

Artificial intelligence in the identification of prognostic DNA methylation biomarkers among patients with cancer: a scoping review of epigenome-wide studies

Tanwei Yuan

Division of Clinical Epidemiology and Aging Research, German Cancer Research Center

Michael Hoffmeister (✉ m.hoffmeister@Dkfz-Heidelberg.de)

Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ)

Method Article

Keywords: artificial intelligence, DNA methylation, cancer, prognosis

Posted Date: January 24th, 2022

DOI: <https://doi.org/10.21203/rs.3.pex-1738/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

DNA methylation signatures are becoming increasingly important in the perspective of precision oncology. Artificial intelligence (AI) represents a powerful tool that could properly analyze high-dimensional DNA methylation data. There has been a surge of epigenome-wide association studies that leverage the power of AI techniques to identify prognostically relevant DNA methylation biomarkers. But the methodological strategies used are quite heterogeneous. This scoping review will comprehensively summarize studies conducted in this direction. We will search PubMed, EMBASE, and Web of Science for studies published as of 30 August 2021 and identify eligible studies by inclusion criteria. We will extract study characteristics, assess methodological quality, and summarize AI methods and pipelines used by included studies. The potential limitation of this study would be the lack of well-established criteria to judge the validity of existing pipelines and to provide suggestions in terms of the selection of pipelines for future studies.

Introduction

Cancer remains a major and long-term public health challenge worldwide. In 2020, cancer accounted for almost 10 million deaths, making it a leading cause of death in nearly every country of the world. Meanwhile, the cancer burden is continuously growing, which is projected to reach 28.4 million cases in 2040, a 47% increase from 2020. This calls for unrelenting efforts and resources for cancer prevention, treatment, and disease management.

Accurate prognosis for patients with cancer is crucial for the planning of individualized treatment and management, which helps improve treatment outcomes and reduce mortality. So far, the tumor-lymph node-metastases (TNM) staging system is the most commonly used to predict the clinical outcome of patients. However, TNM remains inadequate for precise medicine due to considerable heterogeneity in molecular characteristics and clinical behaviors among cancers of the same type and TNM stage.

In contrast, DNA methylation biomarkers hold great promise for improving prognostic accuracy of various cancers. Specifically, DNA methylation involves the addition of a methyl group to the C5 position of cytosine to form 5-methylcytosine, which is one of the most common and important epigenetic changes regulating gene expression. DNA methylation plays a crucial role in carcinogenesis, cancer development, and clinical prognosis. The recent development of DNA methylation microarray platforms enables the analysis of methylation across the genome in a high-throughout manner. Thanks to this profiling technique, a series of DNA methylation signatures with prognostic value for different cancers (e.g., lung, breast, liver, colon and rectum) have been reported in recent years.

Artificial intelligence (AI) is a wide-ranging branch of computer science. It comprises a variety of methods that enable computers to algorithmically learn from data representations and experience, adjust to new inputs, so as to maximize its accuracy of making predictions or classifications. This novel technology has been applied in nearly all walks of life, and there is no exception in molecular cancer research.

Recently, there is a growing body of research that uses both genome-wide analysis approach and AI methods to identify DNA methylation patterns for accurate prediction of prognosis among cancer patients. However, thus far no appropriate review has been published to summarize this recent emerging trend. This is especially important given the fact the type of AI-based methods applied in different studies are heterogeneous, and there might exist some powerful but unexploited AI methods. Therefore, we plan to conduct this scoping review to comprehensively map the studies done in this direction, as well as to identify any existing limitation and research gap.

3) Study aims

Overall, this scoping review aims to summarize how AI methods are being used in genome-wide studies that identify DNA methylation biomarkers for cancer prognosis. The following research questions are formulated:

- (1) What is the most prevalent data processing procedure and AI methods selected?
- (2) What is the characteristics and quality (e.g., source, size) of input dataset?
- (3) How were outcomes evaluated and validated?
- (4) What is the best reported performance for each AI method?
- (5) Is there any study comparing different AI methods using the same dataset? If yes, what is the outcome?
- (6) What is the methodological and reporting quality of existing publications?
- (7) What are the current challenges and limitations of the existing studies, and what are the possible questions that warrant further investigation?

Reagents

Equipment

Procedure

This scoping review will be conducted and reported in accordance with the Joanna Briggs Institute Manual for Evidence Synthesis and PRISMA Extension for Scoping Reviews (PRISMA-ScR).

Eligibility criteria

Studies will be included if they were: reported in English, included patients with cancer of any type at any stage, used an epigenome-wide approach and performed a DNA methylation array, used at least one AI method to explore the association between aberrant methylation and cancer prognosis (i.e., survival, progression, therapy responses).

We will exclude reviews, studies using a candidate-gene approach, multi-omic studies, and studies only investigating diagnostic methylation biomarkers.

Information source and search strategy

We plan to search for eligible studies published between 1 January 1990 and 31 September 2021 in the following databases: PubMed, Web of Science, and Embase. The final search results will be exported into EndNote, and duplicates will be subsequently removed. The electronic database search will be supplemented by searching web search engines (Google Scholar, ClinicalTrials.gov, and Grey Matters) to identify gray literature. References of relevant reviews and full-text articles will be screened for additional studies.

The search strategy will be formulated using both MeSH terms and free-text words related to cancer, AI, epigenetic signatures, and prognosis. Filters will be applied to restrict studies conducted in humans and written in English.

The search term list is as follows:

#1 Cancer

("neoplasms"[Mesh] OR "cancer" [Text Word] OR "tumor*" [Text Word] OR "carcinoma*" [Text Word] OR "pan-cancer" [Text Word] OR "adenocarcinoma*" [Text Word] OR "leukemia" [Text Word])

#2 AI

("ai artificial intelligence [Mesh]" OR "machine learning [Mesh]" OR "decision support [Text word]" OR "clustering [Text word]" OR "deep learning [Text word]" OR "neural network*")

#3 Epigenetic signatures

("methylation"[Mesh] OR "dna methylation*"[Mesh] OR "methyлат*" [Text Word] OR "epigenetic*" [Text Word] OR "epigenetic marker*" [Text Word] OR "DNA-methylation" [Text Word] OR "hypermethylat*" [Text Word] OR "hypomethylat*" [Text Word] OR "CpG" [Text Word] OR "epigenome-wide" [Text Word] OR "methylation-based" [Text Word])

#4 Prognosis

("mortality"[Mesh] OR "survival"[Mesh] OR "survival analysis"[Mesh] OR "mortality"[Text Word] OR "survival" [Text Word] OR "prognos*" [Text Word] OR "predict*" [Text Word] OR "outcome*" [Text Word] OR "recurr*" [Text Word])

Study selection process

The titles and abstracts of retrieved studies will be firstly screened for eligibility by one author. Potentially eligible studies will be retained for full text search, and inclusion will be determined by reading the full text of these studies

A draft charting table/form for data extraction and explanation

The follow study-level key information will be extracted onto a spreadsheet:

1. First author, year of publication, study origin,
2. Endpoints of interest
3. Anatomical site of cancer, data source, data size, types of biological specimens
4. Methodology/methods,
 - (a) Data processing methods and procedures

(b) Type of AI methods

(c) Validation methods

5. Performance measurement

6. Limitations of the study as reported by the authors

Critical appraisal of individual sources of evidence

The reporting quality of all included studies will be assessed by the Reporting recommendations for tumor marker prognostic studies (REMARK). Besides, if a risk prediction model was developed in a study, its reporting quality will be additionally assessed by *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement*, and its methodological quality will be assessed by *PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies*.

Troubleshooting

Step: Identification of eligible studies

Problem and possible reason: Unable to find and download the full text of eligible studies because some journals are not open-access and are not purchased by the research center.

Solutions: Purchase single articles through the research center; contact corresponding authors by email to ask for the full text.

Step: Interpretation of results

Problem and possible reason: Sometimes it is difficult to judge the goodness and the badness of existing pipelines and the statistical methods used due to limited knowledge and lack of benchmark analyses.

Solution: Consult and discuss with experienced statisticians.

Time Taken

The key stages of this scoping review are as follows:

Stage1: Preliminary searches: 1 week (completed)

Stage2: Piloting of the study selection process: 1 week (completed)

Stage3: Formal screening of search results against eligibility criteria: 2-3 weeks (completed)

Stage4: Data extraction and quality assessment: 2-3 weeks (completed)

Stage5: Data charting: 1-2 weeks (ongoing)

Stage6: Manuscript drafting: 3-4 weeks (started)

Anticipated Results

Tables will be used to summarize features extracted from each included study as well as results of quality assessment. The number of publications by year will be summarized in the form of a line chart. The frequency of studies using specific types of AI methods will be presented in a sunburst chart, and the frequency of AI methods used each year will be illustrated by a bubble chart. The pipelines of each study used will be summarized in the form of Sankey diagrams. Descriptive analysis will be used to summarize other important findings.

The findings of this scoping review could be used in future studies to find DNA methylation biomarkers for cancer prognosis in terms of guiding the selection of appropriate analytic approaches, as well as identifying areas for improvement, possible research gaps, and opportunities for innovation.

References