Machine learning in the identification of prognostic DNA methylation biomarkers among patients with cancer: a systematic review of epigenome-wide studies

Tanwei Yuan\textsuperscript{1,2}, Dominic Edelmann\textsuperscript{3}, Ziwen Fan\textsuperscript{1}, Elizabeth Alwers\textsuperscript{1}, Jakob Nikolas Kather\textsuperscript{4,5,6,7}, Hermann Brenner\textsuperscript{1,8,9} and Michael Hoffmeister\textsuperscript{1*}

\textsuperscript{1}Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.

\textsuperscript{2}Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany.

\textsuperscript{3}Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany.

\textsuperscript{4}Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK.

\textsuperscript{5}Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany.

\textsuperscript{6}Gastrounit-Surgical Division, Center for Surgical Research, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark.

\textsuperscript{7}Medical Oncology, National Center of Tumour Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany.

\textsuperscript{8}Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany.

\textsuperscript{9}German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.

*Correspondence to:

Dr. Michael Hoffmeister

Division of Clinical Epidemiology and Aging Research, Molecular Pathological Epidemiology Group, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 581, Heidelberg, 69120, Germany

mhoffmeister@dkfz.de, Phone: +49 6221 42-1303

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Summary

Background: DNA methylation biomarkers have great potential in improving prognostic classification systems for patients with cancer. Machine learning (ML)-based analytic techniques might help overcome the challenges of analyzing high-dimensional data in relatively small sample sizes. This systematic review summarizes the current use of ML-based methods in epigenome-wide studies for the identification of DNA methylation signatures associated with cancer prognosis.

Methods: We searched three electronic databases including PubMed, EMBASE, and Web of Science for articles published until 8 June 2022. ML-based methods and workflows used to identify DNA methylation signatures associated with cancer prognosis were extracted and summarized. Two authors independently assessed the methodological quality of included studies by a seven-item checklist adapted from relevant guidelines.

Results: Seventy-six studies were included in this review. Three major types of ML-based workflows were identified: 1) unsupervised clustering, 2) supervised feature selection, and 3) deep learning-based feature transformation. For the three workflows, the most frequently used ML techniques were consensus clustering, least absolute shrinkage and selection operator (LASSO), and autoencoder, respectively. The systematic review revealed that the performance of these approaches has not
been adequately evaluated yet and that methodological and reporting flaws were common in the identified studies using ML techniques.

**Conclusions:** There is great heterogeneity in ML-based methodological strategies used by epigenome-wide studies to identify DNA methylation markers associated with cancer prognosis. Benchmarking studies are needed to compare the relative performance of various approaches for specific cancer types. Adherence to relevant methodological and reporting guidelines is urgently needed.

**Keywords:** DNA methylation, epigenetic biomarkers, epigenome-wide studies, cancer prognosis, artificial intelligence, machine learning, systematic review
Background

Cancer remains a leading cause of death and is projected to become even more frequent due to population aging and growth in nearly every country of the world (1). Accurate prediction of prognosis for patients with cancer is pivotal for individualized treatment and reduction in mortality. So far, the tumor-lymph node-metastasis (TNM) staging system has been the most commonly used system for treatment decisions and to predict the prognosis of cancer (2-4). However, the role of molecular features is becoming increasingly important for a more precise treatment and prediction of cancer prognosis (3, 4).

DNA methylation involves the addition of a methyl group to the C5 position of the cytosine to form 5-methylcytosine (5). It is one of the most common epigenetic changes that regulate gene expression and plays a key role in carcinogenesis, cancer development, and clinical prognosis (5). DNA methylation signatures hold great promise for improving the prognostic accuracy of different cancers (6-9). Recent advances in DNA methylation microarray platforms enable studies to analyze methylation across the whole genome in a high-throughput manner (10, 11). However, the sample size of these epigenome-wide studies is often less than 1000 patients, much smaller relative to the number of CpG sites investigated (27,578-
Such analyses are prone to the issues of overfitting and multicollinearity when using traditional statistical methods, most of which are intended for the low-dimensional setting (12).

Machine learning (ML) could be a powerful tool to properly handle high-dimensional DNA methylation data. ML is a subset of artificial intelligence used in data analysis, including algorithms that are trained to automatically recognize data representations, learn from experience, and maximize predictive accuracy (13). A branch of machine learning is deep learning, which is based on artificial neural networks (13). ML methods are capable of analyzing high-dimensional, non-parametric data with complex interactions (13, 14). Although these methods are mainly used for analysis of text and images (13), recent years saw a growing body of epigenome-wide studies that leveraged the power of ML to identify DNA methylation signatures associated with cancer prognosis (Supplementary Figure 1). But thus far no appropriate review has been published to summarize this recent emerging trend. However, the types of ML methods selected and workflows used in different studies are quite heterogeneous, and there might be powerful but yet underused ML methods. Therefore, we conducted a systematic review to comprehensively map studies investigating epigenome-wide associations with cancer prognosis using ML, as well as to identify existing limitations and research gaps.
Methods

This systematic review was performed and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The complete PRISMA checklist is shown in Supplementary Table 1. The review protocol was published in advance at


Search strategy

We searched PubMed, EMBASE, and Web of Science for studies published until 8 June 2022. The search strategy used a combination of Medical Subject Headings (MESH) and key words related to ML, cancer, epigenetic signatures, and prognosis, respectively. The full search strategy is provided in Supplementary Table 2.

Reference lists of included studies were also screened for additional eligible studies. The records retrieved from databases were exported to Endnote 20 (Clarivate Analytics, Philadelphia, USA) and duplicates were removed. Subsequently, the titles and abstracts of the remaining records were screened for eligibility by one author (TY), and potentially eligible studies were retained for full-text review.

Selection criteria

Studies were deemed eligible if they were peer-reviewed articles, reported in English, included patients with any type of cancer, used an epigenome-wide DNA methylation array, and used at least one ML method to identify DNA methylation signatures (i.e., CpG sites or methylation-driven genes [MDGs]) associated with cancer prognosis.
(i.e., survival, progression, therapy responses). Multi-omics studies investigating multiple prognostically relevant biomarkers including DNA methylation were also considered. We excluded studies that recruited patients with cancer precursors, used a candidate-gene approach, used ML methods for purposes unrelated to finding prognostically relevant DNA methylation signatures, or simply proposed a customized statistical package or platform based on ML methods.

Quality assessment

Two authors (TY and ZF) independently assessed the methodological quality of the included studies. The methodological quality of the included studies was assessed by a seven-item checklist adapted from 'A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies (PROBAST) (15)' and from the 'Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK) (16)'.

A study was assigned one point each for meeting the following criteria: total follow-up time >3 years; sample size of training set >100; use of an imputation method for missing data; adjustment for confounders including age, sex, and tumor stage when selecting or evaluating DNA methylation signatures; use of cross-validation or bootstrapping to measure model performance. Additionally, a study could score one point for conducting internal validation, and two points for doing external validation. One point was awarded for measuring the discriminative ability (i.e., the ability of a model to differentiate between high-risk patients and low-risk patients) of selected
signatures, and two points in case of additionally measuring calibration accuracy of selected signatures (i.e., the degree of agreement between the predicted and observed prognostic outcomes) of selected signatures.

The quality of each study was rated as low, medium, or high if the score was 0-3, 4-6, or 7-9 points, respectively. Discrepancies between the evaluations of the two authors (TY, ZF) were solved through discussion, and two senior scientists (MH, DE) made the final decision when a consensus was not reached.

**Data extraction and synthesis**

The following study-level information was extracted onto pre-designed spreadsheets: first author, year of publication, geographic origin of training set, median follow-up, anatomical site of cancer, cancer stage, prognostic outcome, biospecimen type, methylation array type, methylation signature type, sample size and source of training set, sample size of testing set, type of internal validation, size and source of external validation set, handling of missing values, prognostic model construction, performance measurement, types of ML methods used, and workflows. We only extracted relevant ML methods and workflows that were used to identify individual DNA methylation signatures associated with cancer prognosis, and disregarded methods used for other purposes. The extracted workflows used in each study were
cross-checked by a second author (ZF) to ensure accuracy, and disagreements were resolved through discussion between the two authors.

The main study characteristics were summarized in tables, and we summarized the different ML methods by a sunburst chart and a bubble chart, respectively. Workflows used in included studies were summarized by Sankey diagrams. These were performed by using R version 4.1.2 with packages tidyverse, ggplot2, and highcharter (17).

Results

Study characteristics

We identified 76 studies that were eligible for this systematic review (Figure 1) (5-9, 18-88).

Included studies were published between 2002 and 2022, and there was an exponential growth in the number of studies published from 2019 to 2021 (Supplementary Figure 1). Characteristics of included studies are summarized in Table 1, and details of each study are available in Supplementary Table 3.

Lung cancer was the most studied cancer type ($N = 16$), followed by colorectal cancer ($N = 11$). The majority of studies recruited patients with stage I-IV cancer ($N = 50$), and overall survival was the most frequently investigated outcome ($N = 52$). Only
six studies based on clinical cohorts reported median follow-up time, ranging from 1.5 to 4.6 years. Nearly all studies determined DNA methylation in tumor tissue ($N = 73$) and used the Illumina 450K BeadChip array ($N = 63$). Fifty-seven studies investigated the role of differentially methylated single CpG sites on cancer prognosis, while 26 studies investigated the prognostic value of MDGs.

In terms of methodological characteristics, more than two thirds of the studies used The Cancer Genome Atlas Program (TCGA) database to identify prognostically relevant DNA methylation biomarkers. The size of the training sets varied from 19 to 890 patients (median: 213). Sample-splitting was the most frequently used internal validation ($N = 42$), and the size of test sets ranged from 33 to 342 patients (median: 143). Twenty-one studies performed external validation, mostly based on the Gene Expression Omnibus (GEO) database. The size of the datasets used for external validation ranged from 12 to 583 (median: 125). Thirty three studies reported imputation of missing values, and K-nearest neighbor was the most frequently used imputation method ($N = 26$). More than half of included studies constructed prognostic models incorporating identified DNA methylation biomarkers.

**Quality assessment**

Based on our seven-item quality assessment checklist, 61% of the studies were rated as moderate quality ($N = 46$), and 12 (16%) studies were rated as low quality. The
summarized results for each quality assessment item are shown in Figure 2, and the detailed results for each included study are available in Supplementary Table 4.

The top three items where most studies failed to score points were use of cross-validation or bootstrapping to measure model performance ($N = 55$), use of an imputation method for missing data ($N = 43$), and adjustment for important confounders like age, sex, and tumor stage when selecting or evaluating methylation markers ($N = 33$), respectively.

**ML methods**

The complete spectrum of ML methods used in the included studies to identify DNA methylation biomarkers associated with cancer prognosis is shown in Figure 3a. The number of ML methods used by each study ranged from one to five, and the majority of studies used one ($N = 41$) or two ($N = 27$) types of ML methods. The time trend of the frequency of these ML methods used is illustrated in a bubble plot (Figure 3b).

Supervised machine learning algorithms and unsupervised clustering were predominantly used by a total of 71 and 46 studies, respectively. The top three most frequently used supervised machine learning algorithms were the Least Absolute Shrinkage and Selection Operator (LASSO, $N = 32$), support vector machine ($N = 11$), and random forest ($N = 10$). The most frequently used unsupervised clustering technique was consensus clustering ($N = 31$), followed by hierarchical clustering ($N =$
8) and K-means clustering ($N = 6$). The application of deep learning is a rather recent trend starting from 2020 (Figure 3b), with autoencoder being the most frequently used ($N = 5$).

**Workflows**

Based on the characteristic stage, we identified three different workflows: 1) unsupervised clustering, 2) supervised feature selection, and 3) deep learning-based feature transformation. These are visualized in the form of Sankey diagrams (Figure 4). Workflows used by each study are detailed in Supplementary Table 5-7.

**Workflow 1: unsupervised clustering**

Among the 39 studies that adopted clustering-based workflows (5, 18-23, 29, 32-35, 37-39, 43, 45-47, 50, 61, 62, 64, 69, 70, 75, 77-80, 83, 84, 87, 88), most started with a preliminary screening of CpG sites (Figure 4.1). The most frequently used pre-selection approach was selecting CpG sites that were significantly associated with prognosis by first performing univariable Cox regression analysis, and then performing multivariable Cox analysis with adjustment for clinical variables. Other pre-screening approaches included the selection of CpG sites above a pre-defined variance threshold, and the selection of significantly differentially methylated CpG sites (dmCpGs) between normal tissue and tumor tissue.
Afterwards, study samples underwent the process of clustering according to their methylation levels of the pre-selected CpG sites, and samples with similar methylation patterns were clustered together. Kaplan-Meier curves and log-rank tests were subsequently used to verify that cancer prognosis across different patient clusters was significantly different. At the third stage, a reduced number of cluster-specific CpG sites were identified, mostly by selecting dmCpG sites across clusters, or by selecting dmCpG sites in the seed cluster (i.e., the cluster associated with good prognosis and containing a great number of dmCpGs). Four studies (50, 56, 62, 79) used weighted gene co-expression network analysis to first identify the co-expression modules having the greatest correlation with the seed cluster, and then to screen for CpGs in that module mostly correlated with the seed cluster.

Based on selected CpGs in the prior stage, eight studies performed additional feature selection, using methods including LASSO (38, 45, 55, 57, 69, 78, 80), stepwise selection (38, 39), and the Akaike information criterion (64, 77), to further refine the number of CpGs. In contrast, four studies (23, 33, 34)(47) directly trained classification models with clusters as labels using Bayesian networks or random forest, while 17 studies constructed prognostic models (37-39, 45, 46, 53-57, 64, 69, 70, 75, 77, 78, 80). Lastly, models developed in the previous stage were mostly internally validated by split-sample validation, albeit four studies validated the model externally in an independent dataset (33, 50, 57, 78).

Workflow 2: supervised feature selection
Similar to clustering-based workflows, an initial screening was generally the first step for the 30 studies using supervised feature selection-based workflows (Figure 4.2) (6-9, 24-28, 30, 31, 36, 40-42, 48, 51, 59, 60, 63, 66-68, 71-74, 81, 84, 85). In the next step, nearly all studies used one to three types of supervised feature selection techniques to pre-select prognostically relevant DNA methylation biomarkers prior to training one or more prognostic models. One study (85) used feature selection (random forest) and dimensional reduction (principle component analysis) in parallel before training multiple supervised ML models.

In the stage of model training, further feature selection was performed by more than half of the 30 studies (8, 9, 24, 25, 28, 31, 36, 40-42, 48, 51, 66, 68, 72-74, 81, 84, 85). Sample-splitting was the most commonly used internal model validation method (N = 12), and external validation was performed in 15 studies (7, 8, 24-26, 31, 42, 48, 59, 66-68, 72, 74, 81).

As to studies using more than one supervised ML model for feature selection, four studies intersected CpGs selected by different ML techniques (9, 31, 48, 84), whereas two studies only selected the biomarkers retained in one model with the best performance (72, 85). Notably, one study (81) first identified methylation-correlated blocks (MCBs, i.e., blocks of closely co-methylated CpGs) and then selected MCBs associated with survival by LASSO. Next, elastic-net model, support vector regression, and Cox regression were separately trained and used for ranking...
MCBs. Finally, a stacking ensemble model combining the three models was constructed based on the top-ranked MCB.

**Workflow 3: deep learning-based feature transformation**

We identified a total of seven studies that applied deep learning-based feature transformation to reduce dimensionality (Figure 4.3). In the first step, five studies used autoencoder (44, 49, 52, 58, 76), and one used bidirectional deep neural networks (65), to transform original input features into a reduced number of new features (i.e., hidden nodes), and then they performed univariable Cox regression analysis to select hidden nodes significantly associated with prognosis. Subsequently, in the five studies (44, 49, 58, 65, 76) that integrated the information of multi-omics biomarkers including DNA methylation, the selected prognosis-associated hidden nodes were used to cluster samples by K-means clustering algorithm. After that, top 20 to 100 significantly differently methylated genes were identified by comparing across clusters. Based on the top features of multi-omics biomarkers, classification models with clusters as labels were constructed by support vector machine or random forest. Cross-validation was used in deep learning-based workflows by five studies, and two studies performed external validation containing DNA methylation information (44, 65).

In contrast, for the one study (52) that exclusively analyzed DNA methylation data, the third step was to calculate the weight of CpGs in every selected hidden node, and
then summed the absolute weight values of CpGs corresponding to the same gene. Subsequently, for every significant node, the ratio of weights for every gene was calculated. Lastly, they selected the intersection between the genes with the highest ratio of weights in all the nodes.

In particular, one study (82) applied one-dimensional discrete wavelet transform to pre-selected key genes, and the transformed results were used as the input for the downward convolutional neural network for feature extraction. Afterwards, the top 200 genes were further refined by univariable Cox regression analysis, which were finally used to construct the prognostic model by multivariable Cox regression analysis.

**Performance evaluation**

The performance evaluation reported by the included studies varied in terms of objects, metrics, and types of validation (Supplementary Table 8). The objects of evaluation were methylation level-based clusters generated by unsupervised clustering, prognostic models constructed by only incorporating identified DNA methylation biomarkers, and prognostic models based on both DNA methylation markers and clinical variables. The most frequently reported metrics were results of Cox regression analysis \((N = 55)\), which were used to evaluate whether there were statistically significant differences in prognosis across clusters, or among patients with different prognostic index. Fewer studies used other metrics, including C-index.
(N = 16) and receiver operating characteristic curves (N = 39) to evaluate the
discriminative power of their prognostic models. Only 18 studies (6, 31, 36, 39-42, 51,
54, 55, 57, 59, 66, 67, 78, 80, 84, 88) evaluated the calibration accuracy of
prognostic models.

Given the large heterogeneity across studies in terms of cancer types, prognostic
outcomes, methodological strategies, and evaluation metrics, we did not summarize
reported metrics. Nevertheless, six studies reported an increase in the discriminative
ability of the methylation-based prognostic model after adding other clinical variables
(e.g., age, cancer stage) to the model (6, 31, 39, 51, 76, 80).

Discussion

This systematic review is, to the best of our knowledge, the first review to date that
summarizes and catalogs the current use of ML-powered analytic methods in
epigenome-wide studies to identify DNA methylation markers associated with cancer
prognosis. Following a systematic approach, we reviewed 76 peer-reviewed articles
and found pronounced heterogeneity in methodological strategies. Workflows used in
prior work were categorized into 1) unsupervised clustering, 2) supervised feature
selection, and 3) deep learning-based feature transformation according to the
characterized ML techniques used. We also identified several common
methodological and reporting issues. This review could potentially serve as a
reference for researchers designing ML-based analytic approaches to find meaningful biomarkers, as well as identifying areas warranting further research.

We found that common to all identified workflows, a combination of different dimensionality reduction strategies, with or without the involvement of ML, was applied in more than one stage. These processes are necessary for high-dimensional microarray data, which typically contain many irrelevant and redundant features. Applying various dimensionality reduction techniques in different stages could improve efficiency by narrowing down the selection scope and reducing computational complexity (89).

Clustering and supervised feature selection were the two main ML-based methods used in the included studies to select a subset of relevant DNA methylation signatures. Consensus clustering, an ensemble-based method of aggregating results from multiple clustering algorithms, was used most frequently, probably because of its relatively robust results compared to other individual clustering algorithms (39). The most popular supervised feature selection method was LASSO, a regularized regression method that could perform variable selection while constraining overfitting (90). We also identified a recent increase in the use of deep learning methods to compress and extract input features, with autoencoder being the most commonly used.
The popularity of a method is not necessarily equivalent to the optimal or best model selection and performance. An objective evaluation of the efficacy of existing methods and workflows could only be achieved by using the same dataset and benchmarking all approaches against each other through direct comparisons (91). A few identified studies performed comparative analysis of the autoencoder framework with other methods (49, 52, 65, 76), and results mostly favored the autoencoder approach. Besides, one study showed that wavelet-based deep learning model outperform traditional LASSO and other wavelet-based approaches (82). However, we did not find benchmarking studies comparing the three types of workflows identified by this review.

For supervised feature selection and unsupervised approaches, each has its own advantages and disadvantages. Specifically, supervised learning uses labeled data in which the targeted variable to be predicted is known. However, it is costly and time-consuming to follow up all patients in large cohorts or databases for a reasonably long time to obtain information about the targeted outcome (89). In comparison, unsupervised learning does not require outcome information and may thus enable analyses in much larger patient cohorts. But it has the downside that the results are not necessarily related to a meaningful clinical outcome (89). As a remedy semi-supervised approaches might be used, which could combine the best of both approaches. Alternatively, studies might use unsupervised and supervised
approaches in conjunction at different stages to select features, as done in some of
the studies (38, 45, 55, 57, 69, 78, 80).

Many methodological and reporting problems were common in the identified studies. First, only five out of the 69 included studies reported median follow-up time (6, 24, 26, 29, 31), which is important information because survival outcomes can only be
decently evaluated after several years of follow-up, especially if patients with early
stages of cancer are included. None of the studies using TCGA data reported median
follow-up time, so little can be said about the data quality and interpretability of the
conducted study. Second, six studies failed to specify the type of survival outcomes
(e.g., overall survival or disease-free survival) (43, 46, 47, 60, 62, 75), which further
limits the interpretation of results.

Third, handling of missing values was not specified in nearly half of the studies, in
which patients with any missing data were likely to be omitted. Six studies explicitly
reported exclusion of patients with missing values (6, 44, 46, 52, 57, 76). Simply
excluding patients with any missing data from the analysis could lead to selection
bias, and proper imputation methods ought to be in place (15, 92).

Fourth, the association between aberrant DNA methylation signatures and cancer
prognosis is only relevant when they provide additional prognostic value not captured
by known prognostic factors such as tumor stage, age and sex. But more than half of
the identified studies did not take these important clinical factors into account.
Fifth, none of the studies mentioned how competing risk events were dealt with in the analysis when the outcome of interest was not overall survival. Sixth, most studies only used one single supervised learning model for variable selection without providing a rationale why choosing that specific algorithm. The reasons for selecting a modeling method should be clearly stated (92). There might have been room for improvement in model performance if a wider range of models had been tried and compared in these studies.

Seventh, the aim of internal validation is to evaluate the performance of a model in a dataset not involved in model training (92). However, one study randomly selected a subset from the training set to validate the model (63), which is not a real internal validation. Such an approach could lead to data leakage and thereby to an optimistically biased evaluation (92).

Most studies randomly split a single dataset into training and test sets. This approach fails to efficiently use all data available for model training, and the model performance can change every run when the dataset is randomly split in different ways (15, 92). Better internal validation methods such as cross-validation and bootstrapping are recommended by guidelines (15, 92). It is also recommended that all analysis steps (imputation, screening etc.) should be incorporated in the cross-validation procedures to prevent data leakage (92). Moreover, external validation is crucial to verify the
generalizability of findings, which is especially important for high-dimensional data with high risk of overfitting.

Eighth, discrimination and calibration are two important metrics to evaluate the performance of prognostic models incorporating meaningful DNA methylation biomarkers (15, 92). However, most studies reviewed focused on discrimination metrics, whereas calibration accuracy was only measured by less than one quarter of included studies. Both discrimination and calibration evaluation metrics are required for all datasets including training, testing, and external validation sets (92). However, 5 studies reported discrimination metrics in both the testing set and external validation set, but reported calibration only in the training set (42, 48, 67, 68, 78).

Lastly, we found that 18 identified studies first constructed a Cox prognostic model based on single selected CpG sites (i.e., methylation score), which was subsequently used as a single predictor, together with other clinical variables, to construct a second Cox prognostic model (6, 31, 36, 39-42, 55, 57, 66-69, 73, 78, 80, 81). It might be an inappropriate approach to construct the first "methylation score", in which the relative weight of individual CpGs were consequently fixed for the second prognostic model. Moreover, four studies dichotomized the continuous methylation score (31, 41, 55, 59), leading to loss of information (15). Given that clinical information and DNA methylation biomarkers are different types of information characterized by different statistical properties, they could be combined by
multimodal deep learning technique. Additionally, traditional nomograms were used by all the 18 studies for the graphical presentation of their prediction model. More user-friendly online risk calculator platforms are recommended over nomograms in today’s digital era (92, 93).

Limitation of the study

This systematic review is the first of its kind and has some limitations. First, we specifically included only studies using a genome-wide array to identify DNA methylation signatures associated with cancer prognosis. We recognize that similar ML approaches can be applied to non-genome-wide studies or to the identification of other biomarkers (e.g., mRNA) relevant to other diseases. Second, in the included multi-omics studies (44, 49, 58, 65, 76), identifying individual prognostically relevant DNA methylation biomarkers were merely side products instead of primary study aims. Third, variations in disease outcomes and the lack of benchmarking studies prevented us from comparing the performance of identified workflows.

Further benchmarking studies are needed to systematically evaluate the comparative performance of various ML-based methods and workflows identified in this review and to identify the best approaches. Novel ML methods not included in this review may also help to reduce dimensionality and to find prognostically relevant biomarkers. Future epigenome-wide association studies intending to use ML
methods should carefully adhere to relevant methodological and reporting standards (15, 92, 94), so as to improve methodological quality and reproducibility.

Conclusions

wide association studies. We identified three major categories of ML-based workflows as well as some common methodological and reporting flaws in existing studies. Benchmarking studies are needed to compare the relative performance of these workflows using a same dataset. Adherence to methodological and reporting guidelines is strongly recommended for future research in this area.

Abbreviations

ML: machine learning; LASSO: least absolute shrinkage and selection operator; TNM: tumor-lymph node-metastasis; PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses; MESH: Medical Subject Headings; MDGs: methylation-driven genes; PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies; RMARK: Reporting Recommendations for Tumor Marker Prognostic Studies; TCGA = The Cancer Genome Atlas Program; GEO: Gene Expression Omnibus; dmCpGs: differentially methylated CpG sites;

Ethics approval and consent to participate

Not applicable.
Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the supplementary materials of the article.

Conflict of interest statement

The authors declare that they have no competing interests.

Funding

Funding support to complete the review was received from the International PhD Program Office, German Cancer Research Center (DKFZ), Heidelberg, Germany. A grant number does not apply. The funder did not have any role in the design of the study or in the explanation of the data.

Authors' contribution

MH conceived the study. MH and TY designed the protocol. DE provided statistical input and consultation. TY conducted study selection, and TY and ZF conducted data extraction and quality assessment. TY made tables and figures and drafted the manuscript. DE contributed to interpretation of findings. DE, ZF, EA, and JNK, and HB critically revised the manuscript. MH supervised the study.
Acknowledgements

We thank Xinyi Zhou for helping retrieve initial records from the EMBASE database.

References


28


Figures and legends
Figure 1. Selection of studies for inclusion

ML = machine learning
Figure 2. Quality assessment of included studies (N = 69) and distribution of assigned scores.

NR = Not reported
Figure 3. (a) Types and composition of ML methods used in included studies; (b) Frequency of each ML methods used by studies each year.

The numbers on the figure do not add up to 76 because they represent the frequency of methods being used, because studies could have used more than one method. In theory, deep learning belongs to machine learning and could be either supervised or unsupervised, but for the purpose of illustration, supervised ML on the figure refers to supervised machine learning algorithms excluding deep learning. The size of a bubble in (b) is proportional to the frequency of the AI method used (the number inside the bubble). ML = machine learning. LASSO = least absolute shrinkage and selection operator.
Figure 4 Workflows used by included studies

At each stage, one study might use more than one method. The number after each method represents the frequency of its being used. ‘Uni’ refers to performing feature selection depending on the P-value of univariable analysis, and features with a P-value < 0.05 were retained; similarly, ‘Multi’ stands for feature selection based on the P-value of multivariable analysis; ‘Cox’ under the stage of model training refers to training a simple multivariable Cox regression model without further feature selection.


Figure 4.1 Workflow 1: unsupervised clustering
Figure 4.2 Workflow 2: supervised feature selection

Figure 4.3 Workflow 3: deep learning-based feature transformation

Table 1. Study and methodological characteristics of included studies.
<table>
<thead>
<tr>
<th>Study characteristics</th>
<th>Studies (N)</th>
<th>Methodological characteristics</th>
<th>Studies (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Geographic origin of training set</strong></td>
<td></td>
<td><strong>Size of training set</strong></td>
<td></td>
</tr>
<tr>
<td>Multi</td>
<td>65</td>
<td>19-100</td>
<td>11</td>
</tr>
<tr>
<td>China</td>
<td>2</td>
<td>100-200</td>
<td>22</td>
</tr>
<tr>
<td>Japan</td>
<td>2</td>
<td>200-890</td>
<td>44</td>
</tr>
<tr>
<td>Norway</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Median follow-up (yr)</strong></td>
<td></td>
<td><strong>Source of training set</strong></td>
<td></td>
</tr>
<tr>
<td>1-3</td>
<td>2</td>
<td>TCGA</td>
<td>60</td>
</tr>
<tr>
<td>3-5</td>
<td>4</td>
<td>Clinical cohort</td>
<td>16</td>
</tr>
<tr>
<td>Not reported</td>
<td>70</td>
<td>GEO</td>
<td>4</td>
</tr>
<tr>
<td><strong>Cancer type</strong></td>
<td></td>
<td><strong>Type of validation</strong></td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>16</td>
<td>Internal validation</td>
<td>46</td>
</tr>
<tr>
<td>Colon and rectum</td>
<td>11</td>
<td>Sample-splitting</td>
<td>42</td>
</tr>
<tr>
<td>Liver</td>
<td>9</td>
<td>Cross-validation</td>
<td>18</td>
</tr>
<tr>
<td>Breast</td>
<td>6</td>
<td>Bootstrapping</td>
<td>4</td>
</tr>
<tr>
<td>Stomach</td>
<td>5</td>
<td>External validation</td>
<td>21</td>
</tr>
<tr>
<td>Ovary</td>
<td>4</td>
<td>Not performed</td>
<td>15</td>
</tr>
<tr>
<td>Bladder</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kidney</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cancer stage</strong></td>
<td></td>
<td><strong>Size of testing set</strong></td>
<td></td>
</tr>
<tr>
<td>I-IV</td>
<td>50</td>
<td>33-100</td>
<td>10</td>
</tr>
<tr>
<td>I-III</td>
<td>3</td>
<td>100-200</td>
<td>24</td>
</tr>
<tr>
<td>Others</td>
<td>3</td>
<td>200-342</td>
<td>9</td>
</tr>
<tr>
<td>Not reported</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Outcome</strong></td>
<td></td>
<td><strong>Size of external validation set</strong></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>52</td>
<td>19-100</td>
<td>13</td>
</tr>
<tr>
<td>RFS</td>
<td>12</td>
<td>100-200</td>
<td>8</td>
</tr>
<tr>
<td>PFS</td>
<td>8</td>
<td>200-583</td>
<td>12</td>
</tr>
<tr>
<td>Unspecified survival</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response to therapy</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not reported</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Type of biospecimen</strong></td>
<td></td>
<td><strong>Source of external validation</strong></td>
<td></td>
</tr>
<tr>
<td>Tumor tissue</td>
<td>73</td>
<td>GEO</td>
<td>10</td>
</tr>
<tr>
<td>Others</td>
<td>3</td>
<td>TCGA</td>
<td>7</td>
</tr>
<tr>
<td><strong>Type of methylation array</strong></td>
<td></td>
<td><strong>Imputation of missing values</strong></td>
<td></td>
</tr>
<tr>
<td>Illumina HM 450K</td>
<td>63</td>
<td>Clinical cohort</td>
<td>4</td>
</tr>
<tr>
<td>Illumina HM 27K</td>
<td>18</td>
<td>Others</td>
<td>4</td>
</tr>
<tr>
<td>Others</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not reported</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Type of methylation</strong></td>
<td></td>
<td><strong>No. ML methods used</strong></td>
<td></td>
</tr>
<tr>
<td>CpGs</td>
<td>57</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>Genes</td>
<td>26</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Prognostic index construction</strong></td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Performance metrics</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discrimination</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cox regression</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C-index</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ROC curves</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Calibration</td>
<td>18</td>
</tr>
</tbody>
</table>

The number of studies under each study-level characteristic might not add up to 76 because some studies contained the characteristic of more than one type. TCGA = The Cancer Genome Atlas Program; GEO = Gene Expression Omnibus; OS = Overall Survival; DFS = Disease Free Survival;
RFS = Relapse Free Survival; PFS = Progression Free Survival; ROC = Receiver operating characteristic; ML = machine learning