. et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).
43. Lu, A. T. et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303 (2019).
This paper presents an influential second-generation epigenetic clock and demonstrates that DNAm predictors of molecular phenotypes, risk factors and exposures can be usefully combined.
44. Belsky, D. W. W. et al. Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. *eLife* **9**, e54870 (2020).
45. Lu, A. T. et al. GWAS of epigenetic aging rates in blood reveals a critical role for TERT. *Nat. Commun.* **9**, 387 (2018).
46. Gibson, J. et al. A meta-analysis of genome-wide association studies of epigenetic age acceleration. *PLoS Genet.* **15**, e1008104 (2019).
47. McCartney, D. L. et al. Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging. *Genome Biol.* **22**, 1–25 (2021).
48. Vetter, V. M. et al. Epigenetic clock and relative telomere length represent largely different aspects of aging in the Berlin aging study II (BASE-II). *J. Gerontol. A Biol. Sci. Med. Sci.* **74**, 27–32 (2019).
49. Joehanes, R. et al. Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).
This paper is the largest EWAS on cigarette smoking in adults with almost 16,000 participants and identifies differential DNAm between current and never smokers at 2,623 CpG sites.
50. Zeilinger, S. et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS ONE* **8**, e63812 (2013).
51. Guida, F. et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum. Mol. Genet.* **24**, 2349–2359 (2015).
52. Maas, S. C. E. et al. Validated inference of smoking habits from blood with a finite DNA methylation marker set. *Eur. J. Epidemiol.* **34**, 1055–1074 (2019).
53. McCartney, D. L. et al. Epigenetic signatures of starting and stopping smoking. *EBioMedicine* **37**, 214–220 (2018).
54. Corley, J. et al. Epigenetic signatures of smoking associate with cognitive function, brain structure, and mental and physical health outcomes in the Lothian Birth Cohort 1936. *Transl. Psychiatry* **9**, 248 (2019).
55. Su, D. et al. Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. *PLoS ONE* **11**, e0166486 (2016).
56. You, C. et al. A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. *Nat. Commun.* **11**, 4779 (2020).
57. Benowitz, N. L. et al. Biochemical verification of tobacco use and abstinence: 2019 update. *Nicotine Tob. Res.* **22**, 1086–1097 (2020).
58. Richmond, R. C., Suderman, M., Langdon, R., Relton, C. L., & Davey Smith, G. DNA methylation as a marker for prenatal smoke exposure in adults. *Int. J. Epidemiol.* **47**, 1120–1130 (2018).
59. Wiklund, P. et al. DNA methylation links prenatal smoking exposure to later life health outcomes in offspring. *Clin. Epigenetics* **11**, 97 (2019).
60. Bojesen, S. E., Timpson, N., Relton, C., Davey Smith, G. & Nordestgaard, B. G. *AHRR* (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax* **72**, 646–653 (2017).
This paper provides a clear example of how DNAm can proxy an established risk factor and out-perform the measurement of that risk factor in predicting morbidity and mortality.
61. Tu, W., Chu, C., Li, S. & Liangpunsakul, S. Development and validation of a composite score for excessive alcohol use screening. *J. Investig. Med.* **64**, 1006–1011 (2016).
62. Joubert, B. R. et al. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *Am. J. Hum. Genet.* **98**, 680–696 (2016).
63. Liu, C. et al. A DNA methylation biomarker of alcohol consumption. *Mol. Psychiatry* **23**, 422–433 (2018).
64. Clarke, T. K. et al. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK biobank (*N* = 112117). *Mol. Psychiatry* **22**, 1376–1384 (2017).
65. Taylor, M., Simpkin, A. J., Haycock, P. C., Dudbridge, F. & Zuccolo, L. Exploration of a polygenic risk score for alcohol consumption: a longitudinal analysis from the ALSPAC cohort. *PLoS ONE* **11**, e0167360 (2016).
66. Philibert, R., Dogan, M., Beach, S. R. H., Mills, J. A. & Long, J. D. AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **183**, 51–60 (2020).
67. Yousefi, P. D. et al. Validation and characterisation of a DNA methylation alcohol biomarker across the life course. *Clin. Epigenetics* **11**, 163 (2019).
68. Wahl, S. et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
This paper provided an early demonstration of the value of DNAm predictors in relation to disease discrimination, by showing that a DNAm score for BMI is associated with incident type 2 diabetes.
69. Dick, K. J. et al. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383**, 1990–1998 (2014).
70. Mendelson, M. M. et al. Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a Mendelian randomization approach. *PLoS Med.* **14**, e1002215 (2017).
71. Reed, Z. E., Suderman, M. J., Relton, C. L., Davis, O. S. P. & Hemani, G. The association of DNA methylation with body mass index: distinguishing between predictors and biomarkers. *Clin. Epigenetics* **12**, 50 (2020).
72. Keller, M. et al. DNA methylation signature in blood mirrors successful weight-loss during lifestyle interventions: the CENTRAL trial. *Genome Med.* **12**, 97 (2020).
73. Crocker, K. C. et al. DNA methylation and adiposity phenotypes: an epigenome-wide association study among adults in the Strong Heart Study. *Int. J. Obes.* **44**, 2313–2322 (2020).
74. Justice, A. E. et al. Methylome-wide association study of central adiposity implicates genes involved in immune and endocrine systems. *Epigenomics* **12**, 1483–1499 (2020).
75. Vehmeijer, F. O. L. et al. DNA methylation and body mass index from birth to adolescence: meta-analyses of epigenome-wide association studies. *Genome Med.* **12**, 105 (2020).
76. Mandaviya, P. R. et al. Association of dietary folate and vitamin B-12 intake with genome-wide DNA methylation in blood: a large-scale epigenome-wide association analysis in 5841 individuals. *Am. J. Clin. Nutr.* **110**, 437–450 (2019).
77. Gensous, N. et al. One-year Mediterranean diet promotes epigenetic rejuvenation with country- and sex-specific effects: a pilot study from the NU-AGE project. *GeroScience* **42**, 687–701 (2020).
78. Ma, J. et al. Whole blood DNA methylation signatures of diet are associated with cardiovascular disease risk factors and all-cause mortality. *Circ. Genom. Precis. Med.* **13**, 324–333 (2020).
79. Do, W. L. et al. Epigenome-wide association study of diet quality in the Women's Health Initiative and TwinsUK cohort. *Int. J. Epidemiol.* **50**, 675–684 (2021).
80. Gomez-Alonso, M del C. et al. DNA methylation and lipid metabolism: an EWAS of 226 metabolic measures. *Clin. Epigenetics* **13**, 7 (2021).
81. Antoun, E. et al. Maternal dysglycaemia, changes in the infant's epigenome modified with a diet and physical activity intervention in pregnancy: secondary analysis of a randomised control trial. *PLoS Med.* **17**, e1003229 (2020).
82. Irwin, R. E. et al. A randomized controlled trial of folic acid intervention in pregnancy highlights a putative methylation-regulated control element at ZFP57. *Clin. Epigenetics* **11**, 31 (2019).
83. Sharp, G. C. et al. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium. *Hum. Mol. Genet.* **26**, 4067–4085 (2017).
84. Howe, C. G. et al. Maternal gestational diabetes and newborn DNA methylation: findings from the Pregnancy and Childhood Epigenetics consortium. *Diabetes Care* **43**, dc190524 (2019).
85. Ouidir, M. et al. Early pregnancy dyslipidemia is associated with placental DNA methylation at loci relevant for cardiometabolic diseases. *Epigenomics* **12**, 921–934 (2020).
86. Agha, G. et al. Adiposity is associated with DNA methylation profile in adipose tissue. *Int. J. Epidemiol.* **44**, 1277–1287 (2015).
87. Huang, Y. T. et al. Epigenome-wide profiling of DNA methylation in paired samples of adipose tissue and blood. *Epigenetics* **11**, 227–236 (2016).
88. Allum, F. et al. Dissecting features of epigenetic variants underlying cardiometabolic risk using full-resolution epigenome profiling in regulatory elements. *Nat. Commun.* **10**, 1209 (2019).
89. Richmond, R. C. et al. DNA methylation and BMI: investigating identified methylation sites at HIF3A in a causal framework. *Diabetes* **65**, 1231–1244 (2016).
90. Sun, D. et al. Body mass index drives changes in DNA methylation: a longitudinal study. *Circ. Res.* **125**, 824–833 (2019).
91. Gudsnuk, K. & Champagne, F. A. Epigenetic influence of stress and the social environment. *ILAR J.* **53**, 279–288 (2012).
92. Cunliffe, V. T. The epigenetic impacts of social stress: how does social adversity become biologically embedded? *Epigenomics* **8**, 1653–1669 (2016).
93. Borghol, N. et al. Associations with early-life socio-economic position in adult DNA methylation. *Int. J. Epidemiol.* **41**, 62–74 (2012).

94. Chen, D., Meng, L., Pei, F., Zheng, Y. & Leng, J. A review of DNA methylation in depression. *J. Clin. Neurosci.* **43**, 39–46 (2017).
95. Vukojevic, V. et al. Epigenetic modification of the glucocorticoid receptor gene is linked to traumatic memory and post-traumatic stress disorder risk in genocide survivors. *J. Neurosci.* **34**, 10274–10284 (2014).
96. Yehuda, R. et al. Lower methylation of glucocorticoid receptor gene promoter 1F in peripheral blood of veterans with posttraumatic stress disorder. *Biol. Psychiatry* **77**, 356–364 (2015).
97. Non, A. L. et al. DNA methylation at stress-related genes is associated with exposure to early life institutionalization. *Am. J. Phys. Anthropol.* **161**, 84–93 (2016).
98. McGowan, P. O. et al. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nat. Neurosci.* **12**, 342–348 (2009).
99. Suderman, M. et al. Childhood abuse is associated with methylation of multiple loci in adult DNA. *BMC Med. Genomics* **7**, 13 (2014).
100. Hostinar, C. E., Sullivan, R. M. & Gunnar, M. R. Psychobiological mechanisms underlying the social buffering of the hypothalamic-pituitary-adrenocortical axis: a review of animal models and human studies across development. *Psychol. Bull.* **140**, 256–282 (2014).
101. Swartz, J. R., Hariri, A. R. & Williamson, D. E. An epigenetic mechanism links socioeconomic status to changes in depression-related brain function in high-risk adolescents. *Mol. Psychiatry* **22**, 209–214 (2017).
102. Clark, S. L. et al. A methylation study of long-term depression risk. *Mol. Psychiatry* **25**, 1334–1343 (2020).
103. Barbu, M. C. et al. Epigenetic prediction of major depressive disorder. *Mol. Psychiatry* **26**, 5112–5123 (2021).
104. Clive, M. L. et al. Discovery and replication of a peripheral tissue DNA methylation biosignature to augment a suicide prediction model. *Clin. Epigenetics* **8**, 113 (2016).
105. Yang, X., Gao, L. & Zhang, S. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Brief. Bioinform.* **18**, 761–773 (2017).
106. Zhang, J. & Huang, K. Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. *BMC Genomics* **18**, 1045 (2017).
107. Tao, Y. et al. Aging-like spontaneous epigenetic silencing facilitates Wnt activation, stemness, and Braf V600E -induced tumorigenesis. *Cancer Cell* **35**, 315–328.e6 (2019).
108. Chen, Y. et al. MGMT promoter methylation and glioblastoma prognosis: a systematic review and meta-analysis. *Arch. Med. Res.* **44**, 281–290 (2013).
109. Wick, W. et al. Temozolomide chemotherapy alone versus radiotherapy alone for malignant astrocytoma in the elderly: the NOA-08 randomised, phase 3 trial. *Lancet Oncol.* **13**, 707–715 (2012).
110. Malmström, A. et al. Temozolomide versus standard 6-week radiotherapy versus hypofractionated radiotherapy in patients older than 60 years with glioblastoma: the Nordic randomised, phase 3 trial. *Lancet Oncol.* **13**, 916–926 (2012).
111. Loeb, S. et al. Overdiagnosis and overtreatment of prostate cancer. *Eur. Urol.* **65**, 1046–1055 (2014).
112. Jørgensen, K. J. & Gøtzsche, P. C. Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *BMJ* **339**, 206–209 (2009).
113. Hulbert, A. et al. Early detection of lung cancer using DNA promoter hypermethylation in plasma and sputum. *Clin. Cancer Res.* **23**, 1998–2005 (2017).
114. Li, L. et al. Diagnosis of pulmonary nodules by DNA methylation analysis in bronchoalveolar lavage fluids. *Clin. Epigenetics* **13**, 185 (2021).
115. Dvorská, D. et al. Aberrant methylation status of tumour suppressor genes in ovarian cancer tissue and paired plasma samples. *Int. J. Mol. Sci.* **20**, 4119 (2019).
116. Majumder, S. et al. Novel methylated DNA markers discriminate advanced neoplasia in pancreatic cysts: marker discovery, tissue validation, and cyst fluid testing. *Am. J. Gastroenterol.* **114**, 1539–1549 (2019).
117. Sanchez-Cespedes, M. et al. Gene promoter hypermethylation in tumors and serum of head and neck cancer patients. *Cancer Res.* **60**, 892–895 (2000).
118. Nakahara, Y., Shintani, S., Mihara, M., Hino, S. & Hamakawa, H. Detection of p16 promoter methylation in the serum of oral cancer patients. *Int. J. Oral. Maxillofac. Surg.* **35**, 362–365 (2006).
119. Nakayama, H. et al. Molecular detection of p16 promoter methylation in the serum of colorectal cancer patients. *Cancer Lett.* **188**, 115–119 (2002).
120. Ooki, A. et al. A panel of novel detection and prognostic methylated DNA markers in primary non–small cell lung cancer and serum DNA. *Clin. Cancer Res.* **23**, 7141–7152 (2017).
121. Guan, Z. et al. Individual and joint performance of DNA methylation profiles, genetic risk score and environmental risk scores for predicting breast cancer risk. *Mol. Oncol.* **14**, 42–53 (2020).
122. Onwuka, J. U. et al. A panel of DNA methylation signature from peripheral blood may predict colorectal cancer susceptibility. *BMC Cancer* **20**, 692 (2020).
123. Walker, R. M. et al. Epigenome-wide analyses identify DNA methylation signatures of dementia risk. *Alzheimer's Dement.* **12**, e12078 (2020).
124. Baglietto, L. et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *Int. J. Cancer* **140**, 50–61 (2017).
125. Zhang, Y. et al. Smoking-associated DNA methylation markers predict lung cancer incidence. *Clin. Epigenetics* **8**, 127 (2016).
126. Wang, L. et al. Methylation markers for small cell lung cancer in peripheral blood leukocyte DNA. *J. Thorac. Oncol.* **5**, 778–785 (2010).
127. Pedersen, K. S. et al. Leukocyte DNA methylation signature differentiates pancreatic cancer patients from healthy controls. *PLoS ONE* **6**, e18223 (2011).
128. Michaud, D. S. et al. Epigenome-wide association study using prediagnostic bloods identifies new genomic regions associated with pancreatic cancer risk. *JNCI Cancer Spectr.* **4**, pkaa041 (2020).
129. Xu, R. H. et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat. Mater.* **16**, 1155–1162 (2017).
130. Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
131. Roy, D. & Tiirikainen, M. Diagnostic power of DNA methylation classifiers for early detection of cancer. *Trends cancer* **6**, 78–81 (2020).
132. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
133. Nassiri, F. et al. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat. Med.* **26**, 1044–1047 (2020).
134. Nuzzo, P. V. et al. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nat. Med.* **26**, 1041–1043 (2020).
135. Guler, G. D. et al. Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA. *Nat. Commun.* **11**, 5270 (2020).
136. Tse, R. T.-H. et al. Urinary cell-free DNA in bladder cancer detection. *Diagnostics* **11**, 306 (2021).
137. Luo, H. et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Sci. Transl. Med.* **12**, eaax7533 (2020).
138. NHS. NHS to pilot potentially revolutionary blood test that detects more than 50 cancers. https://www.england.nhs.uk/2020/11/nhs-to-pilot-potentially-revolutionary-blood-test/ (2021).
139. Klein, E. A. et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann. Oncol.* **32**, 1167–1177 (2021).
This study demonstrates the ability of cell-free DNA polymorphisms and DNAm to discriminate >50 cancer types and tissue of origin.
140. Richard, M. A. et al. DNA methylation analysis identifies loci for blood pressure regulation. *Am. J. Hum. Genet.* **101**, 888–902 (2017).
The largest blood pressure EWAS to date, with information from more than 17,000 participants, which found that a 13 CpG score could explain only between 1% and 2% of the variance of systolic and diastolic blood pressure, respectively.
141. Huang, Y. et al. Identification, heritability, and relation with gene expression of novel DNA methylation loci for blood pressure. *Hypertension* **76**, 195–205 (2020).
142. Fernández-Sanlés, A., Sayols-Baixeras, S., Subirana, I., Degano, I. R. & Elosua, R. Association between DNA methylation and coronary heart disease or other atherosclerotic events: a systematic review. *Atherosclerosis* **263**, 325–333 (2017).
143. Westerman, K. et al. DNA methylation modules associate with incident cardiovascular disease and cumulative risk factor exposure. *Clin. Epigenetics* **11**, 142 (2019).
144. Shen, Y. et al. Epigenome-wide association study indicates hypomethylation of MTRNR2L8 in large-artery atherosclerosis stroke. *Stroke* **50**, 1330–1338 (2019).
145. Dogan, M. V., Grumbach, I. M., Michaelson, J. J. & Philibert, R. A. Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart Study. *PLoS ONE* **13**, e0190549 (2018).
146. Westerman, K. et al. Epigenomic assessment of cardiovascular disease risk and interactions with traditional risk metrics. *J. Am. Heart Assoc.* **9**, e015299 (2020).
147. Nuotio, M. L. et al. An epigenome-wide association study of metabolic syndrome and its components. *Sci. Rep.* **10**, 20567 (2020).
148. Chambers, J. C. et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case–control study. *Lancet Diabetes Endocrinol.* **3**, 526–534 (2015).
149. Cardona, A. et al. Epigenome-wide association study of incident type 2 diabetes in a British population: EPIC-Norfolk study. *Diabetes* **68**, 2315–2326 (2019).
150. Xu, C. et al. Elevated methylation of OPRM1 and OPRL1 genes in Alzheimer's disease. *Mol. Med. Rep.* **18**, 4297–4302 (2018).
151. Wang, C., Chen, L., Yang, Y., Zhang, M. & Wong, G. Identification of potential blood biomarkers for Parkinson's disease by gene expression and DNA methylation data integration analysis. *Clin. Epigenetics* **11**, 24 (2019).
152. Osborne, L. et al. Replication of epigenetic postpartum depression biomarkers and variation with hormone levels. *Neuropsychopharmacology* **41**, 1648–1658 (2016).
153. Guintivano, J., Arad, M., Gould, T. D., Payne, J. L. & Kaminsky, Z. A. Antenatal prediction of postpartum depression with blood DNA methylation biomarkers. *Mol. Psychiatry* **19**, 560–567 (2014).
154. Boks, M. P. et al. SKA2 methylation is involved in cortisol stress reactivity and predicts the development of post-traumatic stress disorder (PTSD) after military deployment. *Neuropsychopharmacology* **41**, 1350–1356 (2016).
155. Kaminsky, Z. et al. A multi-tissue analysis identifies HLA complex group 9 gene methylation differences in bipolar disorder. *Mol. Psychiatry* **17**, 728–740 (2012).
156. Howsmon, D. P., Kruger, U., Melnyk, S., James, S. J. & Hahn, J. Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation. *PLoS Comput. Biol.* **13**, e1005385 (2017).
157. Ju, C. et al. Integrated genome-wide methylation and expression analyses reveal functional predictors of response to antidepressants. *Transl. Psychiatry* **9**, 1–12 (2019).
158. Kuhn, M. & Johnson, K. *Feature Engineering and Selection: a Practical Approach for Predictive Models* (CRC Press, 2019).
159. Zhang, Y., Florath, I., Saum, K. U. & Brenner, H. Self-reported smoking, serum cotinine, and blood DNA methylation. *Environ. Res.* **146**, 395–403 (2016).
160. Rhead, B. et al. Rheumatoid arthritis naive T cells share hypermethylation sites with synoviocytes. *Arthritis Rheumatol.* **69**, 550–559 (2017).
161. Ligthart, S. et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol.* **17**, 255 (2016).
162. Shi, J. et al. Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genet.* **12**, e100649 (2016).
163. Kundu, S. AI in medicine must be explainable. *Nat. Med.* **27**, 1328 (2021).
164. Dye, C. K. et al. Comparative DNA methylomic analyses reveal potential origins of novel epigenetic biomarkers of insulin resistance in monocytes from virally suppressed HIV-infected adults. *Clin. Epigenetics* **11**, 95 (2019).

165. Shen, F. et al. Identification of CD28 and PTEN as novel prognostic markers for cervical cancer. *J. Cell. Physiol.* **234**, 7004–7011 (2019).
166. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning* (Springer, 2013). **This is a standard introductory text to machine learning modelling with some level of mathematical background required and applied programming tutorials**.
167. Hattab, M. W., Clark, S. L. & van den Oord, E. J. C. G. Overestimation of the classification accuracy of a biomarker for assessing heavy alcohol use. *Mol. Psychiatry* **23**, 2114–2115 (2018). **This letter identifies and clearly articulates the issue of data leakage that impacted the approach and inflated the performance statistics of several early DNAm predictors, particularly those developed from large EWAS meta-analyses**.
168. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
169. Cohen, J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**, 213–220 (1968).
170. Jurman, G., Riccadonna, S. & Furlanello, C. A comparison of MCC and CEN error measures in multi-class prediction. *PLoS ONE* **7**, 41882 (2012).
171. Simpkin, A. J., Suderman, M. & Howe, L. D. Epigenetic clocks for gestational age: statistical and study design considerations. *Clin. Epigenetics* **9**, 100 (2017).
172. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, 9 (2019).
173. Chen, I. Y. et al. Ethical machine learning in health care. *Annu. Rev. Biomed. Data Sci.* **4**, 123–144 (2021). **This review identifies the many different ways that uncritical development of prediction models of health characteristics can entrench and exacerbate disparities for vulnerable populations**.
174. Mitchell, M. et al. Model cards for model reporting. In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* 220–229 (ACM, 2018).
175. Thomas, R. & Uminsky, D. The problem with metrics is a fundamental problem for AI. *arXiv*, doi:arxiv.org/abs/2002.08512 (2020).
176. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition* (Springer Science & Business Media, 2009). **This is a canonical text on theoretical and applied machine learning with detailed introductions to linear modelling, many common supervised and unsupervised learning methods, and design considerations for prediction modelling**.
177. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **23**, 40–55 (2022).
178. Bottner, A. et al. Gender differences of adiponectin levels develop during the progression of puberty and are related to serum androgen levels. *J. Clin. Endocrinol. Metab.* **89**, 4053–4061 (2004).
179. Riley, R. D. et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019). **This is an exploration of the key constraints that affect power and sample size in machine learning and prediction settings for binary and time-to-event outcomes**.
180. Riley, R. D. et al. Minimum sample size for developing a multivariable prediction model: part I – continuous outcomes. *Stat. Med.* **38**, 1262–1275 (2019). **This is an exploration of the key constraints that affect power and sample size in machine learning and prediction settings for continuous outcomes**.
181. National Human Genome Research Institute. DNA sequencing costs: data. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (2021).
182. Shafi, A., Mitrea, C., Nguyen, T. & Draghici, S. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Brief. Bioinform.* **19**, 737–753 (2018).
183. Zhang, L. et al. DNA methylation landscape reflects the spatial organization of chromatin in different cells. *Biophys. J.* **113**, 1395–1404 (2017).
184. Lin, N. et al. Genome-wide DNA methylation profiling in human breast tissue by Illumina TruSeq methyl capture EPIC sequencing and infinium methylationEPIC beadchip microarray. *Epigenetics* **16**, 754–769 (2021).
185. Wendt, J., Rosenbaum, H., Richmond, T. A., Jeddeloh, J. A. & Burgess, D. L. Targeted bisulfite sequencing using the SeqCap Epi enrichment system. *Methods Mol. Biol.* **1708**, 383–405 (2018).
186. Liu, Y. et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol.* **22**, 295 (2021).
187. Sakamoto, Y. et al. Long-read whole-genome methylation patterning using enzymatic base conversion and nanopore sequencing. *Nucleic Acids Res.* **49**, e81 (2021). **This study highlights the use of long-read sequencing of DNAm levels without bisulfite conversion**.
188. Shi, J. et al. The concurrence of DNA methylation and demethylation is associated with transcription regulation. *Nat. Commun.* **12**, 5285 (2021).
189. Pinu, F. R., Goldansaz, S. A. & Jaine, J. Translational metabolomics: current challenges and future opportunities. *Metabolites* **9**, 108 (2019).
190. Ignjatovic, V. et al. Mass spectrometry-based plasma proteomics: considerations from sample collection to achieving translational data. *J. Proteome Res.* **18**, 4085–4097 (2019).
191. Shah, S. et al. Improving phenotypic prediction by combining genetic and epigenetic associations. *Am. J. Hum. Genet.* **97**, 75–85 (2015). **This study demonstrates the additive explanatory power of combining polygenic and DNAm-based complex trait prediction, with greater benefit observed when adding DNAm information for traits with greater environmental components**.
192. Shah, S. et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.* **24**, 1725–1733 (2014).
193. Trejo Banos, D. et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nat. Commun.* **11**, 2865 (2020).
194. Zhang, F. et al. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol.* **20**, 107 (2019).
195. Ebrahim, A. et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nat. Commun.* **7**, 13091 (2016).
196. Argelaguet, R. et al. Multi-Omics Factor Analysis — a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
197. Woo, H. G. et al. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nat. Commun.* **8**, 839 (2017).
198. Zhu, B. et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Sci. Rep.* **7**, 16954 (2017).
199. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
200. Gadd, D. A. et al. Epigenetic scores for the circulating proteome as tools for disease prediction. *eLife* **11**, e71802 (2022). **This study highlights the potential of DNAm to index endogenous biomarkers and thus enhance prediction of phenotypes or diseases associated with these biomarkers**.
201. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* **13**, 1 (2015). **This paper details consensus recommendations of best practices for reporting prediction modelling results as developed by an international expert pannel**.
202. Moons, K. G. M., Royston, P., Vergouwe, Y., Grobbee, D. E. & Altman, D. G. Prognosis and prognostic research: what, why, and how? *BMJ* **338**, 1317–1320 (2009).
203. Weber, L. M. et al. Essential guidelines for computational method benchmarking. *Genome Biol.* **20**, 125 (2019).
204. Shmueli, G. To explain or to predict? *Stat. Sci.* **25**, 289–310 (2010). **This paper provides an accessible explanation of the distinctions between explanatory and predictive statistics in terms of aims and methodologies, as well as perspective on why such differences have been persistently confused across fields**.
205. Murray, R. P., Connett, J. E., Lauger, G. G. & Voelker, H. T. Error in smoking measures: effects of intervention on relations of cotinine and carbon monoxide to self-reported smoking. *Am. J. Public Health* **83**, 1251 (1993).
206. Rehm, J. & Spuhler, T. Measurement error in alcohol consumption: the Swiss Health Survey. *Eur. J. Clin. Nutr.* **47** (Suppl. 2), S25–S30 (1993).
207. Subar, A. F. et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. *Am. J. Epidemiol.* **158**, 1–13 (2003).
208. Adab, P., Pallan, M. & Whincup, P. H. Is BMI the best measure of obesity? *BMJ* **360**, k1274 (2018).
209. Greenland, S., Pearl, J. & Robins, J. M. Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48 (1999).
210. Peters, J., Janzing, D. & Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms* (MIT Press, 2017).
211. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. The Morgan Kaufmann Series in Representation and Reasoning (Morgan Kaufmann, 1988).
212. Piccininni, M., Konigorski, S., Rohmann, J. L. & Kurth, T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC Med. Res. Methodol.* **20**, 179 (2020).
213. Austin, P. C. & Steyerberg, E. W. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat. Med.* **38**, 4051–4065 (2019).
214. Korologou-Linden, R., Leyden, G. M., Relton, C. L., Richmond, R. C. & Richardson, T. G. Multi-omics analyses of cognitive traits and psychiatric disorders highlights brain-dependent mechanisms. *Hum. Mol. Genet.* https://doi.org/10.1093/hmg/ddab016 (2021).
215. Tsai, P. C. et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. *Clin. Epigenetics* **10**, 126 (2018).
216. Smith, A. K. et al. DNA extracted from saliva for methylation studies of psychiatric traits: evidence tissue specificity and relatedness to brain. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **168**, 36–44 (2015).
217. Braun, P. R. et al. Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *Transl. Psychiatry* **9**, 47 (2019).
218. Nagy, C. et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).

Acknowledgements

The authors thank G. Hemani for helpful discussions on genetic prediction and K. Tilling for comments on a draft manuscript. The authors' work is supported by the Medical Research Council Integrative Epidemiology Unit at the University of Bristol (MC_UU_00011/1 & 5) and via the Cancer Research UK programme grant (C18281/A29019). The authors' work is also supported by the NIHR Biomedical Research Centre at University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Author contributions

P.D.Y., M.S., R.L., O.W. researched the literature. P.D.Y., M.S. and C.L.R. contributed substantially to discussions of the content. P.D.Y., M.S., R.L. wrote the article. P.D.Y., M.S., G.D.S and C.L.R. reviewed and/or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Genetics thanks Christopher Bell, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.